Centre for
Data Ethics
and Innovation

faculty

Bias identification and mitigation in decision-making algorithms

Final Report

June 20

Faculty Science Ltd, 54 Welbeck Street, London, W1G 9XS | info@faculty.ai
Company Registration No. 08873131 UK | VAT Registration No. GB 202 8674 18

Contents

01 Executive Summary

01.1 Introduction

Building on the Centre for Data Ethics and Innovation's (CDEI) Bias Review, this report systematically reviews the emerging set of approaches for detecting and mitigating algorithmic bias, and explores their practical application, both in technical tooling and in the financial services and recruitment sectors.

The concept of bias is difficult to define. It is intrinsically linked to the broader concept of *fairness*, and indeed CDEI's Bias Review defines bias as "'[referring] to an output that is not only skewed, but skewed in a way which is *unfair*".

Fairness as a concept can be traced to the earliest schools of philosophy, and along with Justice, is perhaps the most debated issue within politics and social policy. In very recent years, fairness has also become a core part of recent developments in AI and Machine Learning, with the world's largest tech firms all making substantial investments in fairness algorithms. This is perhaps not a coincidence. AI and wider algorithmic decision making will need to show it can stand up to the same ethical standards as more classic social policy tools if it is to achieve widespread acceptance and adoption.

By nature, historical notions of bias and fairness have typically not been precisely defined, regarded as abstract qualities, rather than notions that can be quantitatively measured. They are also very context specific: it is difficult to codify hard rules, even in abstract terms, about what is 'fair' or 'unfair', or why.

This does not sit easily with algorithmic decision making. Achieving fairness within algorithms requires not only definition, but measurement, if it is to be practically implemented. It is also unreasonable to expect a single, holistic mathematical definition of fairness to emerge which can have generalised application to all contexts.

In recent years, myriad different approaches have been proposed in the Machine Learning literature for both defining fairness and for intervening to mitigate biases, and practical tooling is starting to emerge which implements these approaches. As this report sets out, these have merit and indeed we believe that approaching fairness as a mathematical/statistical discipline is essential if AI and Machine Learning are to achieve widespread adoption.

It is also important to recognise the limitations, both individually and collectively, of algorithmic fairness approaches, especially for organisations and practitioners seeking to operationalise them. No one definition is holistic by nature. For instance, they are often mutually incompatible; there is as yet no clear decision making framework for selecting between them; and crucially they are generally relatively insensitive to wider philosophical notions of fairness which are by nature hard to codify. Furthemore, the legal framework has not caught up, such that there is a sufficient body of guidance or case law about which algorithmic approach to follow in practice, or what can be considered a legally-compliant approach in different circumstances.

Taken together, this poses a complex set of challenges for practitioners, organisational leaders, and policymakers alike. They need an ability to understand the myriad different algorithmic definitions and intervention approaches, and how they fit together. Crucially, what is really missing is practical guidance about how to create fair algorithms, monitor and audit them using the most up-to-date tools and techniques.

01.2 About this work

To answer this challenge, Faculty was commissioned by the CDEI to assess and compare the various approaches to bias mitigation; to understand the use of these approaches in key sectors of the economy; and to develop practical guidance for firms looking to implement fair approaches and monitoring in their business processes. Faculty has developed three deliverables, which together produced:

- A main report, this document
- An accompanying Implementation Handbook, with how-to-guidance and technical standards for organisations to use to achieve fairness on algorithms in operational use
- A demonstration web application, showing how fairness can be achieved in practice via two reference implementations

Faculty approached this work in four stages:

1. We have undertaken a **systematic literature review** of the latest state of the art on ways of defining algorithmic fairness and intervening to mitigate bias. While unlikely to be fully exhaustive, this brings into one place all of the key definitional and intervention approaches and compares them side-by-side and assesses their strengths and limitations.

2. We have brought all of the key concepts together into a **digestible end-to-end framework,** we believe this is the first time it has been done:
    - Firstly we introduce an overall organising framework for algorithmic fairness, and map the different algorithmic fairness definitions identified in the literature against this
    - Secondly we map these definitions of fairness against ways to intervene to achieve them - considering both the timing of intervention and the different intervention approaches

3. To understand the interpretation of fairness within industry and practical use of fairness tools, we conducted an **Industry Review** focussing on the Financial Services and Recruitment. We undertook interviews with a small sample of firms, academics and industry bodies to understand the revealed practices and current audit approaches in these sectors, summarising our findings and providing detailed case studies of relevant firms. It is worth noting that our consultation is not representative of the industries as whole, but rather provides a snapshot on the practical use of these tools.

4. To support firms and their technical teams as they begin to practically implement algorithmic fairness, we provide an accompanying **Implementation Handbook**. This contains:
    - A generalisable workflow of questions that guides a firm through the process of building, monitoring and reviewing fair algorithm deployments.
    - A corresponding set of technical questions and standards that provide technical teams with the guidance, supporting mathematical notation, and a review of prominent AI fairness tools that they can draw upon.

01.3 Structure of the report

This main report contains the following sections:

- **Section 02**: describes and assesses the different algorithmic notions and definitions of fairness, and places them within a conceptual framework.

- **Section 03**: builds from this conceptual framework to map the different ways of intervening to achieve fairness, including timing of intervention and specific intervention approaches.

- **Section 04**: analyses the overall limitations and inherent challenges with algorithmic fairness approaches, but goes on to explain their central importance and usefulness, stemming from their precise statistical definition and quantification.

- **Section 05**: summaries our deep-dive into the financial services sector, finding that Fairness Through Unawareness is currently pervasive although firms are becoming increasingly sophisticated in their approaches, partly in expectation of further regulation.

- **Section 06**: summarises our deep-dive into the recruitment sector, finding that there is an explosion of interest in this topic and tools becoming available, and that key debates are underway about whether algorithmic decision making should seek to correct historic recruitment biases, or risks exacerbating it.

- **Section 07:** concludes the report.

- **Annex A**: assesses the broader set of algorithmic definitions of fairness identified in the literature, including those not yet ready for operational use at scale.

- **Annex B:** contains a detailed list of the references used and wider literature that we drew on to support the report.

01.4 Acknowledgements

This report would not have been possible without the rich academic literature around mathematical notions of algorithmic fairness. As the references section makes clear, we made full use of this canon - and we pay tribute to the academics and researchers whose contributions enabled our work. In particular, we'd like to thank the Institute for the Future of Work and their recent report on Artificial Intelligence in hiring, which guided our own views on the recruitment sector and equality considerations.

We are also indebted to those who gave up their time and energy to support the writing of this report. We interviewed leaders and data scientists from across the recruitment and financial sectors, drawing on their insights to shape our understanding of practical deployments of algorithms. In particular, we'd like to thank Neil Carberry of the Recruitment and Employment Confederation for his support in this area.

Finally, we are hugely grateful to the wisdom and oversight of the CDEI's Bias Review Steering Group for their steers and drafting suggestions as the report and accompanying material was finalised.

02 Algorithmic Notions of Fairness

Bias, as defined in the CDEI bias review, "[refers] to an output that is not only skewed, but skewed in a way which is unfair". This immediately raises the question of how precisely we define (un)fairness.

Fairness, and other notions such as Justice, are perhaps the most debated issues within politics and social policy, but ever since Aristotle, have typically been left undefined. Furthermore, they are typically regarded as abstract qualities, rather than notions that can be quantitatively measured.

This does not sit easily with algorithmic decision making. Achieving fairness within algorithms requires not only definition, but measurement. And crucially, measurement is necessary if fairness is to be traded-off successfully with other measurable qualities of an algorithm, in particular a model's predictive power, which is typically directly associated with business performance.

We need to recognise that a reasonable qualitative interpretation of what is societally seen as 'fair' (i.e. the abstract quality) is often context-specific. Consequently it is challenging to come up with quantitative formulations of fairness that capture all of the nuances that can arise. As a result the academic literature has seen the introduction of dozens of competing notions of fairness, each with their own merits and drawbacks, and many different terminologies or ways of categorising these notions, none of which are complete.

In this section of the review we introduce many of the more commonly used notions, commenting on some of the trade-offs between different notions that should be taken into account when selecting a particular definition.

For the first time, we also seek to structure these into a single overall framework, capturing the different lenses by which the notion of fairness can be understood, then populate the commonly-used notions in the literature in that framework according to their substantive features.

As shall be explored, partly in this section and then in aggregate in Section 04, these notions of fairness can overlap in different ways, which can add complexity to the policymaker and practitioner. Some are by nature mutually exclusive – it is impossible, for example, to achieve Demographic Parity ('Independence') and Equalised Odds ('Separation') concurrently if there are underlying differences in the features of different groups - whereas some can be achieved simultaneously.

02.1 Algorithmic fairness framework

As above, the literature is awash with not only myriad different notions of fairness, but also myriad different terminologies and ways of describing these. Furthermore, there are different levels that these can be categorised at, for example distinguishing at a high level between 'procedural' and 'outcome' fairness.

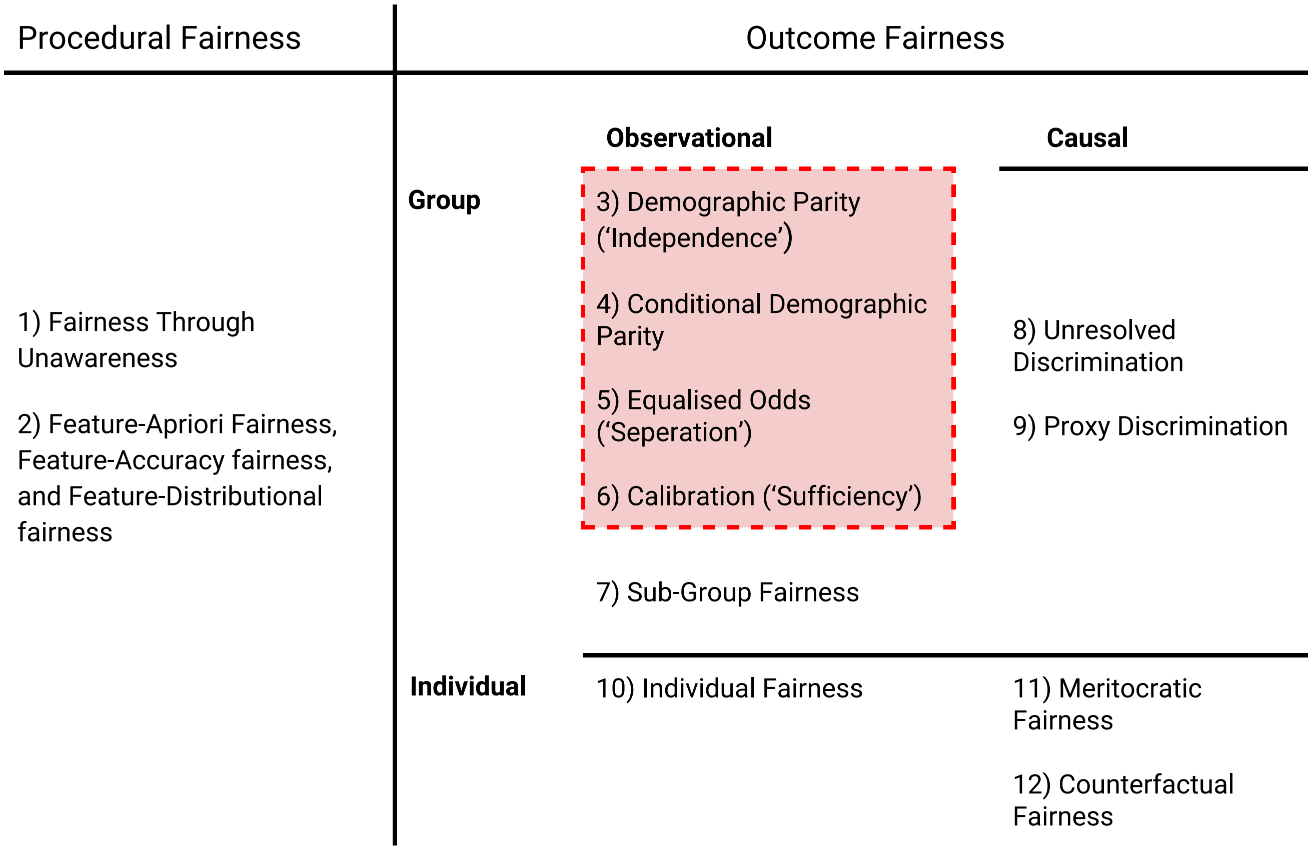
This can be confusing to both data science practitioners and wider stakeholders. We therefore try to set out an overall framework for categorising the different notions of fairness, in a way that can make sense of these different concepts (Figure 3 below).

In Section 03 and Annex A, we then extend this framework out to the different intervention approaches for achieving these notions of fairness. More precisely, in Section 03.1 we introduce intervention methods addressing fairness notions from the framework which have most practical relevance, while in Annex A.1 we present intervention techniques addressing remaining notions.

We then map the current set of available open-source AI fairness tools against these – in doing so, this can be read as an end-to-end framework, giving the practitioner line of sight all the way from high-level concepts of fairness all the way to the tooling available to achieve these.

The below figure summarises our categorisation. *Note: we have capitalised terms defined in the table for clarity throughout the report.*

We distinguish between two schools of thought: Procedural Fairness and Outcome Fairness. Within Outcome Fairness, we can make two additional distinctions between Causal and Observational notions of fairness, as well as Individual and Group notions.



Procedural Fairness

1) Fairness Through Unawareness

2) Feature-Apriori Fairness, Feature-Accuracy fairness, and Feature-Distributional fairness

Outcome Fairness

Observational

Causal

Group

3) Demographic Parity ('Independence')

4) Conditional Demographic Parity

5) Equalised Odds ('Seperation')

6) Calibration ('Sufficiency')

7) Sub-Group Fairness

8) Unresolved Discrimination

9) Proxy Discrimination

Individual

10) Individual Fairness

11) Meritocratic Fairness

12) Counterfactual Fairness

Figure 3: Organising framework of different algorithmic fairness notions
Note: the notions of fairness in the red highlighted box refers to the notions we prioritised, based on how amenable the notions are for practical use.

These are then described and assessed in detail in the remainder of this section.

02.2 High-level distinction: Procedural vs. Outcome fairness

Approaches to formalising the definition of fairness firstly fall into two broad categories: Procedural Fairness and Outcome Fairness.

02.2.2 Procedural Fairness

Procedural Fairness is concerned with fair 'treatment' of people, i.e. equal treatment within the *process* of how a decision is made. In the specific context of Machine Learning this often means considering what information is given to the algorithm with which to make a decision. To date, as set out below, this has often led to Procedural Fairness being interpreted as not including a protected or sensitive attribute in the process when making a decision – so called 'Fairness Through Unawareness' – partly for simplicity but also partly to avoid introducing legal risk. However, as will be explored later, this is not often an effective strategy: indirect discrimination can still occur through the presence of other variables or features within a model which are correlated with the protected attribute in question.

More widely, as a Machine Learning model is embedded into an overall decision-making process, Procedural Fairness can encompass wider elements of the decision-making process. These include aspects such as offering opportunities for individuals to challenge decisions about themselves, and seek redress. In this report, we focus on the Machine Learning aspects of Procedural Fairness only.

02.2.2 Outcome Fairness

By contrast, *Outcome Fairness* is concerned with ‘fair results’ of a decision making process, i.e., equality across different groups with regards to the *outcomes of actually made decisions*. Most of the existing literature on algorithmic fairness falls in this category (Grgić-Hlača et al. 2018).

02.3 Conceptual breakdown of Outcome Fairness

Outcome Fairness can be further divided along two axes: *Individual* vs. *Group* notions of fairness, and *Causal* vs. *Observational* approaches. These two axes are complementary, and any combination of the two choices has been considered in the literature, i.e. both Group and Individual Fairness has been studied from Observational and Causal perspectives.

02.3.1 Group vs. Individual

Fairness can be studied both at the Group level or the Individual level leading to different classes of fairness measure. Group notions of fairness first aggregate outcomes by group, then compare the aggregated group outcomes to determine whether the outcomes are fair. Individual notions of fairness compare the outcomes for individuals to determine whether the outcomes are fair.

In order to clarify this distinction, let us consider an example from hiring: A typical Group fairness notion relevant in this context (Equalised Odds, below) would ask that among the qualified applicants men and women are invited at the same proportion. Individual Fairness would ask that similarly qualified applicants have a similar chance of being invited, then if not, explore how the differences correlate with gender.

It is worth noting that Group and Individual notions are not mutually exclusive: an idealised ‘fair’ algorithm could achieve both simultaneously.

02.3.2 Causal vs. Observational

Many of the notions of fairness that have been introduced are *Observational* in nature, that is they can be formulated fully in terms of the joint distributions of outcomes, decisions, features and sensitive attributes.

The field of causal inference allows us to go beyond this, incorporating knowledge about *how* variables influence each other rather than just measuring correlations. Consequently we can consider the effect of interventions, or counterfactuals (e.g. what would have happened if the sensitive data were different?).

These tools and ideas have been applied to the study of bias in decision making algorithms. There is an attractive alignment between causal notions of bias and intuitive understanding of bias, but a requirement for performing a causal analysis is typically the specification of a causal model of the data, which can be a restrictive requirement. We discuss these advantages and disadvantages in detail in Section 04.

02.4 Assessing algorithmic definitions of fairness against this framework

Below we describe and assess the core algorithmic definitions of fairness which have emerged in the literature and which we have mapped into the conceptual framework above.

We prioritise fairness notions from the framework that are most amenable for practical use and applicability and present those here, while we refer to the remaining ones in Annex A1. It is worth drawing out some general statements about this, to explain our reasoning for inclusion: *(Note: the numbers in parentheses refers to the notions numbered in Figure 3 above)*

- *Group Observational notions (notion numbers 3-6 in the highlighted box, at the centre of the framework in Figure 3)*, lend themselves best to current practical application, as they are simple to compute and provide meaningful measures for aggregated differences between groups. Further, the majority of existing mitigation tools address one of Demographic Parity, Conditional Demographic Parity or Equalised Odds. Besides, we introduce Calibration which is closely related to model performance but also yields a useful measure for aggregated differences between groups relevant for fairness considerations. They are therefore the primary focus of the remainder of this report, the accompanying Implementation Handbook, and the demonstration web application.
- *Fairness Through Unawareness (1)* is in common use, either as a deliberate approach or as a default, including widely in the finance and recruitment sectors. We therefore include an assessment of it in this main report, albeit note its substantial flaws.

By contrast, the fairness notions in the Annex are generally in less common use as the ones we introduce here, and practical implementations of mitigation techniques that address these notions of fairness are fairly limited:

- *Other procedural notions (2)* are promising, but require access to information which might not be available and therefore impedes direct applicability, e.g. detailed input from stakeholder panels (Feature-Apriori Fairness)
- Causal notions (8, 9, 11, 12) are again promising but rely on causal graphs which govern the fairness notion, which are not generally available.
- Individual and Sub-Group notions (7, 10) capture the idea that “similar individuals should be treated similarly”. However, there is no general way to define similarity for individuals, and constructing a suitable metric in a way that doesn’t itself reflect the exact biases we want to mitigate is extremely challenging. As a result individual fairness is generally less practical than other notions of fairness.

Note: *A specific algorithmic definition of fairness is sometimes introduced in the literature under a variety of different names. We acknowledge this fact by stating the existing alternative namings under the appropriate fairness definition, while for our own namings we follow the convention of choosing what has been most widely used in the literature.*

02.4.1 Procedural definitions

1) Fairness Through Unawareness

Fairness Through Unawareness is characterised by withholding from the algorithm access to the protected attributes or any close proxies. The premise is that if the model is “unaware” of the protected attributes then it cannot discriminate with respect to those attributes.

Fairness Through Unawareness is truly a procedural approach in that it requires that protected attributes are not used at all by the decision-making algorithm, and doesn’t make any requirement of the outcomes.

The primary criticism of Fairness Through Unawareness is that it is uncontrolled, and because of correlations in the data, there is no guarantee that the model is not able to discriminate. For example we might withhold information about each individual’s race, but allow the model to use the individual’s postcode to make decisions. If postcode happens to be highly predictive of race then the model could indirectly discriminate.

Also known as: Anti-classification (Corbett-Davies and Goel 2018)

Categorisation: Procedural

02.4.2 Group-based Outcome definitions

3) Demographic Parity

Demographic Parity, also often referred to as 'Independence', is a measure of Group fairness which requires that outcomes for different protected groups are equally distributed. Specifically we require that the probability distributions of protected attributes and model outputs are statistically independent (Calders, Kamiran, and Pechenizkiy 2009; Feldman et al. 2015; Faisal Kamiran and Calders 2012; Zafar, Valera, Rodriguez, et al. 2017).

In the context of an algorithm that assigns a score to individuals, which is then thresholded to obtain a decision – e.g. a risk score in a loan approval algorithm where loans are approved to all who receive a risk score below some threshold – we can apply the Demographic Parity constraint at two different levels: at the score or the decision levels. In the former case, we require that the distribution of scores is the same for all protected groups, which is to say the proportion of individuals from each group receiving a score in a particular range is equal. In the latter case, we require only that the proportion of individuals being approved is the same across all groups. It's not hard to see that the former implies the latter, but not vice versa.

A number of criticisms have been made of Demographic Parity. The first is that the requirement that there be no association between protected attributes and model outputs is often in tension with the underlying task, specifically whenever there is a correlation between the protected attribute and the target variable. In particular, it is possible that a perfect predictor, that is able to correctly predict the true outcome in every instance, would itself not satisfy Demographic Parity.

On the other hand, a model satisfying Demographic Parity may have limited utility. Indeed, the disconnect between model utility and Demographic Parity can perversely lead to increased unfairness between groups. Consider a hypothetical model that correctly identifies qualified candidates from one group, and randomly picks candidates from another group at the same rate. Such a model satisfies Demographic Parity, but can cause harm to the disadvantaged group through poor predictions. In this example, a mathematical notion of fairness has been achieved, but *allocative efficiency* strongly matters: overall utility is reduced if unsuitable candidates are selected. Although a stylised example, it highlights that the disutility effects are such that all groups can be made worse off, even the ones for whom the Demographic Parity condition has defined the outcome as 'fairer'.

Demographic Parity is appropriate in cases where we truly believe that the protected attributes should have no bearing on the outcomes, for example a face detection algorithm – that might be used to zoom / pan a webcam for video conferencing – should not be more likely to detect faces for one race than another. However, in other cases the argument for Demographic Parity may not be so clear. For example in recruitment, certain protected groups may be less likely to have the required qualifications due to socioeconomic disadvantages. Though we may idealise a world in which such systemic biases didn't exist, ignoring them and imposing Demographic Parity in order to redress such imbalances can actually have harmful consequences.

Also known as: Statistical parity (Dwork et al. 2012), Group fairness (Dwork et al. 2012), Independence (Barocas, Hardt, and Narayanan 2019)

Categorisation: Outcome -> Group - > Observational

4) Conditional Demographic Parity

Conditional Demographic Parity is a generalisation of Demographic Parity that takes into account certain "legitimate risk factors" with respect to which we do not consider it unfair to discriminate. Specifically, an algorithm satisfies Conditional Demographic Parity if the model outputs are statistically independent of the protected attributes once the aforementioned legitimate risk factors are taken into account[1] (Faisal Kamiran, Žliobaitė, and Calders 2013; Corbett-Davies et al. 2017).

Let us return to the loan approval example. A legitimate risk factor might be the annual income of the individual. Whereas with Demographic Parity we require that the distribution of scores is the same for all protected groups, here we require that among individuals with a particular annual income, the distribution of scores across the protected groups is the same. Hence if one particular group on average has lower annual incomes, then overall they might receive lower scores, however individuals with the same annual income are treated the same, regardless of their protected attributes.

Much like Demographic Parity, we can apply Conditional Demographic Parity at either the score level or the decision level.

Conditional Demographic Parity addresses some of the weaknesses of Demographic Parity, for example a perfect classifier can achieve Conditional Demographic Parity even if there is a correlation between the protected attributes and the outcomes provided that the legitimate risk factors fully account for this correlation. In the case of the running loan approval example, if protected groups differed in their average ability to pay back a loan, but this was fully accounted for by differences in annual income across the groups then hypothetically a model could be both 100% accurate and still achieve Conditional Demographic Parity.

A drawback of Conditional Demographic Parity is that if the legitimate risk factors contain historical biases, which is likely the case with annual income in our example, then imposing Conditional Demographic Parity could lead to those historical biases being perpetuated.

Categorisation: Outcome -> Group -> Observational

5) Equalised Odds

Equalised Odds (also commonly referred to as 'Separation') is another observational notion of Group fairness, whose introduction was motivated by the tension between performance and fairness that Demographic Parity suffers from. Intuitively it requires that qualified and unqualified candidates are treated the same, regardless of their protected attributes. More precisely, an algorithm satisfies Equalised Odds if the decisions are statistically independent of the protected attributes, conditioned on the outcome (Hardt, Price, and Srebro 2016; Zafar, Valera, Gomez Rodriguez, et al. 2017; Kleinberg, Mullainathan, and Raghavan 2016).

Intuitively, this means that our model scores can depend on the protected attribute, but only in so far as the true outcomes do. Equivalently, the distribution of model outputs across different protected groups are the same when outcomes are held fixed. In the case of binary classification, Equalised Odds can be summarised as the requirement that true positive rates are the same for all protected groups, as are the false positive rates. In other words, the chance that a qualified individual is overlooked, or that an unqualified individual is approved, is the same across all protected groups.

[1] By "taken into account" we mean statistically independence conditional on the legitimate risks.

Equalised Odds admits a perfect model as a valid option, and is further able to deal with unequal base rates between protected groups. While being of significant practical relevance (Larson et al. 2016), Equalised Odds results in different groups being held to different standards (see *Calibration*) when those groups exhibit different risk distributions (Corbett-Davies et al. 2017). The latter can be observed when the risk distribution is related to feature choice, e.g., number of prior arrests being an important indicator for recidivism of convicts while being correlated with race.

Furthermore, where systemic bias is present in the labels, Equalised Odds is at risk of perpetuating this bias. For instance, assume that for the same type of jobs advertised by a company, there is a larger proportion of qualified applications in group one than in group two, while both groups have the same size. Then, adhering to equal opportunity results in hiring a larger proportion of candidates in the first group than in the second. If the advertised jobs are good jobs, they will generally improve the living condition and education for their employees' children, which in turn leads to them having better professional opportunities, and so on.

Also known as: Conditional Procedure Accuracy (Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth 2018), Disparate Mistreatment (Zafar, Valera, Gomez Rodriguez, et al. 2017), Separation (Barocas, Hardt, and Narayanan 2019)
And relaxations of it: Balance for positive / negative classes (Kleinberg, Mullainathan, and Raghavan 2016), Predictive Equality (Chouldechova 2017), Equalised Correlations (Woodworth et al. 2017)

Generalisations

- **Equality of opportunity**: A relaxation of Equalised Odds that requires only that qualified candidates are treated equally across all protected groups. Equivalently, the true positive rates across all protected groups is the same, but there is no requirement made of false positive rates.
- **Predictive equality:** Relaxation of the above checking among negative outcomes (applicable when negative outcome is considered as the advantaged one - e.g. COMPAS)

Categorisation: Outcome -> Group -> Observational

6) Calibration

An algorithm satisfies Calibration if the outcomes are statistically independent of the protected attributes, conditioned on the decision (Crowson, Atkinson, and Therneau 2016; Pleiss et al. 2017; Grgić-Hlača et al. 2018) (first cited in a few papers in medical world - Crowson).

In other words, the decision making algorithm captures all of the influence of the protected attribute on the outcomes, once the model has made a prediction or assigned a score. We would gain no more information about the true outcome than is already contained in the prediction. Consider a model that systematically underpredicts for a particular protected group. This model would not satisfy Calibration, because given a prediction for an individual, if we found out the individual was from that protected group we would adjust our expectations for the true outcome. That is, the group membership gave us some additional information that was not captured by the prediction.

Calibration is closely related to the more common use of the term "Calibration" in Machine Learning. There, Calibration refers to model probabilities being representative of the real world outcomes. Specifically, aggregating all data points for which the model predicted a 70% chance of a particular outcome, we would hope to see that outcome occurring about 70% of the time. Calibration as used by the fairness community is equivalent to the model being well calibrated, in the conventional sense, on each protected group.

A big advantage of Calibration is that it is well aligned with accuracy objectives in many models, in particular many Machine Learning models tend to be approximately calibrated simply as a consequence of the way they have been trained.

On the other hand, this means that imposing Calibration on a model may not represent a particularly significant intervention. Furthermore, since Calibration takes into account the true outcomes, like Equalised Odds it is susceptible to perpetuating historical biases.

Calibration across more complex combinations of multiple protected attributes has been addressed in (Hébert-Johnson et al. 2018; Kilbertus et al. 2018).

Also known as: Sufficiency (Barocas, Hardt, and Narayanan 2019), Conditional Use Accuracy (Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth 2018), Relaxations of Calibration: Predictive parity (Chouldechova 2016)

Categorisation: Outcome -> Group -> Observational

02.5 Key considerations of the different definitions

At this stage we want to introduce two key issues when considering the different algorithmic definitions above: the trade-off between fairness and accuracy, and the mutual compatibility of the different definitions:

02.5.1 The trade-off between fairness and accuracy

Whether in broader social policy, or in algorithmic decision making, there is often (though not always) an inherent trade-off between *fairness* and *accuracy*. That is, whether to maximise the overall objective function (e.g. prediction accuracy of future loan defaults) or ensure that it is fair to different participants. In the case of algorithmic decision making, this trade-off requires quantification.

An attraction of the mathematically-derived algorithmic fairness notions set out above is that they can be measured on a continuous scale, and so allow the practitioner to evaluate quantitatively the degree to which a certain fairness criterion is satisfied. Different definitions of (un)fairness can be expressed via different measures. Often, the degree of unfairness can be formalised as a positive number, i.e., the smaller the number the more fairness holds.

For less fair models, interventions can be applied which increase the model's fairness in its predictions, as explored in section 03 below. However, it is a general paradigm in Machine Learning that the accuracy of a model decreases with increasing fairness, or according to Berk et al. (2018), "demanding fairness of models will always come at a cost of reduced predictive accuracy". The severity of this trade-off depends on a number of factors, such as the underlying task and data at hand, the choice of fairness notion and the way of measuring it and the applied method of mitigating unfairness.

Written theoretically in this way arguably understates both the ethical and practical importance of this trade-off: this trade-off is at the heart of how we want algorithmic models to behave, just as is at the heart of questions of social policy and social justice. This is arguably the core issue when thinking about how to mitigate bias within algorithms in practice. Our choices with respect to accuracy and fairness reveal the aims and desires for how a given model both behaves and is used in context.

Very often, on a practical level, the *accuracy* of models will not only directly affect profitability, but also be a source of competitive advantage, while also being an ethically justifiable aspiration of models in its own right. Understanding and evaluating this trade-off is therefore crucial for creating adaptable policy for AI fairness. At the core of this stands the question what level of fairness is considered necessary and useful given a certain task and fairness notion.

In a simplified view, higher levels of fairness may be desirable for the purpose of equality across society. However, higher levels of accuracy may be desirable for the purpose of efficacy with regards to the primary goal of the underlying model, e.g., a higher performance of a credit scoring model can lead to a higher short-term profitability of its business. That said, this stylised trade-off of societal fairness vs short-term predictive accuracy/profitability is not as clear a dichotomy as this simplified view. Indeed there is increasing recognition across industry of the long-term business benefits, which can come from promoting greater fairness. Details on the tension between model accuracy and fairness are addressed in section 04.

02.5.2 Mutual compatibility between algorithmic fairness definitions

Different notions of fairness can be satisfied simultaneously only in certain cases, e.g. Individual and Group fairness (Zemel et al. 2013). They are typically mutually exclusive, meaning that neither two definitions can be satisfied simultaneously (Kleinberg, Mullainathan, and Raghavan 2016; Pleiss et al. 2017; Corbett-Davies et al. 2017; Lipton, McAuley, and Chouldechova 2018; Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth 2018).

As a consequence, a practitioner aiming to mitigate bias generally has to make a choice as to which notion of fairness to enforce, thereby necessarily trading off other notions (Corbett-Davies et al. 2017).

In general, the decision over which measure of fairness to impose needs an extensive contextual understanding and domain knowledge. One should also understand the sources of bias and downstream consequences of a fairness intervention before imposing it on the model's decisions. This is explored in more detail in section 04 of this report, as we build towards a generalisable approach for practitioners when seeking to achieve fairness within algorithms in the accompanying Implementation Handbook.

03 Fairness interventions

Having defined and assessed the different technical notions of fairness currently in the academic literature, we next set out a framework for summarising and assessing the *interventions* to achieve these, particularly according to the timing at which the intervention is made.

We have also mapped the current suite of leading algorithmic fairness tools available against this framework, set out in the accompanying Implementation Handbook. This gives practitioners and policymakers, we believe for the first time, a structured end-to-end view of latest developments of applying algorithmic fairness approaches in practice, running through each of:

A. High-level distinction between Procedural and Outcome fairness
B. Different conceptual types of Outcome fairness
C. Different algorithmic definitions of fairness, mapped against these
D. Intervention timing, mapped against the different definitions
E. Theoretical intervention approaches mapped against this
F. And finally, practical open-source tools available, again mapped

03.1 Intervention framework

Below in Figure 4, we set out an overall framework for categorising the different intervention approaches to detect and mitigate bias. This categorises intervention approaches according to:

1. The timing of intervention – either 'pre-processing', 'in-processing', or 'post-processing'
2. Which notion of fairness is sought – according to the notions set out in Section 02

In Figure 4, we populate the framework with the latest technical approaches identified in the literature. We summarise each of them briefly and give more details in the Technical Standards, including available open source implementations.

As before, we prioritise for fairness notions with largest practical relevance, and hence focus here only on intervention methods addressing the definitions detailed in section 02.4. The intervention methods addressing the remaining fairness notions are discussed in Annex A.2.

		Pre-processing	**In-processing**	**Post-processing**
Group notions	**Demographic Parity**	**Data reweighting / resampling**: - (Calders, Kamiran, and Pechenizkiy 2009) - (Faisal Kamiran and Calders 2012) **Label modification: -** (Calders, Kamiran, and Pechenizkiy 2009) - (Faisal Kamiran and Calders 2012) - (Luong, Ruggieri, and Turini 2011) **Feature modification:** **-** (Feldman et al. 2015)	**Constrained Optimisation** - (Corbett-Davies et al. 2017) - (Agarwal et al. 2018) - (Zafar, Valera, Rodriguez, et al. 2017) **Regularisation**: - (Kamishima et al. 2012) **Naive Bayes/Balance models for each group:** - (Calders and Verwer 2010) **Naive Bayes/Training via modified labels:**	**Naive Bayes/Modification of model probabilities:** - (Calders and Verwer 2010) **Tree-based leaves relabelling**: - (F. Kamiran, Calders, and Pechenizkiy 2010) **Label modification: -** (Lohia et al. 2019) (F. Kamiran, Karim, and Zhang 2012)

		Optimal clustering / constrained optimisation: - (Zemel et al. 2013) - (Calmon et al. 2017) **Auto-encoding:** - (Louizos et al. 2016) :	- (Calders and Verwer 2010) **Tree-based splits adaptation:** - (F. Kamiran, Calders, and Pechenizkiy 2010) **Adversarial debiasing:** - (Zhang et al. 2018) - (Adel et al. 2019)	
	Conditional Demographic Parity		**Constrained optimisation**: - (Corbett-Davies et al. 2017) **Adversarial debiasing:** - (Zhang et al. 2018) - (Adel et al. 2019) - by passing cond. variable to adversarial	
	Equalised Odds		**Constrained optimisation**: - (Corbett-Davies et al. 2017) (*predictive equality)* - (Agarwal et al. 2018) - (Zafar, Valera, Gomez Rodriguez, et al. 2017) - (Woodworth et al. 2017) **Adversarial debiasing:** - (Zhang et al. 2018) - (Adel et al. 2019)	**Decision threshold modification (ROC curve)/ constrained optimisation:** **-** (Hardt, Price, and Srebro 2016) - (Woodworth et al. 2017)
	Calibration		**Unconstrained optimisation**: - (Corbett-Davies et al. 2017)	**Information Withholding**: - (Pleiss et al. 2017) - *achieves simultaneously a relaxation of Equalised Odds*

		Pre-processing	In-processing	Post-processing
Individual notions (see Annex A)	**Individual Fairness**	**Optimal clustering / constrained optimisation:** - (Zemel et al. 2013)	**Constrained optimisation:** - (Dwork et al. 2012) - (Biega, Gummadi, and Weikum 2018)	**Label modification: -** (Lohia et al. 2019)
	Counterfactual Fairness	**Prediction via non-descendants in causal graph:**		

	- (Kusner et al. 2017)		
Subgroup Fairness		**Two-player zero-sum game:** - (Kearns et al. 2018) - (Kearns et al. 2019)	

Figure 4: Fairness intervention framework

There are four important points to stress, when reading the rest of this section:

- For the intervention approaches set out in Figure 4, we provide a short title description for each, which describes the essence of the approach used (e.g. 'constrained optimisation'). This is not necessarily the terminology used by the authors, but we felt this was better to aid clarity and consistency.
- Several of the notions of fairness have multiple intervention approaches: some notions of fairness have been studied more extensively than others, and may lend themselves to different mitigation strategies, each with their own trade-offs, such as performance and complexity.
- Some notions of fairness are only achievable during algorithm development or retrospectively: for example, for Calibration, the notion of fairness is tied to the classification task (in the sense that it depends on the labels), and consequently can only be performed in- or post-processing.
- We have assessed the intervention approaches relating to Individual notions of fairness in Annex A.2, on the basis they are currently less amenable to widespread application.
- Some of the notions of fairness explored in section 02 do not yet have a practical intervention approach developed, and hence are not yet included in the framework above.

03.2 Intervention time

Building a decision making algorithm is a multi-stage process, so there are numerous opportunities to intervene to correct unfairness. The stage of the process at which an intervention is made is a useful distinction, as it has direct implications on the available methodology and type of intervention. Hence, we shall distinguish between approaches along this axis.

03.2.1 Pre-processing

Pre-processing interventions take place before the model is created. They generally make a modification to the data that the model will be trained on, aiming to remove possible sources of unfairness before the model even sees the data. Examples of preprocessing can be found further down in this section.

There are advantages to taking this approach. First, once data has been pre-processed, it can in principle be used for any downstream task, the intervention must only be made once. Second is that any model that is trained on the data does not need to be modified itself, hence pre-processing interventions are generally fully model agnostic.

However, there are also some general limitations. The first is that while it may seem attractive that data only needs to be pre-processed once before it can be used in multiple downstream applications, this view neglects the fact that the nature of the downstream task usually ought to inform selection of an appropriate notion of fairness. Hence, we cannot really decouple pre-processing and the task requiring the intervention.

Moreover, pre-processing data in an application agnostic way means we cannot incorporate labels into the pre-processing, which in turn means we cannot address notions of fairness that are stated in terms of the class labels such as Equalised Odds or Calibration. Indeed, most of the pre-processing interventions present in the literature do not incorporate outcomes, only model inputs and protected attributes. Hence they are not well suited to addressing notions of fairness, such as Equalised Odds, where the definition is coupled to the outcomes. Instead, they address definitions of fairness, such as Demographic Parity which can be formulated without knowledge of the outcomes.

Finally, these approaches often end up being less performant than other approaches in terms of the trade-off between fairness and accuracy, in part because they often aren't able to co-optimise fairness and accuracy on a specific task. That being said, pre-processing approaches do not need to be used in isolation, and could be a first step in a pipeline that incorporates additional interventions.

Some preprocessing methods only require access to the protected attributes in the training data, however not in the test data (Calders, Kamiran, and Pechenizkiy 2009; Zemel et al. 2013).

Examples of pre-processing algorithms

We order example algorithms according to the fairness notion they address. Pre-processing methods generally intend to achieve Demographic Parity, or as in (Zemel et al. 2013) both Demographic Parity and Individual Fairness.

Demographic Parity

Data reweighting / resampling: The authors (Calders, Kamiran, and Pechenizkiy 2009) present a pre-processing approach that attaches weights to the data, so certain types of observations are more influential during training, thereby balancing out the label distributions across different protected groups. The resulting weights can also be used to resample the data set with replacement to create a fair transformed data set (Faisal Kamiran and Calders 2012).

Label modification: In (Calders, Kamiran, and Pechenizkiy 2009; Faisal Kamiran and Calders 2012), an approach is introduced that changes the labels on qualified data points which are selected according to a ranking algorithm in order to eliminate any disadvantage a protected group may have. The authors (Luong, Ruggieri, and Turini 2011) apply a k-Nearest Neighbours (kNN) approach that flags data points as being discriminated if a significant difference in decision outcomes is found among their neighboring points belonging to the protected group compared to their neighboring points not belonging to it. The labels of flagged points are then flipped.

Feature modification: The approach of (Feldman et al. 2015) adjusts the marginal distributions of each feature across different protected groups so that they agree, thus reducing correlation between features and protected attribute.

Optimal clustering / constrained optimisation: (Zemel et al. 2013) proposes a clustering method which transforms the original data set by expressing points as linear combinations of learnt cluster centres so that the transformed data set is as close as possible to the original while containing as little information as possible about the sensitive attributes. Predicted labels of the transformed data set can be defined so that similar points are mapped to a similar label prediction. In that sense, Individual Fairness is achieved, too. Further, in (Calmon et al. 2017) a probabilistic mapping from the original features and outcomes (however not the protected attributes) is introduced for which the utility of the transformed data is maximised under fairness constraints.

Auto-encoding: (Louizos et al. 2016) introduce a method that learns a latent representation of the original data in a generative model so that the representation is invariant with respect to the protected attributes. A regularisation technique then further removes correlations on the sensitive attributes in the distribution that generates the transformed data set (Calders, Kamiran, and Pechenizkiy 2009; Faisal Kamiran and Calders 2012; Calmon et al. 2017; Zemel et al. 2013; Feldman et al. 2015; Louizos et al. 2016).

Counterfactual Fairness

Prediction via non-descendants in causal graph: The intervention method introduced in (Kusner et al. 2017) learns the feature selection for a model by choosing non-descendants within a causal graph. The intervention achieves that the distribution over predictions for an individual and the distribution over prediction for that individual, if it had been given a different protected attribute in a causal sense, both coincide. Given the fair feature selection a model can then be trained in the usual way.

03.2.2 In-processing

In-processing methods are applied during the training of the model. They typically involve modifying a model's architecture, or modifying the training objective (e.g. by adding a fairness constraint). They seek to ensure that the resulting trained model does not exhibit unfairness.

In-processing approaches are typically able to achieve high performance as the result of co-optimisation of performance and fairness which prevents information bottlenecks.

On the other hand, most such interventions require modification and retraining of the model, which can be a non-trivial undertaking. Furthermore, unlike pre-processing approaches, each model must be intervened on separately, which requires time, human effort and usually computational resources.

Examples of in-processing algorithms

Any of the fairness notions displayed in Figure 4 above can be achieved by an appropriate in-processing method. We present the different in-processing techniques according to the fairness notions they address.

Demographic Parity

Constrained optimisation: In the context of predicting recidivism for criminal defendants (Larson et al. 2016; Dieterich, Mendoza, and Brennan 2016), (Corbett-Davies et al. 2017) formulate algorithmic fairness as maximising a utility function under constraints on group-specific risk thresholds based on which decisions whether to detain a defendant or not are taken. Their formulation addresses Demographic Parity, Conditional Demographic Parity, as well as predictive equality (form of Equalised Odds).

Further, (Agarwal et al. 2018) define fair classification as the minimisation of the prediction error under a general form of linear constraint, which addresses Demographic Parity and Equalised Odds as special cases. The optimisation is solved by a sequence of cost-sensitive classification problems. In (Zafar, Valera, Rodriguez, et al. 2017), a loss function associated with a decision-boundary based classifier is minimised under constraints on the covariance between sensitive attributes and the distance between features and the classifier's decision boundary in order to achieve Demographic Parity.

Regularisation: (Kamishima et al. 2012) develop an approach to achieve Demographic Parity in a logistic regression classifier which is based on maximising the sum between utility expressed via probabilities of classifying data points correctly given their features and further a regularisation term that incorporates the level of unfairness in the classifier.

Naive Bayes/Balance models for each group: (Calders and Verwer 2010) propose a method based on naive Bayes that trains a classifier on each protected group separately. The overall prediction of a data point is given by one of the previously trained classifiers depending on its protected attribute.

Naive Bayes/Training via modified labels: Another approach also presented in (Calders and Verwer 2010) introduces latent variables into the Bayesian model which represent fair labels based on which the model performance is maximised.

Tree-based splits adaptation: (F. Kamiran, Calders, and Pechenizkiy 2010) develop a fair decision tree whose splitting criterion incorporates, in addition to its contribution to the overall model performance, the degree of discrimination it creates with regards to Demographic Parity.

Adversarial debiasing: (Zhang et al. 2018) introduce a fairness intervention based on adversarial learning. Fairness is achieved by training a model to fool a discriminator which aims at identifying the protected attributes from the model prediction. A further adversarial fairness intervention which is based on modifying the architecture of an existing neural network model by adding a discriminator at the top layer is introduced in (Adel et al. 2019).

Conditional Demographic Parity

Constrained optimisation: See (Corbett-Davies et al. 2017) under **Demographic Parity.**

Adversarial debiasing: The adversarial methods in (Zhang et al. 2018) and (Adel et al. 2019) can be used to mitigate for conditional demographic parity by additionally passing the legitimate risk factors to the discriminator.

Equalised Odds

Constrained optimisation: For (Corbett-Davies et al. 2017) and (Agarwal et al. 2018), see **Demographic Parity**. (Zafar, Valera, Gomez Rodriguez, et al. 2017) extend their previous approach from (Zafar, Valera, Rodriguez, et al. 2017), in which the fairness constraint expressed via the classifier's decision boundary allows for bounding the difference in misclassification rates for the different protected groups. Further, (Woodworth et al. 2017) formulate fair classification as the sequential use of an in-processing and a post-processing method. First, the classifier's loss is minimised subject to a relaxed notion of Equalised Odds allowing for computational tractability. Second, a post-hoc fairness correction similar to (Hardt, Price, and Srebro 2016) is applied.

Adversarial debiasing: The adversarial methods in (Zhang et al. 2018) and (Adel et al. 2019) can be used to mitigate for equalised odds by additionally passing the label to the discriminator.

Calibration

Unconstrained optimisation: Removing the group-specific risk threshold constraints in (Corbett-Davies et al. 2017) (see **Demographic Parity**) leads to a classifier that satisfies calibration.

03.2.3 Post-processing

Post-processing algorithms modify the model's outputs, seeking to correct unfairness in the model by applying a post-hoc modification to its decisions.

Post-processing interventions are typically very flexible, often only requiring scores or decisions from the original model as well as corresponding protected attributes or labels. As a result, post-processing approaches are usually fully model-agnostic, and do not require models to be modified or retrained. Post-processing approaches often perform well, exceeding the performance of pre-processing approaches and rivaling the performance of in-processing interventions in some cases.

That said, unlike in-processing approaches, they are by nature simpler and do not typically assess the root cause of the unfairness, hence only effectively treating its symptoms not the original cause. Furthemore, the accuracy / fairness threshold cannot be pre-set (*in contrast to e.g. in-processing constrained optimisation approaches)*.

Examples of post-processing algorithms

Demographic Parity

Naive Bayes / Modification of model probabilities: (Calders and Verwer 2010) propose an approach that modifies the classification probabilities of a naive Bayes estimator to achieve demographic parity, without significantly altering the number of assigned positive labels.

Label modification: (F. Kamiran, Karim, and Zhang 2012) introduce an intervention method which relabels data points for which the model prediction is not decisive. Such data points are assigned the positive label if they belong to the unprivileged group and the negative label otherwise. A similar approach is proposed in (Lohia et al. 2019), however instead of selecting data points with large model uncertainty for relabelling, data points are selected which are likely to suffer from individual bias.

Tree-based leaves relabelling: In (F. Kamiran, Calders, and Pechenizkiy 2010), the authors present a postprocessing algorithm specific to decision trees that allows the user to relabel leaves of a decision tree to achieve Demographic Parity.

This algorithm is specific to decision trees. It may generalise to ensembles of trees (e.g. random forests) but the authors do not investigate this.

Equalised Odds

Decision threshold modification (ROC curve)/ constrained optimisation: (Hardt, Price, and Srebro 2016) introduce the notion of Equalised Odds and equality of opportunity. If only the decisions are available, they randomly choose either the original decision or fixed outcome in a way that ensures agreement across both protected groups. If a score function is available, they choose between two carefully chosen thresholds with a particular probability to ensure agreement of true and false positive rates.

This method can be proved to be the optimal postprocessing algorithm for Equalised Odds, however the randomness introduced into decision making – which in particular could mean two identical individuals receive different outcomes – might clash with intuitive notions of fairness.

Calibration

Information Withholding: (Pleiss et al. 2017) introduce a method for achieving a relaxed version of Equalised Odds, while maintaining Calibration by withholding information. In particular a proportion of the advantaged group is predicted according to the base rate without considering the model inputs. This preserves Calibration but allows us to bring the error rates for the two classes closer together.

This method is attractive in that it achieves one notion of fairness and approximately achieves another. However, similarly to the intervention of Hardt et al., it introduces randomness into decision making that might not be compatible with individual notions of fairness. Furthermore, the method requires as input calibrated classifiers, it does not offer a way to achieve Calibration, only to preserve it.

03.2.4 Interdependencies between pre-, in- and post-processing

One mitigation technique can be understood as in more than one category: It is noteworthy that some intervention approaches have features which span time categories. For example, Zemel et al. is in most literature considered pre-processing, however due to the fact that the new data representation is basically due to training a new model, new predictions are automatically obtained as well, which relates more to in-processing.

Further, interventions at different timings can be combined: for example it would be possible to both pre-process and post-process in most cases. This might be attractive if, for example, it is important to achieve a baseline level of fairness in a model via pre-processing, but then retrospectively post-process for particularly sensitive or important decisions.

03.3 Existing tools to mitigate bias

There has been an explosion in algorithmic fairness tools coming onto the market in the last one to three years, that are either commercial or open source, that help with measuring and mitigating unfairness in machine learning models.

We set out the key open-source tools available today below, and map them to our framework:

Tool	Strategies implemented	Implementation
IBM AI Fairness 360[2]	**Pre-processing** **Optimal clustering / constrained optimisation:** - (Zemel et al. 2013) - (Calmon et al. 2017) **Feature modification:** - (Feldman et al. 2015) **Data reweighting:** - (Faisal Kamiran and Calders 2012) **In-processing** **Adversarial debiasing:** - (Zhang et al. 2018) **Regularisation**: - (Kamishima et al. 2012) **Post-processing** **Information Withholding**: - (Pleiss et al. 2017) **Decision threshold modification (ROC curve)/ constrained optimisation:** - (Hardt, Price, and Srebro 2016; Woodworth et al. 2017)	Python, R

[2] https://github.com/IBM/AIF360

	Label modification: - (F. Kamiran, Karim, and Zhang 2012)	
FairLearn[3]	**Post-processing** **Decision threshold modification (ROC curve)/ constrained optimisation:** - (Hardt, Price, and Srebro 2016) **In-processing** **Constrained optimisation**: - *(Agarwal et al. 2018)*	Python
Algorithmic Fairness[4] *- BlackBoxAuditing* *- Fairness-comparison*	**Pre-processing** **Feature modification:** - (Feldman et al. 2015) **In-processing** **Regularisation**: - (Kamishima et al. 2012) **Naive Bayes/Balance models for each group:** - (Calders and Verwer 2010) *(slight alteration to original algorithm as latter may fail to stop)* **Constrained optimisation**: - *(Zafar, Valera, Rodriguez, et al. 2017)* - (Zafar, Valera, Gomez Rodriguez, et al. 2017)	Python
Fairclassification[5]	**In-processing** **Constrained optimisation**: - *(Zafar, Valera, Rodriguez, et al. 2017)* **-** (Zafar, Valera, Gomez Rodriguez, et al. 2017) Python	Python
Fairness-aware Data Mining[6]	**In-processing** **Regularisation**: - (Kamishima et al. 2012) **Naive Bayes/Balance models for each group:** - (Calders and Verwer 2010) (slight alteration to original algorithm as latter may fail to stop)	Python
FairSight[7]	**In-processing** **Prediction via non-descendants in causal graph:** - (Kusner et al. 2017) **Post-processing** **Reranking:** - (Zehlike et al. 2017) *(Allows for other mitigation techniques to be plugged in)*	Standalone Application

[3] https://github.com/fairlearn
[4] https://github.com/algofairness
[5] https://github.com/mbilalzafar/fair-classification
[6] https://github.com/tkamishima/kamfadm
[7] https://github.com/ayong8/FairSight

04 Limitations of Algorithmic Fairness Approaches

In sections 2 and 3, we introduced measures of fairness and mitigation strategies and discussed some of the individual limitations of each. In section 4, we collect and discuss some of the wider criticism and problems with algorithmic fairness that are not specific to any particular notion of fairness or any particular mitigation strategy.

At the end of this section, we then consider how they can still be useful in practical application, in light of these limitations.

04.1 Accuracy-fairness trade-off

The accuracy of a model and its fairness are in general in tension as briefly addressed in Section 02.5.1. That means, enforcing a mathematical notion of fairness generally decreases the performance of the model, which itself is a valid objective. In the context of constrained optimisation (Agarwal et al. 2018; Corbett-Davies et al. 2017; Zafar, Valera, Rodriguez, et al. 2017), the trade-off is intuitive: instead of optimising purely for performance, constraints must satisfy the decrease in the space of possible solutions. Hence, the resulting solution has the same or, typically, worse accuracy than the unconstrained solution. As a consequence, practitioners need to understand the severity of this trade-off for the specific task at hand in order to make informed decisions for modeling or in policy (Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth 2018).

04.1.1 Cases of undesired trade-off choices

In critical applications, such as present in healthcare and justice, the choice between fairness and accuracy can seem impractical or unethical (Corbett-Davies et al. 2017; Chen, Johansson, and Sontag 2018). It is advisable that, where possible, bias is mitigated while avoiding the necessity of this trade-off (Chen, Johansson, and Sontag 2018). This applies when bias exists on the basis of differently well represented groups in the training data. A subsequently trained classifier tends to expose bias toward the majority group and rather poor accuracy in the minority group (He & Garcia 2009), which itself can be a procedural form of bias.

For example, facial recognition algorithms deal with such problems (C. Huang et al. 2019), due to prevailing imbalances in common benchmark data (G. B. Huang et al. 2008; Kemelmacher-Shlizerman et al. 2016). Bias due to underrepresentation can be mitigated by the collection of additional data and an appropriate model choice (Chen, Johansson, and Sontag 2018). However, in most situations, bias is not due to group imbalance in the data only, but the consequence of historical, human and other sources of bias which cannot be mitigated by tailored data collection alone.

04.1.2 'Biased accuracy'

There can be instances where the source of bias in a classifier's decisions is not due to underlying factors, but rather due to bias in the underlying data, or more precisely, in the data labels[8] it was trained on. As a result, in such circumstances, the reported accuracy of a biased classifier is also biased. The true accuracy is generally unknown. Hence, enforcing fairness constraints onto decisions need to take this effect into account in order to produce the desired equalising effects.

That said, in such circumstances, the usual accuracy/fairness trade-off is potentially reversed: i.e accuracy and fairness being positively associated. It is possible and indeed likely that there will be ways to improve fairness which simultaneously increase the model's *true* accuracy (i.e. the actual accuracy of the model, unmeasured, not the reported measured accuracy).

[8] Data labels refer to the specific features applied to individual data points - e.g. applying a label in a recruitment data set to each individual whether they have a university degree or not.

04.2 Incompatibility of fairness notions

As has been discussed in Section 02.5.2, some of the key algorithmic definitions of fairness are typically mathematically mutually incompatible. This makes generalised practical application very difficult for policymakers and practitioners: definitions will be context-specific, but often seemingly contradictory.

For example, if labels and protected attributes are marginally associated, such as recidivism and race (Larson et al. 2016), then generally equality of opportunity (generalised form of Equalised Odds) and Calibration cannot be reconciled (Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2017). To see this, note that if a model satisfies Calibration, then in each risk category, the proportion of defendants who reoffend is the same, regardless of race. However, the only way of achieving this if the recidivism rate is higher for one race, is if more individuals from that race are predicted to be high-risk. Consequently, this means that the model will make more false positives for that race than others, meaning Equalised Odds cannot be satisfied. Similarly, Calibration or Equalised Odds, and Demographic Parity or Conditional Demographic Parity, cannot be satisfied simultaneously in practice, and satisfying one typically degrades the other (Corbett-Davies et al. 2017). To continue with the above example, if a recidivism model satisfies Demographic Parity, then the chance a defendant ends up in any particular risk category is the same, regardless of their race. If one race has a higher recidivism rate than the others, that means the model must make more false negatives for that race in order to maintain Demographic Parity, which as a result means Equalised Odds cannot be satisfied. Similar arguments apply for other notions of fairness, see ((Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2017) for additional details.

The COMPAS model, a well-known example in the US to predict offender recidivism, brings this point to life. For a given risk score in the criminal recidivism model (Dieterich, Mendoza, and Brennan 2016), the proportion of defendants who reoffend is roughly the same independently of the protected attribute (Calibration). If that weren't the case, the model would likely disagree with the equal protection clause, i.e., it would apply a different standard between black and white defendants (a risk score of 8 for a white person would mean something different for a black person). Propublica's criticism was that black defendants that did not reoffend are approximately twice as likely to be given medium/high risk. However, ensuring equal risk scores among defendants who didn't (re-)offend (Equalised Odds) would result in losing Calibration at least to some degree. That means, one can't be fair in both manners (Corbett-Davies et al. 2017).

Although different notions of fairness are generally incompatible and lead to trade-off choices for practitioners, there are at least two exceptions:

- The information withholding post-processing method introduced (Pleiss et al. 2017), achieves both Calibration and a relaxed notion of Equalised Odds. The algorithm starts with a model that achieves Calibration, and considers calibration-preserving modifications to the model. It then chooses the calibration-preserving modification that most improves a relaxed notion of Equalised Odds. It is necessary to relax the definition, as Calibration and the standard definition of Equalised Odds are incompatible (see above).
- There are multiple approaches aiming on achieving both Individual Fairness and Demographic Parity: the clustering pre-processing approach from (Zemel et al. 2013) aims at achieving Individual Fairness combined with notions of Group fairness. Similarly, by selecting samples that likely suffer from individual bias and changing their predicted label, (Lohia et al. 2019) achieve Individual Fairness and Demographic Parity simultaneously. Finally, Demographic Parity can be achieved while treating similar individuals as similarly as possible as proposed in (Dwork et al. 2012).

04.3 Non-holistic nature of algorithmic fairness definitions

Closely tied to their mutual incompatibility, another key limitation of algorithmic fairness definitions is that they are by nature not holistic. No one definition will be right in all circumstances.

Furthermore, algorithmic definitions struggle to encompass or codify wider more qualitative characteristics of fairness, which as discussed early on, do not lend themselves well to mathematical precision but are nonetheless extremely important.

04.4 Lack of techniques to choose between fairness notions

As above, we can conclude that the practitioner is forced, to some extent, to choose / mathematically trade-off between different fairness definitions owing to their incompatibility. Therefore, techniques are needed to help practitioners carry out that weighing up.
However, as stands, we conclude that these techniques are lacking in the Machine Learning literature – this is an area which hopefully will develop further, as algorithmic fairness intervention becomes a more applied discipline.

By far, most of the literature on fairness mitigation approaches follow the principle of first choosing a single suitable notion of fairness, if not already determined by the method itself, and then achieving the specified fairness notion by applying the intervention. There seems to be a gap in the literature regarding the development of intervention methods seeking trade-offs between different notions of fairness. One exception is (Kleinberg, Mullainathan, and Raghavan 2016).

Given how important it is for practitioners to have a way of selecting between fairness methods in practice, in the accompanying Implementation Handbook, we set out a practical way of doing this. However, we hope that this can be iterated and improved upon as the field evolves.

04.5 Non-holistic nature of algorithmic fairness definitions

Closely tied to their mutual incompatibility, another key limitation of algorithmic fairness definitions is that they are by nature not holistic. No one definition will be right in all circumstances.

Furthermore, algorithmic definitions struggle to encompass or codify wider more qualitative characteristics of fairness, which as discussed early on, do not lend themselves well to mathematical precision but are nonetheless extremely important.

04.6 Limits of responsibility for correcting systemic unfairness

Unfairness in algorithms generally occurs due to the perpetuation or exacerbation of a variety of existing biases in the underlying data set on which an algorithm is trained, including historical, measurement and population bias (Mehrabi et al. 2019; Olteanu et al. 2019). Mitigating unfair decision making must be addressed in order to create a sustainable and equal society. However, in addition to identifying the right steps towards long-term fair decision making, a core issue is the question about whom should be viewed as responsible for achieving fairness, and what are the consequences on a legal or professional level of arguably unfair decision making.

The question arises to what degree commercial organisations can be considered responsible for the mitigation of bias across society. In many situations, algorithmic decision making simply perpetuates existing biases, which are based on historical artifacts, whose roots lie in the past and are not due to any current company procedures or cultures.

04.7 Sources of unfairness and causality

The algorithmic definitions and intervention approaches discussed in sections 02 and 03 generally do not help with identifying *sources or causes* of unfairness, only observing symptoms in practice.

Defining fairness based on causal inference (Kusner et al. 2017; Kilbertus et al. 2018) has been picked up in the literature only to a limited extent (Garg et al. 2019; Chiappa and Gillam 2018) due to the difficulty of knowing the underlying causal graph, i.e, validating the causal implications assumed for the task at hand. In order to avoid this problem, (Russell et al. 2017) propose a method which achieves approximately fair predictions with respect to multiple possible causal models at once.

04.8 Long-term impact of fairness mitigation strategies

The definitions of fairness introduced in section 02 are "static", in the sense that we generally measure them on a snapshot of the population at a particular moment in time. We should expect however that making an intervention into model predictions in order to impose fairness constraints will cause that population to change over time. In the financial and recruiting sectors under consideration in this report, for example, allocation of financial support and jobs will alter the distribution of wealth and opportunity within the population.

Failing to account for these dynamics risks leading to interventions that are actively counter-productive, indeed there are cases where a supposedly fair intervention leads to greater unfairness than without the intervention (Kusner et al. 2017; Liu et al. 2018).

Algorithm designers must also consider the possible effects of strategic manipulation, i.e. individuals take action to change their attributes in order to achieve a more favourable outcome. For example, hypothetically a Natural Language Processing-based algorithm used for screening job applicants may evaluate particular words or phrases more favourably when included in a CV or application form. As a result, if known, individuals could be incentivised to use these words to achieve a more favourable evaluation, even if it does change their underlying ability to do the job. However, the cost of manipulation may typically be higher for the disadvantaged group. In the case of our example, knowledge about how to 'game' the algorithm may be only obtainable by better networked groups. The differing costs of manipulation can thus result in disparities between protected groups being exaggerated (Hu et al. 2019).

On causal modelling

One may argue that in order to correct bias, one needs to understand the underlying sources. Hence, part of the problem is that most debiasing approaches do not capture how their intervention propagates through and influences the world it is applied to in terms of *causal* relationships (Kusner et al. 2017).

The causal approaches, introduced in section 02 and set out in more detail in Annex A, seek to address this precise issue. However, they are often impractical as they require the knowledge of a *causal graph*[9] for the task at hand, which is typically not fully known or too complicated to capture in many real-world situations. Seeking to understand this causal graph can itself introduce further bias, as often these relationships will be less well understood for minority groups with less data. Authors (Russell et al. 2017) try to address the problem of exact causal specification by a framework that integrates multiple competing causal graphs at once, and that subsequently trains a model that is approximately fair across all of them.

In order to make causal models more tractable, we often make even more restrictive assumptions that aren't always realistic. There are some approaches to getting around having to make such assumptions, but they are themselves very complicated.

[9] In simple terms, a causal graph maps the dynamics of how an input feature drives the outcome of interest.

On reinforcement learning

A static view of fairness neglects that most decisions in the real world are actually taken in sequence. For example, a company hiring new employees over time will factor the current employee pool when making future recruitment decisions. Besides the already discussed causal approach, consequences of these individual decisions which generally influence the dynamics of the world can also be addressed within an online learning setup (Jabbari et al. 2017). The ideal decision policy is the one that promotes fairness in the long run.

04.9 Unfairness when fair algorithms are combined

It is generally not well understood in which case a system made up of several machine learning algorithms, which are fair individually, satisfies fairness guarantees. For instance, (Dwork and Ilvento 2018a, [b] 2018) show that both Group and Individual Fairness can degrade when two fair predictors for two original tasks are composed into one task by making these tasks compete with each other. This means that different tasks which are, each separately considered, addressed in a fair way may finally be addressed unfairly as part of a larger pipeline of independent tasks, which might lead to significant social implications.

Following (Dwork and Ilvento 2018a), let us for example assume that there are two different advertisers on one website competing for users who visit the site, say, one for goods purchases and the other for a job. It is further assumed that each advertiser, considered in isolation, bids in an advertising auction on potential customers fairly. However, if both advertisers compete with each other on the same website, the resulting bidding algorithm is unfair. Intuitively, this can be understood as follows. Two individuals with a similar job qualification may effectively have different chances to see the job advertisement depending on the desirability for them being advertised purchase goods, that is, if one of them is desirable to be advertised purchasing goods and therefore be claimed by the purchase good advertiser, then that individual will not see the job advertisement.

In such cases a fairness intervention approach would then be needed at the level at which those two separate procedures co-exist, introducing complexity and concepts of joint responsibility/liability between independent actors, which does not lend itself well to regulatory policy/enforcement.

04.10 Usefulness of algorithmic fairness approaches

In light of the above limitations, it may be natural to conclude that algorithmic fairness approaches have limited usefulness for practical application, however, we do not believe that this is the case.

Rather, we believe that algorithmic fairness approaches will be *essential* to facilitate widespread adoption of algorithmic decision making, and hence ultimately allowing AI to achieve its potential to positively transform society.

In particular, the usefulness of algorithmic fairness approaches stems from their precision:

1. **Quantification:** algorithmic approaches allow precise quantification of fairness against different notions, often ascribed a single number. This takes a previously ill-defined quality, and turns it into quantitative measure which, for example, can be included in RoI calculations or impact assessments. While there are potential downsides (e.g. the single number may not capture all types of unfairness), this is potentially extremely important to ensure that fairness is given due weight as a goal in its own right.
2. **Definition of terms**: algorithmic approaches force the policymaker or decision maker to be precise about different terms or notions of fairness: providing clearer structure to an otherwise abstract debate, and forcing them to self-inspect their institutional goals accordingly.
3. **Trade-off with accuracy**: algorithmic approaches also allow a precise quantification of the fairness / accuracy trade-off within the model itself – again, potentially elevating the importance of achieving fairness within overall decision making.

What is needed is to find a way to harness this quantitative and definitional power, and combine it with wider but equally legitimate notions of fairness into an overall generalisable approach, which can be applied in operational contexts by practitioners and other decision makers. We set out a proposed method for doing this in the accompanying Implementation Handbook, for further iteration and debate.

05 Algorithmic Bias in the Financial Sector

Algorithmic decision making is widespread in the financial services sector and its use will only continue to increase in the future. As a result, there is significant potential for algorithmic bias to have a detrimental effect not only on individuals, but also on businesses. This section is based on feedback received from fintech, banking and other organisations providing financial services, and has been complemented by additional desk research. Our findings also build on CDEI's industry review.

05.1 Overview of revealed practices

The use of machine learning algorithms in financial services has grown rapidly in recent years. These tools and techniques are being used across a wide range of areas, including personalised finance, marketing optimisation, loan applications, trading, risk modelling, fraud analysis and robo-advice.

Findings from the industry review suggest that fintech companies have more flexibility and space to deploy machine learning algorithms in innovative ways. Feedback from smaller financial firms revealed that supervised learning has been in use for at least the past five years. As a result, fintech companies are generally more advanced than traditional banks in terms of using algorithmic decision making.

Although established banking institutions also make use of these techniques, industry experts explained that it is still a conservative sector, considerably constrained by strict regulatory requirements. Established banking institutions collect vast amounts of customer data, thus are more risk-averse than the fintech sector. Further to this, legacy core banking systems are out of date – legacy technology means that banks have limited abilities to interface with other systems, thus restricting a bank's ability to rapidly deliver new tools and techniques. That being said, stakeholders confirmed that banks generally rely on simple AI tools and/or classical statistical techniques to automate decision making.

While clear governance frameworks to audit for bias have been identified across the industry (Section 05.3), it is clear that companies place more emphasis on detecting and mitigating bias in the pre-processing stages (i.e. by carefully selecting variables and involving human judgement in the loop). The use of governance and scrutinising fairness tends to diminish at the stage of testing models and assessing the outcomes or impacts of an algorithm.

It is also important to highlight that some companies are arguably insufficiently concerned about potential bias, because their models factor out protected characteristics, particularly in a business-to-consumer context (B2C); or because they do not collect or use protected characteristics in their algorithms, especially in a business-to-business (B2B) context. Our findings suggest that there is a difference in the way corporate and individual discrimination are monitored. Irrespective of company size and business model, the use of protected characteristics is generally avoided in the financial sector – this suggests that Fairness Through Unawareness dominates the financial sector.

05.2 Sector specific issues

05.2.1 Differences between B2B and B2C businesses

B2B and B2C financial services both favour Fairness Through Unawareness, however the context in which they detect and mitigate bias in algorithmic decision making largely differs due to operational differences.

The B2B fintech companies we consulted scrape publicly available data (e.g. Companies House), and to some extent, collect corporate data internally (e.g. bank transaction details, credit reports). These companies reported being less concerned about bias entering their models because protected characteristics are not *directly* relevant to their products and services. Interestingly, stakeholders in a B2B context widely reported that they were operating in a 'grey area': algorithmic driven processes might not consider discrimination, even though it will be considered by humans.

On the other hand, large banks or financial services, that process large amounts of personal data, are more concerned about bias entering their models. Broadly speaking, customer-facing companies are faced with stricter regulations, since algorithmic bias has the potential to create reputational damage or perpetuate systemic discrimination.

One unexplored area of overlap is for B2B models that are making decisions about sole traders/ other small companies (such as whether to extend a loan) and are to an extent relying on the credit history/ personal information of the business owner, rather than the business itself. These models are open to issues of bias based on a business owner's personal information, but the outcome will impact a business rather than an individual.

05.3 Current audit approaches

The majority of companies we consulted have governance frameworks in place to ensure that their decision making is not biased. Globally, companies highlighted the lack of guidance and transparency on auditing approaches. This pushes companies to develop their own interpretations and approaches to audit for bias, which involves a combination of human oversight and technical tools.

05.3.1 Use of governance

The research identified three stages where bias detection and mitigation takes place. Based on the interview feedback, companies follow to some extent the following typology - which broadly maps onto pre-, in-, and post-processing:

1. **Data:** analysis of the data before a model is built
2. **Model:** analysis of the model predictions on test data
3. **Impact:** measure of the long-term impact of imposing fairness

According to interview feedback, all companies analyse the data they have; discuss the variables that should be included and excluded from the model; and discuss potential risks of bias entering the system. It is at this stage that companies integrate mitigation techniques in their models, such as excluding protected characteristics (Fairness Through Unawareness) or integrating calibration techniques so bias monitoring becomes a natural part of the system. Generally, the inputs and outputs of the model undergo a sign-off process overseen by a human. This approach was common in the data-savvy fintech firms that we spoke to.

A smaller proportion of companies proceed in analysing model predictions on test data, such as representative synthetic data or anonymised public data. We were not able to identify a commonly-used approach across the industry and companies noted that they approach issues on a case-by-case basis. For example, a fintech company monitors bias by checking whether new features affect the model calibration. Another company shared that they drop problematic models without trying to fix the bias. In other cases, companies will rely on common sense or human judgement to monitor the direction and relationships generated by machine learning. One stakeholder recommended companies should systematically challenge their models with counter-factual synthetic data to test model predictions.

Measuring the long-term impact of imposing fairness is difficult to conceptualise because it depends on the context. The feedback we received was fragmented: in some cases, companies have strict explainability guidelines, in other cases, companies are not able to interpret the drivers of their decisions. In some cases, a human gets involved for quality assurance, particularly for the latter.

05.3.2 In-house vs. external auditing tools

Our interviews did not reveal a commonly used approach in the industry to audit bias – companies use a mix of in-house and external auditing tools. However, there is a preference for adapting open-source tools internally. For example, a fintech company used the IBM toolbox (AI 360), but built a similar in-house tool to tailor it for their needs. Amongst the companies we consulted, none had developed in-house tools from scratch. There is therefore an appetite for open-source auditing tools.

In banking, companies also refer to the Three Lines of Defence for effective risk management and control. Although this is not specific to AI and algorithmic decision making, interview feedback revealed that banks adapt and apply this approach to algorithmic risks. A high-level overview of the Three Lines of Defence is presented below (The Institute of Internal Auditors 2013):

1. In the first line of defence, management control is first in line to deal with risk management. 2. The various risk control and compliance oversight functions established by management are in the second line of defence. This will include staff acting in compliance with GDPR and other regulatory requirements.
3. In the third line of defence, an internal audit provides independent assurance.

The lines of defence also play an important role within a company's wider governance

framework. 05.4 Case studies

Case study 1: Utilising corporate data

- A B2B fintech company is specialised in solving problems in the sphere of international trade and cross-border financial activity.
- They use algorithms to credit grants to corporates.
- No personal data is used in this context. The models are trained from the aggregation of publicly available corporate data, however the final decision (whether to grant a loan or not) is made by a human.
- Even though they do not use personal data, their algorithms are built to differentiate for companies that are most likely to default (i.e. based on historical data, certain companies are more likely to default than others in certain geographical regions).
- In this case, differentiation is based on performance-relevance: models have learnt to differentiate against certain variables, such as geography or specific types of industries. The fintech company wants to find companies with similar profiles.
- The expected probability of defaulting is common across all geographies – therefore, geography is taken as an input, but it is not a single qualifier/disqualifier for business loans. It is weighed against many other financial and non-financial corporate attributes to deliver predictions.
- The company considers that the bias they introduce is not unethical because it is corporate data. ● This case study reveals the difference in considering bias between firms that use corporate data and customer-facing companies that use personal data. Indeed, there is a difference in the way corporate and individual-level bias or differential outcomes are monitored: algorithmic-driven processes might not consider differentiation in a B2B scenario (although it will be considered by humans).

Case study 2:
Using human scrutiny to supplement Fairness Through Unawareness

- Faculty held an interview with one of the leading fintech companies in London, which provides unsecured personal loans.
- The company has the ability to be flexible and technically innovative. Supervised ML is applied across the company, they consider themselves to be ahead of banks in terms of ML use. The two typical applications of ML are: (1) predicting whether people are able to repay personal loans; (2) Fraud detection.
- Although the fintech company regularly update their models and scrutinise internal auditing (i.e. regulation/compliance), they are less concerned with traditional human bias because their supervised ML models are trained on facts (e.g. whether the customer has defaulted), rather than on human decisions (e.g. whether they were declined by underwriters). In line with 'treating customers fairly' requirements, the firm applies Fairness Through Unawareness (i.e. they explicitly prohibit the input of any protected characteristics into the decision models).
- One important tool is to prevent bias is ongoing monitoring: monitoring processes at the fintech company were mainly designed to detect model inefficiency in terms of predicting credit risk. This fits the notion of sufficiency when it is applied on protected characteristics. The company explained that the same setup can be used to investigate separation, however, they chose to focus on sufficiency, as it is a regulatory requirement on financial services to have an accurate estimate of credit risk.
- The company also relies on human judgement. They make sure that protected characteristics aren't included and interpret the direction in which their models are going. They found that understanding the latter helps mitigate bias – if a variable doesn't fit the pattern, they reject it or transform it. Humans ensure the direction of the models is sound.
- This case study illustrates a company that adopts Fairness Through Unawareness in the model build stage. The fintech firm recognises that this does not solve all problems, therefore human analysts and auditors also actively scrutinise the models in the model validation and monitoring stages.

06 Algorithmic Bias in the Recruitment Sector

The use of algorithms to support hiring processes has started to increase in recent years, and is likely to become pervasive before long. Robotic process automation (RPA), and other methods to automate non-decision-making recruitment processes - such as candidate reminders and interview scheduling - are now widespread. A recent report has estimated that over 98% of Fortune 500 companies use 'Applicant Tracking Systems' of some kind in the hiring process (Sánchez-Monedero, Dencik, and Edwards 2020), a trend that we'd expect to see in similar surveys of large UK firms.

Building on that, machine learning and other AI approaches are starting to be deployed in recruitment decision making processes - albeit primarily as intelligent decision-support. They have been deployed in the sourcing, engagement, selection and onboarding stages of hiring decision making. At the same time, there has been a growing number of consultancies and start-ups in the UK offering machine learning based hiring solutions and tools. Industry leaders expect this trend to continue, and it could accelerate if a period of substantial unemployment significantly increases applicant/job ratios, making the efficiency gains from AI-assisted recruitment essential rather than desirable.

The prevalence of bias in recruitment processes is also highly topical. The history of human bias in recruitment has been extensively documented - in its impact on individual fairness and in the creation of homogenous workforces that lack the benefits of diverse backgrounds, skills and ideas. Some vendors propose machine learning tools as a solution to these historical biases. Introducing AI can help humanise the recruitment process: AI approaches can be applied to deliver complex tasks, personalise processes at a larger scale and reduce bias in the selection of candidates, while increasing the support to humans (Nordmark 2020). Indeed, by outsourcing the initial screening of applicants, human prejudice is detached from the first screening. However, there is also growing awareness of the risk of bias within AI tools: bias can enter through the data that is used to train the model or by the constraints set up by humans when designing the models.

This section draws on feedback from recruiters and recruitment consultancies, as well as recommended readings and desk research, to examine these issues of bias within the recruitment sector. We build on the CDEI's industry review and recent analysis from Institute for the Future of Work (IFOW).

06.1 Drivers behind use of recruitment algorithms

The use of algorithms in different stages of the recruitment process has grown rapidly in recent years. In part just as in other sectors, this is likely to be a function of the availability of data and technology that can be used to drive decision making. With many time-consuming, repetitive and data-driven tasks, the recruiting sector is an area with considerable potential for an AI-revolution.

In addition, our industry review has revealed some sector specific factors that have led to increased uptake in the field of recruitment. Namely, a stakeholder noted that the biggest shift will be when AI will be used to aid recruitment decision making, which could be at different points along the recruitment cycle, such as automating the process of targeted ads or CV sifting, among other uses. It was also noted that large Managed Service Providers (MSP) are investing the most in the adoption of AI, since they have scale and vast amounts of data suitable for training algorithms.

Our review found that large national and international firms were the earliest and most significant users of algorithmic processes to aid recruitment. These firms possess large graduate schemes with thousands of applicants - equipping them with the need to sift through large numbers of applications as well as growing repositories of data. This was also true for government departments and other public sector organisations.

We might expect to see the use of algorithms and AI tools accelerate if there is to be a period of high unemployment. Industry leaders have already seen substantial increases in average applicants-per-role ratios during the Covid-19 pandemic, and if this persists the large recruiters are likely to have to rethink business models and how to drive efficiencies. This might see recruitment firms having to rely on algorithms more for initial screening, and/or use AI-based decision support to help a human reviewer to review tens or hundreds of CVs or job applications more quickly.

06.2 Overview of revealed practices

06.2.1 Outsourcing of tools and approaches

Our review found that of the firms deploying algorithms to support hiring, most bought in systems and tools from established consultancies and other vendors, rather than developing in-house capability. It was posited that this reflects common practice across other Human Resources functions in firms – where payroll services, contract services, internal communications systems etc are also more likely to be bought in. This contrasts with Financial Services, where we saw a mixture of in-house and outsourced solutions.

A survey of 18 algorithmically driven pre-employment assessment vendors found that they were internationally focussed, with half based in the United States, but with customers across developed economies (Raghavan et al. 2020). Most of these vendors provide off the shelf tools, with customizable segments and functions for customers, whilst a smaller number provide more bespoke services, building and adapting a tool to the needs of the hiring organisation.

Based on feedback from our industry review, some vendors may claim their tools are free from bias, however, this is unlikely certain to be the case in practice. Although this is likely to change in coming years, the concept of algorithmic bias is still considered to be relatively nascent in the recruitment industry.

Indeed, a stakeholder noted that the recruitment tech market is under-regulated, in a way which limits incentives to consider bias. There is also lack of clarity about the allocation of responsibility or liability between tech vendors and firms using the tech, when algorithms fail or demonstrate bias. For example, some firms offer regulated recruitment services in practice but are not labelling them as such, which means they are out of scope. It was also noted that the sector is not yet sophisticated enough as a buyer of tech.

06.2.2 Stages of hiring

Our review identified four stages of the hiring process where algorithms are regularly deployed:

Sourcing: support to determine which channels (e.g. jobs boards) should be used for hiring; test the most effective ways to promote jobs on the chosen channel; and how much to spend on promotion activities.

Engagement: tools to increase the engagement of prospective applicants with a job advert. This could include chat bots; using machine learning to personalise content for applicants; and identifying and removing words that might put off applicants from particular backgrounds.

Selection: tools that recommend which applicants should be shortlisted, interviewed or hired. This is the most significant stage. Tools here are varied - some machine learning models train on the data of existing staff to select which applicants resemble high performing workers, whilst others use utilise results from competency and psychometric tests. Others are used to support the candidate through the process, as an evolution of the widespread use of RPA - e.g. Natural Language Processing tools which identify and flag to candidates when their application is incomplete or fails to meet an essential criterion. In general, selection-based AI tools are still at the stage of being used for intelligent decision support, rather than for outright algorithmic decision making.

Onboarding: tools that use data to personalise the onboarding process.

06.3 Sector specific issues

06.3.1 Correcting for human bias

One salient debate that arose from the review was the purpose of tackling 'bias' in the recruitment sector. Several interviewees argued passionately that deep biases have existed in recruitment since the inception of organised and formalised hiring processes - and that these arise from inbuilt prejudices in human decision-making. For several vendors of recruitment tools, the deploying of data, algorithms and machine learning was an endeavour to correct these human biases - by making objective, dispassionate decisions that can be mathematically evaluated. This is not a defense of biased models, but a reminder that the alternative to an algorithm-led approach is not a bias-free world.

However, some firms argue that switching to machine learning is a flawed solution for avoiding bias. This is because for all 'selection' process designs described above, there are still important human decisions needed - such as subjective judgements about preferable traits that are themselves prone to bias. Further, any training data or baseline assumptions used by models are likely to be based on existing employees or past employment practices that were themselves the beneficiaries of human bias. These consultancy firms argue that it is more important to train recruiters to recognise and accommodate their own biases than to replace human recruiters with models. The response of companies that develop machine learning tools is to base their models on more objective criteria such as test score performance.

06.03.2 Dispute over fairness objective

The argument above is sometimes taken even further in the pursuit of a 'fair' workplace, to argue that hiring processes needed to be consciously biased in order to correct for past biases in the other direction. This belief in diversity - that diverse backgrounds and ideas improve organisational performance - can lead firms to seek characteristics that are correlated with groups that they feel are underrepresented in their workplace, such as women and ethnic minorities. This debate is adjacent rather than central to a discussion of algorithmic bias, though it remains connected, as attempts to monitor and mitigate model bias may cause problems for organisations seeking to use machine learning to positively discriminate.

06.4 Current audit approaches

06.4.1 Audit within outsourced solutions

The review found that vendors of machine learning tools used in recruitment all had established processes for auditing their models - both off the shelf tools and the bespoke tools they developed for clients. The most elaborate audit process had three stages: Pre-deployment checks with dummy data and or sampled real-world data to adjust a model prior to deployment; post deployment checks where anonymised data from customers was used for further adjustments and correction of over-fitting; and third-party audits conducted by academic institutions particularly focussed on identifying sources of bias. Firms used a mixture of proprietary techniques and open-source software to test their models - Audit AI was used as an example open source tool.

06.4.2 Influence of overseas regulation

It's notable that the auditing approach of vendors may be limited by overseas regulations or standards, as they seek to harmonise their auditing approach across international deployments. For example, one firm noted that in the US they are prohibited from asking for (anonymised) data containing protected characteristics from customers that could be used to test, audit and validate their tools.

Case study 3:
In-house and third-party auditing

- Faculty approached **pymetrics**, a US-based company that uses behavioural data and AI to generate fair and predictive algorithms for recruitment and talent mobility. Fairness is a core value of the company, and **pymetrics** takes a leadership role in the field of bias detection and mitigation.
- The company uses algorithms that are trained on high-performing employees at a company and then builds a profile of a company's top performers to select the best fit candidate for a job. These algorithms are then audited to remove any gender or ethnic bias.

● Decision making in recruitment looks at inputs (i.e. how do you choose data that reflects the actual qualifications of an applicant?) and outputs (i.e. which measures are accurate and fair?). ● **pymetrics** have also developed a three-step model to guide algorithmic decision making: ● What makes a person good at their job? What behavioral factors predict success? ● What measurements enable robust identification of these candidates?

● How do you link each candidate to their ideal role without introducing bias?

- **pymetrics** avoids measures that are known to be problematic, such as facial analysis, educational history, and historically biased assessments. Instead, they focus on people's aptitudes through a series of cognitive tests that have been validated as predictive of job performance.
- Interestingly, the company adopts three steps to audit bias. The steps are as follows: ● **Pre-deployment checks**: When developing a new model, they implement a robust de-biasing algorithm to separate signal (performance) from noise (bias). They then confirm the performance of their model against real-world data and make tweaks where required.
    - **Post-deployment checks: pymetrics** monitors each model after deployment, and actively improves models wherever possible. In addition, they confirm success across models through long-term meta-analysis.
    - **Third-party audit: pymetrics** invites experts from academia and practice to review their code and validate that their approach is unbiased.
- This case study illustrates a company with high auditing standards. Including third-party auditing and on-going monitoring means the company detects bias early so it can quickly mitigate it. Fairness is not only at the core of the company's culture, but they have also developed a recruitment process that completely ignores protected characteristics.

Case study 4:
Automated evaluations and other in-practice techniques

- Oleeo produces recruitment software, which is deployed to large corporate organisations and central Government.
- In this case, the recruitment platform uses algorithms for (i) sourcing; (ii) engagement; (iii) selection; and (iv) onboarding.
- The stakeholder commented that interest in understanding the technicalities, explainability or transparency of the software tool is mixed among its customers.
- During interviews and in application forms, employers commonly use competency questions (e.g. "demonstrate how you have achieved strong commercial outcomes?") to screen out applicants. Real time monitoring over a period of 6 months showed that, for some competency questions, recruiters were rejecting responses from proportionately twice as many people of black ethnicity as those of white ethnicity
- Oleeo's data sets include 1 million+ responses to competency questions. Oleeo trained a natural language algorithm to be able to score the responses on the same basis as the recruiters. Oleeo also measured the diversity impact of recruiter's decisions and the algorithm scoring. Since Oleeo found bias, the firm adjusted the way the algorithm learned from the recruiters, using a process that optimised both on its fit with the human decisions and the diversity outcome. In this way the algorithm only learned the human decision making behaviours that did not lead to bias.
- Oleeo then asked the algorithm to mark 10,000 new responses, where recruiters had rejected proportionately twice as many people of black ethnicity as those of white ethnicity. The algorithm was unbiased passing equal proportions of black and white ethnicities, resulting in a 7% improvement on overall diversity. The algorithm was also found to be more reliable than humans who in addition to unconscious bias, suffer from fatigue and stress resulting in more quality candidates: 12% of all answers that were of high quality were no longer rejected.
- The algorithm was productionised as a "virtual panel member": once the recruiter decides whether an answer is a pass or fail, the virtual panel member (i.e. the algorithm) displays how it scored the response. When the recruiter's and algorithm's opinions diverge, the recruiter is prompted to re-read the candidate's response and make a final decision.
- Oleeo provides explainability on the working mechanisms of the algorithm (both for themselves and the client). They integrate automated evaluations, the 4/5th rule (Demographic Parity) and corresponding fairness checks into their models. Oleeo also applies validation techniques to make comparisons over time before the model goes into practice.
- To monitor bias, the firm uses real-time visualisation tools (i.e. Tableau) to see what the algorithm is doing. In the report, the overall performance on diversity (4/5th rule) is a key metric. Metrics on gender, age or ethnicity are correlated with the outcome. Oleeo is therefore able to understand which variables are correlated with diversity and how this impacts decision making.
- If some variables are having an adverse impact, they are removed. Otherwise, data scientists will use an optimisation process, where they change the parameters until the bias is removed, typically boosting diversity by 5% to 20%.
- This is a clear example of a company using different types of in-house techniques to detect and mitigate bias.

09 Conclusions

Based on our work, we make the following conclusions.

Firstly, we have identified a set of challenges:

1. 'Fairness' by nature is an abstract concept and highly context-specific, which does not easily lend itself to mathematical definitions and practical implementation.

2. There are myriad different algorithmic definitions of fairness and associate intervention approaches in the Machine Learning literature to tackle this, as well as an increasing number of open-source tools on the market. These generally centre around Group Observational notions of fairness, which are currently most amenable to practical application. However, collectively ,the fast-changing nature of this landscape risks creating confusion for organisational leaders and practitioners.

3. Further, there are a range of practical challenges of applying these definitions in practice. They suffer from mutual incompatibility: there is no singular definition which will work in all cases; and the literature has yet to come up with a satisfactory way of helping practitioners to select between definitions and approaches. This also does not lend itself well to crisp regulatory policy as different sectors grapple with what to expect organisations to do on fairness in their algorithms.

We have also seen from our industry deep-dives into the financial services and recruitment sectors that algorithm and AI use is already reasonably widespread and increasing fast, but that the sectors are relatively nascent in their approach to algorithmic bias. In both sectors, we identified appetite for greater engagement on algorithmic fairness, more clarity on what firms should do in practice, and crucially tools and best-practice practical guidance to put this into reality.

With these in mind, we have sought to do two things:

1. We created an **end-to-end organising framework** for algorithmic fairness, set out in this report, giving for the first time a full line of sight from high-level concepts of fairness through to the range of available open-source tools, in a structured way.

2. Further, in the accompanying **Implementation Handbook**, we have set out the first version of a generalisable workflow for organisational leaders and technical practitioners to use to implement fairness in practice, including how to balance algorithmic fairness against wider considerations, and how to select between different algorithmic definitions. This is underpinned by **Technical Standards**, defining the different terms for data science practitioners. We hope this will be iterated and improved with further engagement, and will prove useful for achieving fairness in practice.

Annex A: Further Fairness Definitions and Interventions

A.1 Remaining algorithmic fairness notions from the framework in Section 02.1

In Section 02.4 we discussed in detail the algorithmic fairness notions which are most relevant to organisational leaders and practitioners, as they are most amenable to practical application. Here, we give details on the other definitions contained within the overall organising framework set out in Figure 3 of Section 02.

Despite their drawbacks in general applicability, some of those notions might still be a suitable choice in a specific situation, and as being subject to current research, future developments may make them more widely used. The definitions below follow the numbering convention from Figure 3.

A.1.1 Procedural definitions

2) Feature-Apriori Fairness, Feature-Accuracy Fairness, and Feature-Disparity Fairness

Fairness Through Unawareness has been combined with quantitative measures of Procedural Fairness (Grgić-Hlača et al. 2018). Rather than making an absolute requirement about which features should be used, we can quantify the fairness of the features the algorithm has access to by surveying relevant stakeholders.

More specifically, the authors introduce three measures - Feature-Apriori Fairness, Feature-Accuracy Fairness, and Feature-Disparity Fairness - each of which assigns a number between zero and one to every input to the algorithm, corresponding to the proportion of members of a panel that believed use of that feature in the model was fair. Feature-Apriori Fairness asks the panel this question in an absolute sense, whereas Feature-Accuracy Fairness asks if the use of the feature in question would be fair if it increases the accuracy, and Feature-Disparity Fairness asks if use of the feature in question would be fair if it increases the disparity in outcomes. In the latter case the line between Procedural Fairness and Outcome Fairness is slightly blurred.

This approach offers a way to make Procedural Fairness more precise, and offers an interesting way to quantify the notion of a "fair process". By making human input the basis for the definition, these notions of fairness are able to capture subtleties in perceptions of fairness that are otherwise difficult to quantify, including relevant contextual information that will automatically be built in.

Perhaps the primary limitation of this approach is that it does not lend itself well to optimisation. We can measure unfairness according to each of these definitions, but the only mechanism to change the level of unfairness that is available to us is adding or removing features from the list of model inputs. Removing a large proportion of the inputs has implications for performance, as noted in the paper. Additionally, whether a feature is considered fair to use in a model will typically depend on how it is used. While this is partially captured by the notions of feature-accuracy fairness and feature-disparity fairness, the model builder is generally unable to guarantee how a feature will be used if included, so these measures of fairness can miss some of those subtleties.

Categorisation: Procedural

A.1.2 Group-based Outcome definitions

7) Subgroup Fairness

Subgroup Fairness was introduced in (Kearns et al. 2018) as a compromise between Individual and Group fairness, and can be seen as Group fairness (e.g., Demographic Parity or Equalised Odds) on a typically large number of structured groups.

It addresses the problem of the arguably restrictive assumptions required for Individual Fairness, such as the availability of a suitable similarity metric. It also addresses the limitation of Group fairness to incorporate fairness across intersections of different groups. For instance, mitigating bias against women is not adequately addressed if it still permits bias against women who are mothers (Dwork and Ilvento 2018a).

Main issues for Subgroup Fairness lie in finding an appropriate selection of subgroups, in the exponential number of possible subgroups, as well as in the computational complexity when implementing Subgroup Fairness auditing and mitigation (Kearns et al. 2018, 2019).

Categorisation: Outcome -> Group -> Observational

8) Unresolved Discrimination

Unresolved Discrimination is one of two causal notions of Group fairness introduced by (Kilbertus et al. 2018). Similar to Conditional Demographic Parity, it takes into account a collection of “resolving variables” with respect to which we do not consider it unfair to discriminate.

A motivating example is given by Pearl’s commentary on claimed gender discrimination in admissions to UC Berkeley. The data appeared to show that women were admitted at lower rates than men, however when department choice was accounted for women in fact experienced a slightly higher admittance rate. In this case, we might consider department choice a resolving variable, as individuals should be free to apply to any department they choose, and if it so happens that women tend to favour competitive departments then that is a legitimate reason for a disparity in outcomes.

Unresolved Discrimination formalises this idea using the language of causal inference, a more detailed discussion of which we defer to the technical standards.

As is the case with other causal notions of fairness, the definition is attractive in its alignment with our intuitive understanding of fairness, however it requires the specification of a causal model which is generally a very restrictive assumption. Moreover, the mitigation strategy presented in the paper requires an additional linearity assumption which places further restrictions on the modeller.

Categorisation: Outcome -> Group -> Causal

9) Proxy Discrimination

Proxy Discrimination is the second of two causal notions of fairness introduced by (Kilbertus et al. 2018). It complements the notion of unresolved discrimination in the sense that rather than requiring that any influence of the protected attributes on the outcome is mediated by resolving variables, we disallow only the influence of the protected attribute on outcomes that arises through the use of certain proxy variables. To quote the paper “this viewpoint acknowledges that the influence of [the protected attributes] … may be complex and it is too restraining to rule out all but a few designated features”.

As with Unresolved Discrimination, the language of causal inference is required to make this precise, so we defer to the technical standards for a more detailed discussion.

As is the case with the other causal notions of fairness that we have considered, proxy discrimination requires the specification of a causal model which is generally a very restrictive assumption. Moreover, the mitigation strategy presented in the paper requires additional linearity assumption which adds further restrictions.

Categorisation: Outcome -> Group -> Causal

A.1.3 Individual-based Outcome definitions

10) Individual Fairness

Individual Fairness, explored in detail by (Dwork et al. 2012), can be summarised as the idea that "similar individuals should be treated similarly". Differing treatment of individuals, rather than groups of people, forms the basis for the determination of fairness.

More precisely, if we have measures of similarity of individuals, and of outcomes, then Individual Fairness requires that the outcomes for two individuals are at least as similar as the individuals are.

The primary challenge with applying Individual Fairness is determining suitable measures of similarity, both for individuals and for outcomes. There are a few natural choices for outcomes, but no canonical approach for individuals, and the level of fairness could be sensitive to these choices.

The paper (Dwork et al. 2012) suggests for example that a measure of similarity between individuals in a loan application process could be based on comparison of their credit scores. In practice, there is a risk that use of such a metric could hide historical or systemic biases that feature in the design of the similarity measure itself. Thus we find that many of the considerations that would ordinarily be important for the design of a fairness measure, instead now become relevant for the similarity measure instead.

Categorisation: Outcome -> Individual -> Observational

11) Meritocratic Fairness

Fairness considered in an online/reinforcement learning context, in which decisions are actions taken in sequence and can lead to different quantitative rewards (Joseph et al. 2016; Jabbari et al. 2017; Joseph et al. 2017). A fair decision is one that always favours higher quality among a group of individuals, where quality is defined by a higher expected reward according to some function of interest (in loan application, e.g., function that outputs how much an individual is able to repay). This fairness notion is, analogously to Individual Fairness, defined through a metric (reward), however the notion is explicitly oriented towards the performance goal.

Example in loan application: the individual with higher repayment rate should be offered a higher loan (Saxena et al. 2019).

Categorisation: Outcome -> Individual -> Causal

12) Counterfactual Fairness

Counterfactual Fairness is a notion of fairness introduced by (Kusner et al. 2017) that provides a principled way to impose a constraint that an individual should not have received different treatment, had their protected attributes taken different values. Making this precise requires the language of causal inference, a discussion of which we defer to the technical standards.

While Counterfactual Fairness is a principled approach to capturing an intuitive notion of fairness, the requirement that we specify a causal model for the data and outcomes is generally a very restrictive assumption.

Categorisation: Outcome -> Individual -> Causal

A.2 Remaining intervention methods from the framework in Section 03.1

Below we set out and assess the currently-known intervention approaches contained within the literature for achieving these other definitions of fairness - specifically Individual Observational definitions.

A.2.1 Pre-processing

Individual Fairness

Optimal clustering / constrained optimisation: For the details on (Zemel et al. 2013) we refer to Section 03.2.1.

A.2.2 In-processing

Individual Fairness

Constrained optimisation: (Dwork et al. 2012) introduce a classification method that minimises a loss function associated with the classifier subject to quantitative similarity constraints, i.e., so that predictions are close to each other if the underlying data points are close to each other according to predefined metrics. Further, (Biega, Gummadi, and Weikum 2018) propose an algorithm based on online optimisation which achieves individual fairness in the context of ranking algorithms.

Subgroup Fairness

Two-player zero-sum game: (Kearns et al. 2018) define fairness with respect to every combinatorial subgroup of protected groups, and develop an algorithm that seeks the equilibrium in a zero-sum game between a learner and an auditor in order to achieve a desired bound on statistical parity and equal opportunity with respect to the subgroups. In a follow-up publication (Kearns et al. 2019), the authors apply their algorithm in an extensive empirical study.

A.2.3 Post-Processing

Individual Fairness

Label modification: Due to the relabelling of data points which likely suffer from individual unfairness**,** (Lohia et al. 2019) achieves besides demographic parity also individual fairness.

Annex B: References

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. "A Reductions Approach to Fair Classification." In *International Conference on Machine Learning*. Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness in Machine Learning*. fairmlbook.org. Biega, Asia J., Krishna P. Gummadi, and Gerhard Weikum. 2018. "Equity of Attention: Amortizing Individual Fairness in Rankings." In *ACM SIGIR Conference on Research & Development in Information Retrieval*.

Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy. 2009. "Building Classifiers with Independency Constraints." In *IEEE International Conference on Data Mining Workshops*.

Calders, Toon, and Sicco Verwer. 2010. "Three Naive Bayes Approaches for Discrimination-Free Classification." *Data Mining and Knowledge Discovery*, September.

Calmon, Flavio, 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*.

Chen, Irene, Fredrik D. Johansson, and David Sontag. 2018. "Why Is My Classifier Discriminatory?" In *Advances in Neural Information Processing Systems*.

Chiappa, Silvia, and Thomas P. S. Gillam. 2018. "Path-Specific Counterfactual Fairness." In *AAAI Conference on Artificial Intelligence*.

Chouldechova, Alexandra. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data*.

Corbett-Davies, Sam, and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *arXiv:1808.00023*.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Crowson, Cynthia S., Elizabeth J. Atkinson, and Terry M. Therneau. 2016. "Assessing Calibration of Prognostic Risk Scores." *Statistical Methods in Medical Research*.

Dieterich, William, Christina Mendoza, and Tim Brennan. 2016. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity." *Northpointe Inc*.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness through Awareness." In *Innovations in Theoretical Computer Science Conference*. Dwork, Cynthia, and Christina Ilvento. 2018a. "Group Fairness under Composition." In *Conference on Fairness, Accountability, and Transparency*.

———. 2018b. "Individual Fairness Under Composition." In *Conference on Fairness, Accountability, and Transparency*.

Feldman, Michael, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. "Certifying and Removing Disparate Impact." In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Garg, Sahaj, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. "Counterfactual Fairness in Text Classification through Robustness." In *AAAI/ACM Conference on AI, Ethics, and Society*.

Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. "Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning." In *AAAI Conference on Artificial Intelligence*.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." In *Advances in Neural Information Processing Systems*.

Hébert-Johnson, Úrsula, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. 2018. "Calibration for the (Computationally-Identifiable) Masses." In *International Conference on Machine Learning*. Huang, Chen, Yining Li, Chen Change Loy, and Xiaoou Tang. 2019. "Deep Imbalanced Learning for Face Recognition and Attribute Prediction." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huang, Gary B., Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. "Labeled Faces in the Wild: A Database Forstudying Face Recognition in Unconstrained Environments." hal.inria.fr.

https://hal.inria.fr/inria-00321923/.

Jabbari, Shahin, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. “Fairness in Reinforcement Learning.” In *International Conference on Machine Learning*.

Joseph, Matthew, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. “Fair Algorithms for Infinite and Contextual Bandits.” *arXiv:1610.09559*.

Joseph, Matthew, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. “Fairness in Learning: Classic and Contextual Bandits.” In *Advances in Neural Information Processing Systems*. Kamiran, Faisal, and Toon Calders. 2012. “Data Preprocessing Techniques for Classification without Discrimination.” *Knowledge and Information Systems*.

Kamiran, Faisal, Indrė Žliobaitė, and Toon Calders. 2013. “Quantifying Explainable Discrimination and Removing Illegal Discrimination in Automated Decision Making.” *Knowledge and Information Systems*.

Kamiran, F., T. Calders, and M. Pechenizkiy. 2010. “Discrimination Aware Decision Tree Learning.” In *IEEE International Conference on Data Mining*.

Kamiran, F., A. Karim, and X. Zhang. 2012. “Decision Theory for Discrimination-Aware Classification.” In *International Conference on Data Mining*.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. “Fairness-Aware Classifier with Prejudice Remover Regularizer.” In *Machine Learning and Knowledge Discovery in Databases*. Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness.” In *International Conference on Machine Learning*.

———. 2019. “An Empirical Study of Rich Subgroup Fairness for Machine Learning.” In *Conference on Fairness, Accountability, and Transparency*.

Kemelmacher-Shlizerman, Ira, Steven M. Seitz, Daniel Miller, and Evan Brossard. 2016. “The Megaface Benchmark: 1 Million Faces for Recognition at Scale.” In *IEEE Conference on Computer Vision and Pattern Recognition*.

Kilbertus, Niki, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2018. “Avoiding Discrimination through Causal Reasoning.” In *Advances in Neural Information Processing Systems*.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” *arXiv:1609.05807*.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. “Counterfactual Fairness.” In *Advances in Neural Information Processing Systems*.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. “How We Analyzed the COMPAS Recidivism Algorithm.” 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Lipton, Zachary, Julian McAuley, and Alexandra Chouldechova. 2018. “Does Mitigating ML’s Impact Disparity Require Treatment Disparity?” In *Advances in Neural Information Processing Systems*. Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. “Delayed Impact of Fair Machine Learning.” In *International Conference on Machine Learning*.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Louizos, Christos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2016. “The Variational Fair Autoencoder.” In *International Conference on Learning Representations*.

Luong, Binh Thanh, Salvatore Ruggieri, and Franco Turini. 2011. “K-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention.” In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. “A Survey on Bias and Fairness in Machine Learning.” *arXiv:1908.09635*.

Nordmark, Viktor. 2020. “The Ultimate Guide to AI Recruiting.” Hubert AI. 2020. https://hubert.ai/blog/the-ultimate-guide-to-ai-recruiting.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. “Social Data:

Biases, Methodological Pitfalls, and Ethical Boundaries." *Frontiers in Big Data*.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. "On Fairness and Calibration." In Advances in Neural Information Processing Systems.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." In *Conference on Fairness, Accountability, and Transparency*.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research*. Russell, Chris, Matt J. Kusner, Joshua Loftus, and Ricardo Silva. 2017. "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness." In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc.

Sánchez-Monedero, Javier, Lina Dencik, and Lilian Edwards. 2020. "What Does It Mean to 'Solve' the Problem of Discrimination in Hiring? Social, Technical and Legal Perspectives from the UK on Automated Hiring Systems." In *Conference on Fairness, Accountability, and Transparency*.

Saxena, Nripsuta, Karen Huang, Evan DeFilippis, Goran Radanovic, David Parkes, and Yang Liu. 2019. "How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness." In *AAAI/ACM Conference on AI, Ethics, and Society*.

The Institute of Internal Auditors. 2013. "The Three Lines of Defence in Effective Risk Management and Control." https://global.theiia.org/standards-guidance/recommended-guidance/Pages/The-Three-Lines-of-Defense-in-Effective-Risk-Management-and-Control.aspx.

Woodworth, Blake, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. "Learning Non-Discriminatory Predictors." In *Conference on Learning Theory*.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment." In *International Conference on World Wide Web*.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. "Fairness Constraints: Mechanisms for Fair Classification." In *International Conference on Artificial Intelligence and Statistics*.

Zehlike, Meike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. "Fa* Ir: A Fair Top-K Ranking Algorithm." In *ACM Conference on Information and Knowledge Management*.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. "Learning Fair Representations." In *International Conference on Machine Learning*.