Financial Conduct Authority

Research Note

11/12/2024

A Literature Review on Bias in Supervised Machine Learning

Daniel Bogiatzis-Gibbons, Lawrence Charles, Harry Dewing, Camilla Gretschel, Maria Jomy, Annette Reid, Ryan Slack

FCA research notes in financial regulation

The FCA research notes

The FCA is committed to encouraging debate on all aspects of financial regulation and to creating rigorous evidence to support its decision-making. To facilitate this, we publish a series of Research Notes, extending across economics and other disciplines.

The main factor in accepting papers is that they should make substantial contributions to knowledge and understanding of financial regulation. If you want to contribute to this series or comment on these papers, please contact Kieran Keohane (Kieran.Keohane@fca.org.uk).

Disclaimer

Research notes contribute to the work of the FCA by providing rigorous research results and stimulating debate. While they may not necessarily represent the position of the FCA, they are one source of evidence that the FCA may use while discharging its functions and to inform its views. The FCA endeavours to ensure that research outputs are correct, through checks including independent referee reports, but the nature of such research and choice of research methods is a matter for the authors using their expert judgement. To the extent that research notes contain any errors or omissions, they should be attributed to the individual authors, rather than to the FCA.

Authors

All the authors at the time of writing work for the FCA, and Daniel is a PhD candidate in Politics at Birkbeck College, University of London.

Acknowledgements

We would like to thank Edward Barrett, Wenjin Li, Lucas Lafone, Freddie Levett, Brittany Bainbridge, Ayesha Dayaji, Vicki Jordan, Henrike Mueller, and Dara Jeyasundaram for their in-depth reviews of the content of this note.

All our publications are available to download from www.fca.org.uk. If you would like to receive this paper in an alternative format, please call 020 7066 9644 or email publications_graphics@fca.org.uk or write to Editorial and Digital Department, Financial Conduct Authority, 12 Endeavour Square, London E20 1JN.

Contents

Non-Technical Summary

This note provides a review of available academic and ‘grey’ literature to explore how biases may arise and may be mitigated in Supervised machine learning models used to make predictions or assist in decision-making about individuals. It is a research piece as part of a broader academic conversation.

In this note, bias refers to *unjustified differences in predictions or decision-making based on the demographic characteristics, or wider life or social circumstances of a person, such as characteristics of vulnerability* (see Chapter 2). These disparities may arise from several sources, discussed in the literature, including historical biases in data, modelling choices, or how humans selectively use predictive models (for instance, if human judgment is used, advertently or not, more to help a particular group).

This note focusses on the technical measurement and mitigation aspects of bias. We do not provide any opinion on what constitutes unfair treatment or discrimination from a legal standpoint, or more generally what is required of organisations to meet any applicable legal requirements, including data protection law.

There are a range of different ways to measure bias outlined in the academic literature (see Chapter 3), which reflect different interpretations of what constitutes fairness. One promising metric is conditional demographic disparity, which asks whether the same proportion of two or more groups would, for example, get a loan based on similar characteristics (e.g., both being employed). However, any measurement of bias will inevitably need to be considered in fuller context of the social and economic situation a model is deployed in and not used just mechanically.

When data on demographic characteristics or wider life or social circumstances can be collected or proxied, several methodologies that may help mitigate bias (see Chapter 4) were identified from the literature, for example by flagging cases for human review where potentially biased or less confident predictions are made.

In cases where demographic characteristics or wider life or social circumstances cannot be measured, we identified two alternatives from the literature that could help mitigate the potential for bias to occur (see Chapter 5).

- Firstly, using existing evidence on potential biases to inform decision-making, for example identifying features that may lead to bias based on a market level view.
- Secondly, intuition-based evaluation of generated variables or rules which help identify potential biases.

Overall, addressing bias in machine learning models requires careful consideration of context, as this will ultimately inform which demographic characteristic/s if any might be of concern, as well as the methods that are appropriate (if any) to measure and mitigate any potential bias.

1 Introduction

Background

Artificial intelligence (AI) models represent a potentially important growth industry for the UK. Within financial services, AI has the potential to improve consumer outcomes, for example through better targeting of consumer support. However, AI does not come without risks.

In its AI Update (FCA 2024), the FCA highlighted how key elements of its regulatory framework relating to consumer protection would be relevant to upholding fairness in the use of AI systems. As part of this, the Update flags the FCA’s Guidance on the Consumer Duty which highlights that firms using AI technologies in a way that embeds or amplifies bias, leading to worse outcomes for some groups of consumers, might not be acting in good faith for their consumers, unless differences in outcome can be justified objectively.

This note focusses on bias in supervised machine learning models, which are in essence cases where some available outcome (such as future risk of defaulting on a loan) is predicted based on past training data. Specifically, it focusses on how differences in predictions for groups sharing different demographic characteristics or characteristics of vulnerability might be measured, understood, and mitigated.

For the purposes of this note, when referring to “demographic characteristics” we mean:

- the set of protected characteristics protected by equalities legislation (age, disability, gender reassignment, marriage or civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation), and
- broader demographic characteristics (such as socio-economic status or region)

By “characteristics of vulnerability” we mean the indicators set out in our Guidance for firms on the fair treatment of vulnerable customers (FCA, 2021) related to the 4 key drivers of health, life events, resilience, and capability.

Empirical work on AI shows how there could be potential differences in predicted risk or outcomes in financial services based on demographic characteristics. For example, a recent paper shows that more complex credit scoring models disadvantage minority borrowers in a US context, partly by inferring race, and partly through more accurate targeting of individual risk (Fuster et al., 2022). In a UK context, there has been historical work on patterns of redlining by landlords in the UK which could generate potential for long-run differences in the riskiness of areas in which members of minority ethnic groups tend to live (Wetherell, 2020).

In this note, we focus on one kind of AI called supervised machine learning, where a set of known inputs like prior credit history is used to predict a fixed future output, such as the risk of defaulting on a new loan or credit card (Hastie, Tibshirani, and Friedman, 2001). We will usually refer to cases of binary classification i.e., where there are only two possible outcomes such as an individual defaulting on a loan or not, in part because this is where the academic literature is best developed. We also focus on differences in average outcomes by different demographic groups, which relates to what is called distributive justice in political science, or who gets what outcome or is charged what price on the basis of some inalienable characteristic of that person (see e.g. Gabriel, 2022).

By bias we specifically mean unjustified differences by group membership in predictions of risk or other future outcomes like willingness to pay, that typically feed into decisions about the prices to charge and decisions to provide products to individual consumers with adverse impacts for the impacted groups. Bias is intentionally defined here more narrowly than ethical notions of fairness.

Our underlying research focused on the following technical question: how can these differences in predictions be measured when using supervised learning, and how can they be mitigated in that context? We focus on both how existing literature suggests this can be done when demographic characteristics or characteristics of vulnerability can be measured for or proxied (that is predicted using a probabilistic model based on some other data such as a person’s name), as well as alternatives when this is not feasible. However, we are aware that even if differences in predictions and ultimate outcomes exist for persons from different demographic groups, there are then separate questions around the causes of and any justifications for such differences. For this, it will be important to consider the broader context in which a model is being used and, where appropriate, any non-discriminatory business or other justifications for any measured differences.

Given the focus and aims of the work presented here readers should be aware that this note:

- constitutes research to spark debate and contribute to academic discussion, rather than any form of guidance or direction about what firms or practitioners should do.
- does not set out any particular expectations for how firms should approach managing AI risks (in all cases, firms will need to consider the risks relating to AI adoption in the context of their specific use cases and in light of applicable requirements))
- is not a comment or statement about the direction of the broader debates on what constitutes fairness or discrimination in the AI space.
- will not cover risks from language models (including Large Language Models such as ChatGPT and older models such as BERT) often stemming from biases in their underlying embeddings, which are covered in a separate Research Note Dwyer, Francis, and Tyagi (2024).

This note is primarily intended for those with some understanding of statistics or AI but is written in non-technical language. This Research Note should help signpost important considerations in the technical development of these models, as well as first-line mitigation strategies, that is those before a broader set of ethical, commercial, and legal considerations would need to be explored. Readers may also want to consider the RTA (2020) report into bias in algorithmic decision-making, as well as ICO (2023) guidance.

Key points

We identified the following key findings from our research:

1. Systematic differences in a supervised learning context in predictions or decision-making across demographic characteristics or characteristics of vulnerability can result from human inputs, historical biases in data, or choices in the modelling process itself.
2. There are a variety of different metrics that can be used to measure these differences in supervised learning predictions, which reflect varying notions of what bias (and ultimately fairness) means. Based on existing literature, one useful notion of bias is conditional demographic disparity, which asks, for example, whether the same proportion of two groups of people get a loan, on the basis that they have suitable characteristics such as being employed.
3. If a demographic characteristic can either be collected or proxied, there are several feasible approaches for removing bias. These include methods for pre-processing data (including removing biased features or adjusting features to remove bias), adjusting modelling to choose a less biased alternative model, and post-processing (including flagging cases where the model makes biased or less confident predictions for human review).
4. Finally, in the circumstance that characteristics or life circumstances cannot be measured or proxied there are two alternatives: creating industry knowledge bases, and intuition-based consideration of whether variables or rules generated might cause biases.

2 What is bias and how does it come about?

What is bias?

The general idea of bias

For the purpose of undertaking this literature review, we first wanted to consider what bias is, for which we have drawn on extensive reading of literatures on related concepts such as fairness and discrimination in machine learning, AI, and statistics. We did this in order to frame the rest of the review and to set broad parameters for the types of bias we wanted to consider. We set out in this section a concept of bias that underpins this piece, although this is a contested area of research, and other notions are required for example when using pricing or language models.

In this sub-section, we define bias qualitatively and in general, before proceeding in the next two sub-sections to explore how it arises, and how it can be measured.

In general, in the behavioural and decision theory literatures, a bias is generally understood as any systematic decision-making error. For the purposes of this review, we will focus on biases resulting from a predictive modelling process that occur based on some demographic characteristics or wider life or social circumstances of a person, such as characteristics of vulnerability.

The nature of the bias that may arise and the impact it could have on end users and consumers is likely to vary depending on the context in which a particular model is used, and the customers served. For example, more significant impacts could arise in contexts such as loan decisions.

We focus on bias in the form of unjustified differences in predictions or outcomes for groups of consumers with particular demographic characteristics or life circumstances.

In terms of the scope of the characteristics considered in this review, it could denote any characteristic of a person or group of people or any fact about their wider life circumstances. For illustrative purposes, we opt to discuss three sets of groupings in this review.

Protected Characteristics

The first grouping are the protected characteristics set out by the *Equality Act 2010*:

- Age,
- Gender reassignment,
- Being married or in a civil partnership,
- Being pregnant or on maternity leave,
- Disability,
- Race including colour, nationality, ethnic or national origin,
- Religion or belief,
- Sex, and

- Sexual orientation.

Vulnerability

The second grouping that one might consider under a bias lens could encompass the concept of vulnerability, as in the FCA's Vulnerable Consumer Guidance (2021). The FCA's general view is that vulnerability is a spectrum of risk, as depending on life circumstances any person can be vulnerable. However, when considering bias in this area, one might use what the FCA terms "characteristics of vulnerability" which are generally associated with four key drivers:

- Health – health conditions or illnesses that may make it difficult to undertake day-to-day tasks.
- Life events – such as bereavement, job loss, or relationship breakdown.
- Resilience – low ability to withstand financial or emotional shocks.
- Capability – low knowledge of financial matters or low confidence in managing money, literacy, numeracy, or digital skills.

Other Demographic Characteristics

Finally, there are wider demographic characteristics that may nonetheless be worth considering in the context of bias like region of the UK, occupation, or income level.

There are a number of overlaps and interactions between these three distinct categories of characteristics. For example, health conditions will be relevant both to the protected characteristic of disability as well as characteristics of vulnerability. There may also be correlations between the numbers of vulnerable persons within different demographic groups (i.e., those sharing particular protected or wider demographic characteristics). However, we thought setting these different groupings out would provide useful framing for our note.

Bias in an algorithmic (vs. human) context

Group bias is a property both of *predictive algorithms* (of which we only address supervised machine learning examples here), that is rules for making statements about the likelihood of some future event, and *decisions* about individuals, possibly made with the assistance of an algorithm. This is usually measured on average across the members of the different subgroups.

For example, in comparing white and non-white individuals, group bias in a *predictive algorithmic* context would mean the differences by group membership in prediction, like the risk of defaulting on a loan. Those differences could refer to different levels of accuracy for different subgroups, or different predictions made about people with otherwise similar profiles, for example but for their race (see Verma and Rubin, 2018, or Section 2).

When human decision-makers become involved, it would mean the difference in *final decisions* made. For instance, in the approval of a loan this could involve consideration of whether a human decision maker overrides algorithmic recommendations differently for white and non-white customers. Angelova, Dobbie, and Yang (2023) find in a bail-decision context that algorithmic overrides increase the disparity between black and white defendants in release probabilities. Whilst, for the purposes of this review, we focused on model bias in a supervised machine learning context, it is important to note that the biases that can be introduced through human input into decision-making can also be significant.

One critical point is that there may be observed baseline differences in outcomes between groups which do not by themselves suggest unfair or discriminatory outcomes once we account for other factors. For instance, if one group lives in predominantly rural areas and another in predominantly urban areas this may have justified implications for outcomes in the context of insurance pricing. Similarly, there may be either affordability or business reasons why, for instance, income level might be associated with different outcomes in relation to lending decisions, but it may be of interest to understand why these occur and if they are justified.

What contributes to model bias?

In this section, we refer to supervised machine learning models and note that the situation would be somewhat distinct in an unsupervised (e.g., clustering) case.

In this review, we found in the literature documentation of how bias can be introduced into the predictive algorithms built using a supervised machine learning approach throughout the building process from problem framing to production i.e., building and employing the algorithm. These findings demonstrate how bias may arise at distinct stages of the modelling process and may therefore need to be measured at multiple stages, given changes over time. That can occur, as Ferrara (2024) points out because machine learning models can exhibit 'Butterfly Effects', where small changes to initial inputs cause large changes in model outputs due to shifts in market or consumer behaviour or model instability. Feast (2020) suggests that bias might be measured at:

- the time of data collection or model building
- after human intervention (as human overrides of model predictions could either reduce or increase bias)
- on an ongoing basis, as if new applicants are not similar to historical training data, then bias might change on this basis also.

Measuring bias at multiple stages might then, as Feast discusses, point to whether it can be mitigated, and whether there is a good justification for the bias that results.

The following section will discuss literature on how bias can be introduced into an algorithm during the problem-framing, data collection and algorithm-building steps.

Problem-framing

The first challenge of any algorithm's creation is how to frame the problem that requires solving. This involves considering the problem, any available inputs, and the output. During this step, one could consider the bias arising from using particular features (also known as covariates or independent variables), as they could be associated with demographic characteristics or characteristics of vulnerability.

In the financial industry an example could be creating an algorithm to predict which cash machines can be removed. An input could be the amount of money that individuals withdraw from the machine where each transaction with a lower amount suggests the machine is more likely to be removed. This might negatively impact low-income areas, potentially also correlating with demographic characteristics such as race, age or health, given disparities in income (ONS, 2023).

That is because, for example, lower-income individuals, old age pensioners, or disabled persons on benefits might only withdraw enough for each week or shop using cash to manage their budget. An alternative input could be the number of withdrawals or the number of individuals using the machine. This could allow a company to remove underused machines and not disproportionally affect lower-income, older, or disabled individuals. When framing the problem, a data practitioner must consider how different input sources could introduce bias into the algorithm.

Similarly, seemingly unrelated characteristics might be close proxies for demographic characteristics or characteristics of vulnerability, especially when sophisticated predictive algorithms such as supervised machine learning models are used. This should be thought about carefully. For example, Kusner et al. (2019) discuss the examples of living in a high-crime region or area with poor exam results which might correlate with socio-economic status or race.

As part of the problem framing a data practitioner may also consider what supervised learning method or model would be most appropriate, given what target the modeller is trying to predict and the features they have available to them.

Data collection

After having framed the problem and chosen the data sources, the next step is data collection for algorithm training and testing. Algorithms are trained to find and replicate patterns in the dataset by looking at how different variables link to the outcome. Without intervention by modellers, biases in the dataset are likely to be reproduced by the algorithm. Practically, a dataset can exhibit two different biases: representation and past beliefs.

Firstly, if the data does not accurately represent the population covered in the problem this can lead to some groups being underrepresented or overrepresented, which is termed "representation bias" in the literature (see e.g. Shabhazi et al. 2023). Underrepresentation of a group can lead to large generalisations and inaccurate predictions leading to worse outcomes for some groups. For instance, there has been analysis of the issues with past data that some facial recognition algorithms rely upon (though it is worth noting that these rely on deep learning, rather than supervised machine learning). Buolamwini and Gebru (2018) found that of three facial recognition algorithms, the worst inaccurately recognised 0.8% of lighter-skinned males compared to almost 35% of women with darker skin. This inaccuracy in some groups is due to the underrepresentation of those with darker skin in the dataset compared to those with lighter skin.

For the finance industry, a hypothetical example of a representation problem could occur if spending data were used from an open-banking app with mainly male customers. If the algorithm generalised the spending of men across the entire population, it may be unlikely to be able to predict female spending patterns with the same accuracy. Data that does not accurately represent the population to whom the algorithm will be applied can lead to negative outcomes for some groups when used to make decisions about them. For instance, Blattner and Nelson (2021), find that credit scores are noisier estimates of default risk for minority applicants as they are more likely to have shorter or no credit histories.

It could also happen due to selection problems of the type first explored, where for example in credit scoring only those who get credit form part of the training dataset. If certain groups of applicants get more expensive or no credit historically, then they may be denied credit or get more expensive credit in the future. This is exacerbated by the fact that past data represents the noisy judgments of human decision-makers, which may encodes attendant biases (Cowgill and Tucker, 2019).

Secondly, a dataset can propagate existing social biases or biases in the data collection process itself. Examples of documented existing biases in financial services outside a UK context include female investors being given lower quality investment advice than males (Bucher-Koenen et al., 2021) and higher interest rates on mortgages for some ethnicities (Bayer, Ferreira, and Ross, 2014). Using the investment advice example, if a supervised machine learning model was trained on the types of advice that had been previously given to customers, biases might result even if the customer's sex was not used in the model (a strategy Kusner et al. 2019 term "fairness through unawareness").

More broadly, a hidden assumption that training on past data embodies is that the data-generating process is invariant, that is it does not change over time. An example of a data collection bias might be that certain customers are more likely to give permission for their data to be retained, which might lead to a highly imbalanced dataset.

Algorithm training

Having framed the problem and collected the data for training the next step is to begin training the algorithm. The process of model training encompasses various stages, each with the potential to introduce or perpetuate bias, starting from pre-processing of data to the model's deployment.

Pre-processing techniques such as weighting and handling missing data (Zhang and Long 2023) are pivotal stages where biases can seep in. Of course, this is also part of generally good practice for model risk management, even outside AI assurance or AI debiasing procedures. In a supervised machine learning context, the weighting of each data input is done either manually or by using different searching algorithms. A higher-weighted data point will have a greater impact on the algorithm output than a lower-weighted one. Usually, these weights are picked to maximise the accuracy of the model, but as Chouldechova (2017) showed in the context of recidivism in criminal justice, this can lead to biases along racial lines even when race is not explicitly considered.

Further along the model building pipeline, dimensionality reduction and feature selection can inadvertently remove important context that might be crucial for making fair or unbiased predictions. An example of this is Amazon's roll out of same-day delivery across cities in America (Bloomberg, 2016), where ZIP codes were selected with high concentrations of Prime members for the same day delivery access. While ostensibly a neutral selection criterion, lower concentrations of Prime members in majority-Black neighbourhoods then meant that these were less likely to be chosen for same-day delivery access. This then meant for those customers in majority-black areas they were likely to be paying the same price as customers in majority-white areas but receiving a lower quality service. The choice of model is also critical; reusing old models can perpetuate existing biases, while more complex models, although potentially more accurate, are harder to interpret, making it challenging to identify and address biases within them.

Lastly, model drift refers to changes in the data or context that a model is applied to over time. Without careful monitoring and updating, models can become biased as the world changes around them, underscoring the utility of continuous oversight and re-evaluation of model performance and fairness. For example, this can occur if there are what Ensign et al. (2018) term "feedback loops" when algorithms that learn from human biases then alter future human decisions made with the algorithms, in this case that certain groups of applicants do not get credit and therefore might continue to be missing from the data.

To recap, bias can be introduced into algorithms from sources throughout the model-building process: from the way data is collected, generated, processed, weighted, or used.

Bias metrics and what they measure

In this sub-section, we focus on measuring bias in a binary prediction context, for example whether a person will default on a loan or not, or whether they will make an insurance claim or not. Binary prediction is important because many financial decisions are binary, such as whether a person will buy a product or not, or whether they will fall behind on payments, or whether they will make an insurance claim. It is also the simplest outcome type to consider bias in, which is likely why it has garnered the most attention in the academic literature.

There are generalisations of the notions discussed here to look at continuous outcomes like claim or loan sizes, for example looking at differences in means, differences in quantiles (Liu et al., 2022), differences in entire distributions (Mary, Calauzenes, and Karoui, 2019), and finally a specific literature on different notions of what constitutes fair pricing (see e.g. in an ML context Cohen, Elmachtoub, and Lei, 2020, and the FCA's publication Starks et al., 2018).

Further, for any chosen metric, Lum, Zhang, and Bower (2022) show that especially if some groups are very small, then the statistical estimate of group bias in a given dataset may be significantly different from the "true" bias. The authors provide a corrected variance estimator that allows for more rigorous quantification of the uncertainty around estimates of bias.

We concentrate on two metrics for bias in a supervised machine learning context here for illustration, but it is worth noting that there are substantially more, each with their own advocates. We take the example of a hypothetical loan provider who wants to model whether an applicant is likely to default or not, given information on their past credit history and then examine bias based on gender. Explanations here are based on this example and in words, for more mathematical details please see e.g., Castelnovo et al. (2022).

Our first metric is demographic parity (Barocas, Hardt and Narayanan, 2023), which involves enforcing equality of outcomes across groups. Here that means the same proportion of men and women will be predicted to not default on a loan by the predictive model, and therefore the same proportion qualify for a loan. This metric is simple and intuitive, but it does not adjust for individual characteristics such as income or past credit history, and therefore may not be considering legitimate differences in circumstances.

An alternative is conditional demographic parity, which produces a conditional quality of outcomes on relevant characteristics of people and will account for historical biases in the data. It was invented in Kamiran, Žliobaitė, and Calders (2013), and has been advocated for by Wachter, Mittelstadt, and Russell (2020). It means that the same proportion of male and female applicants qualify for loans given a set of characteristics, such as that they are employed, or have an income over £50,000, or have equal credit histories. It does adjust for individual characteristics and is outcomes focussed. However, there are always more possible variables to adjust for, and so this requires careful consideration of what features to "condition" on or put more simply what characteristics of a person are legitimate to term it a like-for-like comparison.

Conditional demographic parity treats any individual characteristics that it accounts for as having developed free of other forms of bias, or at least outside of the control of the modeller. That may not be a socially reasonable assumption, for instance women see poorer employment outcomes on average for a range of reasons (see e.g., in a UK context ONS 2018). However, wider structural inequalities or biases are outside the modeller's control and are not taken into account for the purposes of this metric of bias provided they are justified to take into account.

Therefore, a possibility would be to causally model the relationship between the outcome, features, and the demographic characteristic. We would then consider what biases result as compared to a person who instead lived in the counterfactual universe where they were born as the opposite sex (Kusner et al. 2017). However, this is subject to two key problems.

- first, writing down such a causal model is likely infeasible for most relevant scenarios, especially as most human behaviours have a range of determinants, and their causes are not known sufficiently well.
- second, it in effect transfers social responsibility for issues entirely outside the person or firm control to the person or firm.

In general, if there is thought to be no selection bias or historical patterns of discrimination, then sufficiency or separation (see Glossary) could be considered as metrics for bias. However, conditional demographic disparity is likely preferrable given selection bias is present for many modelling contexts and because it adjusts for important characteristics to compare "like-for-like" persons without requiring complex or infeasible causal models to be constructed.

3 De-biasing algorithms where data on demographic characteristics or characteristics of vulnerability are available

What is de-biasing in a supervised machine learning context in the current academic literature?

When we refer to "de-biasing" methods from the academic literature, this term does not imply achieving a completely fair predictive algorithm resulting from a supervised machine learning modelling process without any kind of bias because that is generally impossible. Therefore, oversight and a human-in-the-loop are considered to be important checks on the bias of models. Following the work of Balayn & Gürses (2021), what we mean is that a method has been implemented to guarantee adequate "statistical parity" based on one or more bias measures outlined in Section 2. This involves ensuring that different demographic groups have approximately equal predictions, after any adjustments control for characteristics that can justifiably be taken into account.

It is important to emphasise at the outset that, at least in the current literature, there is no one-size-fits-all strategy to debiasing, and that any process could carefully think through why bias might be occurring and whether a particular strategy will therefore be likely to mitigate any sources of bias. Further, mitigating bias against one group may reduce the accuracy of the model, potentially reducing the overall benefits to society of having the model and/or leading to bias on other demographic characteristics or characteristics of vulnerability.

In this section, we review de-biasing strategies from the academic literature which rely on collecting or proxying data on demographic characteristics or characteristics of vulnerability alongside model training data. Most of these methods do not rely on the availability of data on demographic characteristics or characteristics of vulnerability or proxied data for new customers to which a model might be applied to generate new predictions. However, they do rely on data on demographic characteristics or characteristics of vulnerability (or proxies) in the training data.

To implement the methods of measuring bias, outlined in Section 2, and mitigating bias, outlined in this section, organisations need access to demographic data pertinent to the bias they are seeking to test. Debiasing methods require collection of demographic characteristic data on at least the sample used to train the model, and some require collection for all consumers.

However, organisations may face several practical, legal, and technical challenges in obtaining data of this nature, which is often sensitive. Driven by an increasing need for high-quality demographic data to assess biases, there is growing interest and research into ways of overcoming these challenges. We have considered in Section 4 some of the techniques that may be available where the relevant data cannot be obtained.

In the next section, we consider the case when such data or proxied data is not available.

De-biasing techniques

Corrales-Barquero, Marín-Raventós, and Barrantes (2021) classify methods of de-biasing into three stages:

- pre-processing methods which de-bias the data directly,
- in-processing methods that attempt to de-bias model construction, and
- post-processing methods that mitigate bias after model building or during model deployment.

The authors note that there might be more than one demographic characteristic to account for, with many methods and metrics not accounting for all of these. A more complete list of techniques can be found in that paper or in Leslie et al. (2023), we concentrate here on a selection of the main types of techniques.

It is worth noting at the outset that these solutions are all partial and come with important limitations which we discuss here in some detail. However, in some circumstances they may represent better solutions to the problem of model bias than realistic alternatives, such as either not using predictive models or omitting using model features based on intuition.

Pre-processing

Pre-processing techniques seek to address issues around bias before a classifier is built using a supervised machine learning process. A first way this can occur is by focussing on individual features that may be associated with a demographic characteristic and therefore to bias against a particular group. In general, the techniques around individual features have significant limitations under most reasonable scenarios.

Individual features can be "suppressed" or more simply excluded from a predictive model construction if there is a strong association between that feature and a demographic characteristic (Kamiran and Calders, 2012). This is often ineffective as unless all potentially biased features are excluded—which may lead to a sharp loss in accuracy—bias may just transfer onto other included features.

Individual features can also be modified, for example through the Disparate Impact Remover of Feldman et al. (2015), who provide a method for shifting the distributions of features of different groups under a demographic characteristic like between different racial groups or income levels. This does have the advantage that it means important features can be retained, and the "unbiased portion" of them still contribute to the model. However, it has three key disadvantages:

- it is likely to impair the explainability of the subsequent model as explanation will be in terms of adjusted features.
- it is likely to be seen as procedurally unfair given it uses different inputs for different groups.
- unlike most de-biasing methods, it requires knowledge of the demographic characteristic in subsequent model deployment to continue to adjust features.

A second way is to reweight or resample the data to fix differences in how accurate predictions are made for minority groups (Kamiran and Calders, 2012). These techniques are broadly applicable and do not suffer from the issues around adjusting or removing individual features.

One simple weighting scheme is choosing weights to ensure that the training sample is representative of the population that the classifier applies to, for example for a general use product, this could be the population of British adults. This is likely to be effective if the problems stem from insufficient minority individuals in a training sample, for example if certain groups of applicants tended not to be successful or even apply for certain products. If certain groups of applicants had different paths to say become in arrears, then the overall model reflecting their experiences might be important. It is worth noting that this assumes that observed individuals are reasonably representative of any demographic characteristic groups they are a member of, for example that observed disabled adults in the data are representative of disabled adults in the population.

Kamiran and Calders (2009) instead propose reweighting samples by taking any disadvantaged subgroup and up-weighting cases where those individuals are successful in achieving an outcome, while down-weighting successful cases from advantaged subgroups. One issue with any weighting scheme is that not all machine learning estimators are designed to work with weighted data. An alternative proposed by Kamiran and Calders, 2012 is to change the sampling of the data, for example to up-sample cases where applicants from disadvantaged subgroups are successful, that is by replicating those data points.

In-processing

In-processing techniques aim to mitigate biases during model training, with two principal techniques that are primarily used to achieve this: adversarial debiasing and regularisation. Adversarial debiasing stems from Zhang, Lemoine and Mitchell (2018). The idea is to imagine an adversary who is trying to learn the demographic characteristics or wider life circumstances of individuals from the predictions of your model. If the modelling process is set up so that the model tries to minimise the adversary's ability to do this while maximising accuracy, then this should de-bias a model's predictions. Berk et. al. (2017) demonstrate a general method for regularisation (a term which means altering an optimisation problem to add a penalty for complexity), which similarly attempts to trade-off bias and accuracy by adding a tuning parameter to the model building process.

Both these techniques tend to work by adjusting the relationships between features and the outcome that are learned from the data through imposing a form of penalty for producing biased models in addition to the usual task of maximising a chosen accuracy metric. They are likely to be effective or useable in a range of contexts, and the resulting model can be interpreted as it would be normally. This is unlike, say, the Disparate Impact Remover. That said, it might be useful to interpret a model constructed both with and without an in-processing technique to understand its effects – for example, by comparing importance plots or average local effects plots (see Molnar, 2022 especially Chapter 8).

Post-processing

Post-processing techniques aim to mitigate bias after model construction. One method is similar to in-processing in spirit but adjusts the outputted probabilities after modelling has occurred. A method that performs this adjustment is Chzhen et al. (2020), who transform the output of a regression to satisfy demographic parity. They demonstrate that their method performs well in terms of reducing bias while minimising the loss of accuracy across distinct problem domains. In principle, this need not lead to a loss of interpretability if measures like feature importance are calculated on the transformed model.

A second method of post-processing is to highlight cases where the model is uncertain about the prediction it makes or bias risk is high such as in Kamiran, Karim, and Zhang's (2012) method of reject option classification, which could then be used to either adjust predictions or to flag cases for human review. However, it is important to note that human review itself may need auditing, as otherwise prejudice or subconscious bias against groups may worsen the bias in the decision system.

Considerations

First, it can be the case that debiasing worsens predictive accuracy because it may remove genuinely predictive relationships between a feature and the outcome in the process of removing bias. Empirical evidence on this point is mixed. Kingsman (2021) finds that attempting to ensure no bias across multiple different demographic characteristics and ways of measuring bias tends to worsen accuracy. However, this relationship is not a simple, linear one and it does not occur in all the datasets they studied. Rodolfa, Lamba, and Ghani (2021) find negligible trade-offs from using bias mitigation methods across educational, criminal justice, and housing safety domains. Lee and Floridi (2021) find that for four of the five classes of machine learning models they studied, more complex models with generally higher accuracy rates in a U.S. context disadvantaged black mortgage borrowers compared to models with generally lower accuracy rates. In practice, this trade-off must be tested for a given model, given that evidence is ambiguous, often from the substantially different U.S. context, and dependent on modelling choices and chosen accuracy metrics.

A second important consideration is that reducing bias present on one demographic characteristic such as race can worsen bias on a different demographic characteristic such as sex. For example, if a loan approval model is adjusted to ensure equal approval rates across racial groups, the model might overcorrect by relying more heavily on other factors, like income or employment status, which may correlate with gender. This could lead to more men being approved for loans than women, thus increasing gender bias. It is important in the modelling process then to consider which demographic characteristics or characteristics of vulnerability are prioritised for mitigation efforts, based on the risk of bias arising based on those characteristics.

Unfortunately, most of the techniques we reviewed for mitigating bias in supervised machine learning models work based on a single characteristic, and some of them assume that characteristic is binary. Therefore, further academic research in this area would be welcome, as often more than one characteristic needs to be balanced against each other. One exception is Singh et. al (2021), who propose a generalisation of this idea called DualFair to mitigate bias and handle considerations of how bias and accuracy can be in tension and a new metric called Alternative World Index (AWI) to measure it. DualFair works by subsetting a dataset into different combinations of demographic characteristics, and either oversampling and under sampling from each outcome within this subset to get a class-balanced dataset with no selection bias. Next, it uses situation testing, which will test all combinations of these protected attributes on the results from the ML models and will remove any value where the prediction in that data subset is different from the others' subsets.

Finally, there will also be privacy considerations to take into account when seeking to reduce bias in modelling. As mentioned, these techniques rely on having access to data on individuals' demographic characteristics and wider life circumstances. Organisations would need to navigate data protection law requirements when collecting and processing personal data for this purpose. There is also the relatively theoretical risk of "privacy attacks" where someone's race or sex is inferred from the predictions given by a bias-mitigated model (see e.g., Chang and Shokri, 2021), but in practice this requires access to model predictions, which for privately held models then becomes more of a cybersecurity than data science concern.

In conclusion, there are a variety of techniques to debias supervised machine learning algorithms provided that per the previous section demographic characteristic data is available or can be accurately proxied. Some of the most promising techniques here are adversarial debiasing which try to make a demographic characteristic hard to predict using the output of the original model, and post-processing techniques which apply changes to predictions to ensure they are not biased after a model is initially developed. There are important considerations which model developers would need to consider around the effects of debiasing a model while trying to maintain accuracy, on how addressing bias on one characteristic may worsen bias on other demographic characteristics or characteristics of vulnerability, and on data privacy.

4 Mitigating bias when data on demographic characteristics or characteristics of vulnerability cannot be collected or proxied

In this section, we review two imperfect substitutes suggested by Veale and Binns (2017) for measuring and mitigating bias that can be used even when data on demographic characteristics or characteristics of vulnerability cannot be collected or proxied for, notably exploratory bias analysis, and collaborative online platforms for sharing insights into fair algorithms.

They are imperfect because they involve judgments rather than direct measurement, and as should be clear from the discussion in Section 2, bias can result from quite nuanced, seemingly neutral technical decisions. At their heart, these strategies involve trying to either remove features or modify algorithms to remove relationships that are suspected to cause bias. Where they vary is in terms of how insights or tentative hypotheses as to how to do this are shared.

The reason for the reliance on a single paper here is that this scenario, despite how common it is likely to be due to trust and other considerations, is not to our knowledge analysed elsewhere in academic work on AI. That is because any solutions here would need to be indirect, and it is difficult to measure how effective they would be. As will be highlighted in the conclusion, more concrete proposals for bias-mitigation strategies in the absence of data on demographic characteristics or characteristics of vulnerability are therefore crucial.

However, Veale and Binns' proposals are certainly more widely applicable than what Kusner et al. (2017) term "fairness through unawareness", that is the claim that an algorithm is unbiased simply because it does not include a demographic characteristic as a predictor. That is unlikely to be a tenable strategy to prevent bias, because it would require that an algorithm designer (who may have their own cognitive biases in thinking about the modelling process) is strongly convinced that none of the issues described in Section 2 apply.

Collaborative online platforms

Veale and Binns (2017) call for the establishment of online fora to share experiential knowledge about the construction of more ethical algorithmic systems, especially given the likely relative comfort of data scientists in using collaborative tools like StackExchange and wikis, and more broadly research summarisation initiatives like The Cochrane Collaboration in health. The establishment of knowledge sharing fora seems especially important given the ever-expanding and often highly technical literature on bias, and related questions on accuracy, explainability, and transparency. The depth and difficulty of this literature is likely to cause barriers to entry to new data scientists seeking to understand and implement solutions to these issues.

Information this could contain includes research on human behaviour in a given field (such as retail insurance, payments, or credit), common features and associations with demographic characteristics or characteristics of vulnerability, and the effect of model complexity on bias.

Summarising contextual research

A knowledge sharing forum could summarise research about how human behaviour in a given field is likely to determine outcomes, and how these vary across demographic characteristics or characteristics of vulnerability. For example, this could include information on how disability status might have affected financial product choice, or how historic housing discrimination might change patterns of where non-white people live.

Summarising social scientific research here would need to take into account research gaps, potential contradictions between different studies, the issues with failed replications in research findings (for a lucid summary within psychology, see Korbmacher et. al., 2023), and issues with peer review processes.

However, there is a significant volume of this kind of research, much of it in the only partially analogous US context. This would require a significant amount of human effort to summarise, although appropriate use of technologies such as Large Language Models (LLMs) may to a degree mitigate this problem.

Common features and their associations with demographic characteristics or characteristics of vulnerability

Further, a knowledge sharing forum could include common features in each domain (for example, location for motor insurance or payment history variables for credit) and how they might influence the bias of models. This could occur using publicly available aggregated datasets such as survey data, where available, or where a third-party has published their modelling results on a privately held dataset. The list of features could be used to consider omitting some features from a model that are strongly associated with a demographic characteristic, or their replacement by potentially less strongly associated alternatives. That would be particularly true if those features were not key to making business or affordability decisions but had been included merely as extra training data without a clear justification. The optimal approach here is likely to be demonstrating which variables influence the bias of a predictive model the most, for instance what variables most influence the racial bias of a credit risk model. Recent innovations such as feature influence functions (Ghosh, Basu, and Meel, 2023) allow quantification of the marginal influence of features on model bias.

However, this may not be feasible for two reasons. First, the findings on the bias attributable to any specific feature within a given model will depend heavily on the specific customer base and to the features chosen, and so therefore may not generalise well. Second, there may not be data available which contains all of the potentially biased features, the demographic characteristics or characteristics of vulnerability, and the target variable for prediction. More commonly, publicly available data will only contain information on potentially biased features and demographic characteristics or characteristics of vulnerability. For instance, the published aggregated data on the UK census only has information on local-area ethnicity statistics (ONS, 2021b) which could be associated with apparently unrelated information like local-area deprivation published in the IMD (CDRC, 2024). But neither the census nor the IMD have credit information available. Further, as noted previously historical datasets may have been biased in terms of their data collection practices.

Therefore, an alternative would be to look at the association between a feature and a demographic characteristic, which might be more commonly available. However, in the context of this note, any feature which causes bias must be associated with a demographic characteristic or characteristic of vulnerability, but this is not sufficient for contributing to bias as it must also be included in and important to a model's predictions. Further, another limitation is that associations which exist in aggregated data (for example, between local-area ethnicity and local-area vehicle accidents) may not exist in individual data or in other levels of aggregation due to Simpson's Paradox.

Model complexity

Finally, these databases could include information on the impact of model complexity, like the choice of gradient-boosted regression trees instead of logistic regressions, on bias issues in related applications. This might subsequently affect the best model from a de-biasing perspective. As noted in the introduction, in a credit scoring context, recent empirical research has found more complex models tend to disadvantage minority borrowers in a US context (Fuster et al., 2022). That said, reducing model complexity may not be the best way to achieve debiasing. For instance, Blattner, Nelson, and Spiess (2023), find that targeting specific variables that cause misalignment between accuracy and bias is more effective than only using simpler models.

Exploratory bias analysis

What Veale and Binns (2017) term "exploratory fairness analysis" (in our language, exploratory bias analysis) involves considering how a model was made and its outputs and thinking through whether that process might induce forms of bias. Necessarily, this involves an exercise combining statistics and intuitive judgment, and is not a full replacement for other forms of analysis. However, exploratory bias analysis may be the only option if demographic characteristics or characteristics of vulnerability can be collected for no published datasets in each market. This area is not well established in the literature, but the authors provide two possible examples.

First, clustering individuals together and investigating the bias according to those cluster assignments. Based on other information, it might be possible to roughly proxy for demographic characteristics or characteristics of vulnerability based on other variables. This could give an indication of potential bias issues, but this kind of analysis seems fraught with the potential for inaccurate guesswork, and there has been no empirical validation of whether it does remove bias.

Second, creating interpretable models and thinking through which decision rules might be likely to cause bias, based on research about the relationship between different variables and demographic characteristics or characteristics of vulnerability. This could potentially be more promising, as engaging with the logic of the model is likely to be important for model assurance, and there might be more possibility of spotting obvious cases of rules that cause potential biases.

Therefore, while the non-availability of individual level data containing the model's features, target/outcome variable, and the demographic characteristic (or an accurate proxy) does impair bias mitigation there are partial alternatives around either establishing knowledge fora which may be sector-specific on debiasing and exploratory bias analysis.

5 Conclusion

We have aimed to provide a literature review on AI bias in the context of supervised machine learning. In this review, as a reminder, bias means unjustified differences by demographic characteristics or characteristics of vulnerability in predictions and subsequently in decisions made in either an automated or algorithm-assisted way. The literature shows that this can result from a variety of sources, notably human inputs, data collection practices, or modelling choices. It is important to note that so-called "raw" differences in outcomes for different groups, that is those without adjusting for appropriately justified risk or economic characteristics, may give a misleading impression of bias.

The literature then suggests these questions to consider in supervised machine learning model construction and deployment for bias measurement and mitigation:

- *What demographic characteristics or characteristics of vulnerability could be studied?* This could be based on considering intuitively and possibly with regard to academic research what biases might exist in the model development pipeline.
- *What metric/s could be used to measure any bias?* Based on the current existing literature, one useful metric highlighted in the literature is conditional demographic parity/disparity, because it accounts for other differences between people and focusses on final outcomes. However, choice of metrics requires consideration of context, and this area is ever evolving.
- *Is it feasible and appropriate to collect or proxy data on demographic characteristics or characteristics of vulnerability on an individual level, either for some or all our customers?* Here, there are a range of considerations including data protection law, the accuracy of the data collected, and whether direct or proxy (sex and race only) are the appropriate alternatives if data is collected.
- *Is it feasible to mitigate biases and if so, how?* There are several methods in the pre-, mid-, and post-processing stages to mitigate bias. The literature suggests modellers could consider what is most feasible, and trade-offs with the explainability of any resulting model.
- *If demographic characteristics or characteristics of vulnerability cannot be measured or proxied, are there any realistic alternatives?* We outline two possibilities: creating industry-wide knowledge bases, and exploratory bias analysis. These have the drawback that they do not provide the same guarantees as mitigation techniques where measuring or proxying the actual demographic characteristics or characteristics of vulnerability of consumers is possible, but they do provide some possibilities of addressing bias issues.

For **researchers including academics,** some important considerations are likely to be:

1. Do the strategies that Veale and Binns (2017) propose for debiasing in the absence of data on demographic characteristics or characteristics of vulnerability work (or other such plausible strategies)? One possible strategy for testing this would be first to construct a suitable knowledge forum, and then conduct a

randomised controlled trial on the levels of bias from models built by developers with and without access to that knowledge forum.

2. Information on practical case studies of debiasing models and decision processes, especially beyond the well-studied case of credit scoring to include cases like general insurance and retail investments.
3. Detailed studies on the processes that might lead to bias within the UK financial services context, and evidence for the impact of those biases on datasets. At present, a significant amount of the AI literature uses American datasets, which come from different historical circumstances and practices of exclusion.
4. What characteristics should be included to account for differences between persons to ensure bias measures compare "like-for-like" persons both from a social scientific and ethical perspective?

Annex 1 - References

Angelova, V., Dobbie, W., and Yang, C.S. (2023). Algorithmic Recommendations and Human Discretion. *NBER Working Paper*, no. 31747. doi: https://doi.org/10.3386/w31747.

Balayn, A. and Gürses, S. (2021). Beyond Debiasing: Regulating AI and its inequalities. [online] Available at: https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine learning.* Cambridge, MA: MIT Press.

Bayer, P., Ferreira, F., and Ross, S. (2014). Race, Ethnicity, and High-Cost Mortgage Lending. *NBER Working Paper,* no. 20762. doi: https://doi.org/10.3386/w20762.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. doi: https://doi.org/10.48550/arXiv.1706.02409.

Blattner, L., and Nelson, S. (2021). How Costly is Noise? Data and Disparities in Consumer Credit. [online]. Available at: https://arxiv.org/pdf/2105.07554.pdf.

Blattner, L., Nelson, S., and Speiss, J. (2023). Unpacking the Black Box: Regulating Algorithmic Decisions. [online]. Available at: https://arxiv.org/pdf/2110.03443.pdf.

Bloomberg (2016). *Amazon Doesn't Consider the Race of Its Customers. Should It?.* [online]. Available at: https://www.bloomberg.com/graphics/2016-amazon-same-day/.

Bucher-Koenen, T., Hackethal, A., Koenen, J., and Laudenbach, C. (2021). Gender differences in financial advice. [online]. *Leibniz Institute for Financial Research SAFE Working Paper Series*. Available at: https://ideas.repec.org/p/zbw/safewp/309.html.

Buolamwini, J., and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine learning Research: 2018 Conference on Fairness, Accountability, and Transparency*, pp. 1-15.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. A., and Cosentini, A.C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12. doi: https://doi.org/10.1038/s41598-022-07939-1.

Chang, H., and Shokri, R. (2021). On the Privacy Risks of Algorithmic Fairness. *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, Vienna, Austria, 2021, pp. 292-303, doi: https://doi.org/10.1109/EuroSP51992.2021.00028.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. [online]. Available at: https://www.andrew.cmu.edu/user/achoulde/files/disparate_impact.pdf.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020). Fair regression with Wasserstein Barycenters. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, pp. 1-11.

Cohen, M.C., Elmachtoub, A.N., and Lei, X. (2020). Price Discrimination with Fairness Constraints. *Management Science*, 68(12): 8536-8552. doi: https://doi.org/10.1287/mnsc.2022.4317

Corrales-Barquero, R., Marín-Raventós, G. and Barrantes, E.G., 2021. A Review of Gender Bias Mitigation in Credit Scoring Models. *2021 Ethics and Explainability for Responsible Data Science (EE-RDS)*, pp.1-10. doi: https://doi.org/10.1109/EE-RDS53766.2021.9708589

Cowgill, B., and Tucker, C.E., 2019. Economics, fairness, and algorithmic bias. *preparation for: Journal of Economic Perspectives*.

Department for Science, Innovation, and Technology. (2022). A pro-innovation approach to AI regulation. [online]. Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. and Venkatasubramanian, S., 2018, January. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability, and transparency* (pp. 160-171). PMLR.

Equality Act 2010. Available at: https://www.legislation.gov.uk/ukpga/2010/15/contents.

FCA (2021). FG21/1 Guidance for firms on the fair treatment of vulnerable customers. [online]. Available at: https://www.fca.org.uk/publication/finalised-guidance/fg21-1.pdf.

FCA (2022). PS22/9 A new Consumer Duty. [online]. Available at: https://www.fca.org.uk/publication/policy/ps22-9.pdf.

FCA (2023). CP23/20 Diversity and inclusion in the financial sector – working together to drive change. [online]. Available at: https://www.fca.org.uk/publication/consultation/cp23-20.pdf.

FCA (2024). AI Update. [online]. Available at: https://www.fca.org.uk/publications/finalised-guidance/guidance-firms-fair-treatment-vulnerable-customers.

Feast, J. (2020). Root Out Bias at Every Stage of Your AI-Development Process. [online]. Available at: https://hbr.org/2020/10/root-out-bias-at-every-stage-of-your-ai-development-process.

Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S. (2015),. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259-268). doi: https://doi.org/10.48550/arXiv.1412.3756

Ferrara, E. (2024). The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications*, 15. https://doi.org/10.1016/j.mlwa.2024.100525.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2022). Predictably Unequal? The Effects of Machine learning on Credit Markets. *The Journal of Finance*, 77: 5-47. https://doi.org/10.1111/jofi.13090.

Frontier Economics. (2021). The economic impact of trust in data ecosystems. [online]. Available at: https://theodi.org/insights/reports/the-economic-impact-of-trust-in-data-ecosystems-frontier-economics-for-the-odi-report/

Gabriel, I. 2022. Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, 151 (2): 218–231. https://doi.org/10.1162/daed_a_01911.

Ghosh, B., Basu, D., and Meel, K.S. 2023. "How Biased are Your Features?": Computing Fairness Influence Functions with Global Sensitivity Analysis. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, pp. 138–148. https://doi.org/10.1145/3593013.3593983.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). https://doi.org/10.1609/aaai.v32i1.11296.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc.

ICO. (2023). *What about fairness, bias, and discrimination?*. [online]. Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/what-about-fairness-bias-and-discrimination/#:~:text=to%20your%20problem?-,How%20do%20bias%20and%20discrimination%20relate%20to%20fairness?,of%20one%20solution%20over%20another.

Kamiran, F., and Calders, T. (2009). Classifying without discriminating. *Proceedings of the 2009 2nd International Conference on Computer, Control and Communication*, Karachi, Pakistan, 2009, pp. 1-6, doi: 10.1109/IC4.2009.4909197.

Kamiran, F., and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge Information Systems*, 33: 1-33.

Kamiran, F., Karim, A., and Zhang, X. (2012). Decision Theory for Discrimination-Aware Classification. *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, Brussels, Belgium, pp. 924-929. doi: https://doi.org/10.1109/ICDM.2012.45.

Kamiran, F., Žliobaitė, I., and Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems* 35(3): 613-644. doi: https://link.springer.com/article/10.1007/s10115-012-0584-8.

Kingsman, N., 2021. Debiasing Credit Scoring using Evolutionary Algorithms. *arXiv preprint arXiv:2110.12838*.

Korbmacher, M. et al. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology* 1(3). https://doi.org/10.1038/s44271-023-00003-2.

Kusner, M., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. *30th Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA. doi: https://doi.org/10.48550/arXiv.1703.06856

Lee, M.S.A., and Floridi, L. (2021). Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs. *Minds & Machines* 31: 65–191. doi: https://doi.org/10.1007/s11023-020-09529-4.

Leslie, D., Rincón, C., Perini, A., Jayadeva, S., Borda, A., Bennett, SJ. Burr, C., Aitken, M., Katell, M., Fischer, C., Briggs, M., Wong, J., and Kherroubi Garcia, I. (2023). *AI Fairness in Practice*. [online] The Alan Turing Institute. Available[at: https://www.turing.ac.uk/sites/default/files/2023-11/ai-fairness.pdf.

Liu, M., Lei, D., Dengdeng, Y., Wulong, L., Linglong, K., and Bei, J. (2022). Conformalized Fairness via Quantile Regression. *Proceedings of 35th Advances in Neural Information Systems Processing (NeurIPS 2022),* New Orleans, pp. 11561-11572. Available at: https://proceedings.neurips.cc/paper_files/paper/2022/file/4b52b3c50110fc10f6a1a86055682ea2-Paper-Conference.pdf.

Lum, K., Zhang, Y., and Bower, A. (2022). De-biasing bias measurement. *Proceedings of 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22),* June 21-24, 2022, Seoul, Republic of Korea. Available at: https://arxiv.org/abs/2205.05770. doi: https://doi.org/10.48550/arXiv.2205.05770.

Mary, J., Calauzènes, C., and El Karoui, N. (2019). Fairness-Aware Learning for Continuous Attributes and Treatments. *Proceedings of the 36th International Conference on Machine learning*, Long Beach, California, PMLR 97. Available at: https://proceedings.mlr.press/v97/mary19a/mary19a.pdf.

Molnar, C. (2022). *Interpretable Machine learning: A Guide for Making Black Box Models Explainable* (2nd Edn). Self-published, available at: https://christophm.github.io/interpretable-ml-book/.

Rodolfa, K.T., Lamba, H., and Ghani, R. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature: Machine Intelligence*, 3, 896–904 (2021). doi: https://doi.org/10.1038/s42256-021-00396-x.

RTA. (2020). A review into bias in algorithmic decision-making. [online]. Available at: https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making.

RTA. (2023). CDEI publishes research on AI governance. [online]. Available at: https://www.gov.uk/government/publications/cdei-publishes-research-on-ai-governance.

Starks, M., Reynolds, G., Gee, C., Burnik, G., and Vass, L. (2018). Price discrimination in financial services: How should we deal with questions of fairness?. [online]. *FCA Research Notes.* Available at: https://www.fca.org.uk/publication/research/price_discrimination_in_financial_services.pdf.

Singh, J., Singh, A., Khan, A. and Gupta, A. (2021). Developing a novel fair-loan-predictor through a multi-sensitive debiasing pipeline: DualFair. *arXiv preprint arXiv:2110.08944*.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the international workshop on software fairness* (pp. 1-7). doi: https://doi.org/10.1145/3194770.3194776.

Wachter, S., Mittelstadt, B., and Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review,* 41: 1-72. doi: https://doi.org/10.1016/j.clsr.2021.105567. Pre-print: https://arxiv.org/abs/2005.05906.

Wetherell, S. (2020). [online]. *Redlining the British City*. Available at: https://renewal.org.uk/wp-content/uploads/2020/10/renewal28.2_13wetherell.pdf

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2). https://doi.org/10.1177/2053951717743530.

Zhang, Y. (2016). Assessing Fair Lending Risks Using Race/Ethnicity Proxies. *Management Science*, 64(1). doi: https://doi.org/10.1287/mnsc.2016.2579.

Zhang, B.H., Lemoine, B. and Mitchell, M., 2018, December. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335-340).

Zhang, Y., and Long, Q. (2021). Assessing Fairness in the Presence of Missing Data. *Advances in Neural Information Processing Systems*, 34: 16007-16019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9043798/

Annex 2 – Glossary

Alternative World Index: a fairness metric that can handle multiple demographic characteristics or characteristics of vulnerability, which works by taking a count of biased data points.

Bias-accuracy trade-off refers to the observed phenomenon where efforts to reduce bias or increase fairness in a machine learning model can sometimes lead to a decrease in the model's overall accuracy. This could occur if debiasing removes genuinely predictive relationships between features and the outcome.

Binary Classification: type of statistical task where an algorithm is trained to categorize data into one of two distinct classes or groups.

BISG: Bayesian Improved Surname and Geocoding is a methodology that combines geocoded address and surname as a way of inferring race when self-reported data is unavailable.

Black-Box Models: computational algorithms whose internal workings are not transparent or easily understood by the user, making it difficult to interpret how inputs are transformed into outputs.

RTA: Responsible Technology Adoption Unit, previously called Centre for Data Ethics and Innovation (CDEI), a UK government body that aims at supporting innovation in both public and private sectors by developing tools, guidance, and standards for AI and data-driven technology, ensuring they perform as expected and fostering public trust in their use.

Conditional Demographic Parity: it is a variation of demographic parity that aims for equal probability of positive predictions across groups, given certain legitimate attributes. For example, the acceptance rates of a loan application for men and women are equal given their income.

Debiasing: refers to the application of select methods to address bias by achieving certain forms of approximate statistical parity.

Demographic parity: also known as Independence, is a fairness definition that states that the proportion of each segment of a protected class (e.g., sex) should receive a positive outcome at equal rates. As an example, it would ensure that the acceptance rates of a loan application for men and women are equal.

Disparate Impact Remover: a pre-processing technique that edits feature values to increase group fairness while preserving individuals ranking within their group.

Distributive justice: is concerned with how goods, honours, and obligations are distributed within a community. Distributive claims can be (and have been) justified based on the need for particular essential goods, or on the moral standing of human beings, as well as numerous other bases.

DualFair: a pre-processing technique that can handle debiasing when multiple demographic characteristics or characteristics of vulnerability are involved through oversampling and under sampling techniques that target root causes of bias.

Equality Act: The Equality Act 2010 legally protects people from discrimination in the workplace and in wider society. It replaced previous anti-discrimination laws with a single Act, making the law easier to understand and strengthening protection in some situations. It sets out the different ways in which it is unlawful to treat someone.

Fairness through Unawareness: a debiasing technique whereby the sensitive characteristic is excluded from the model. However, there may be proxy variables or historical biases in the data that would still bias it.

Feedback loop: refers to a cycle whereby AI learns from human bias (e.g., historically biased data), from which these biased outputs affect human decision-making and outcomes, which are then fed back into the system.

ICO: Information Commissioner’s Office is the UK's independent body set up to uphold information rights.

IMD: the Index of Multiple Deprivation are small area measures of relative deprivation across each of the constituent nations of the United Kingdom.

In-processing: debiasing techniques that modify the algorithms during model training.

Large Language Models: is a type of artificial intelligence (AI) algorithm that uses deep learning techniques and large data sets to understand, summarize, generate, and predict text-based content.

Post-processing: debiasing techniques applied after a machine learning model has made its predictions, to adjust these predictions or decision thresholds to correct for biases and ensure fairness across different groups defined by protected attributes.

Predictive algorithm: models used to predict future events or outcomes by analysing patterns in a given set of input data.

Pre-processing: debiasing techniques applied prior to modelling through data transformation.

Procedural justice: fairness of processes used by those in positions of authority to reach specific outcomes or decisions.

Protected Characteristic: refers to a trait or feature of an individual that is legally protected against discrimination and unfair treatment in the Equality Act 2010. These characteristics commonly include race, sex, age, disability, religion, and sexual orientation.

Proxy variable: an indirect measure or substitute for a variable of interest that cannot be directly observed or measured.

Regularisation: an in-processing technique that introduces a range of either implicit or explicit penalties to the optimisation problem, which here is balancing accuracy with fairness, through adjustable weights.

Reject Option Classification: a post-processing technique that identifies and withholds decisions on instances near the classification boundary where prediction uncertainty—and potential bias—is highest. Thus, it temporarily defers the decision and potentially subjecting them to further human review.

Selection Bias: occurs when individuals or groups in a study differ systematically from the population of interest leading to a systematic error in an association or outcome.

Separation: a fairness definition that states that the both the true positive rate and the false positive rate are similar across different groups within a protected class. For example, the proportion of men and women who will not default (non-defaulters) and are correctly predicted to be accepted for loans (true positives) is similar and the proportion of men and women who will default (defaulters) but are incorrectly predicted to be accepted for loans (false positives) is also similar.

Simpson’s Paradox: a statistical phenomenon where an association between two variables in a population emerges, disappears, or reverses when the population is divided into subpopulations.

Sufficiency: a fairness definition that focuses on Positive Predictive Value and Negative Predictive Value should be consistent across groups, focusing on the reliability of the model's positive and negative predictions. For example, for individuals predicted by the model as low risk (likely to repay a loan), the actual proportion of loan repayment (those who truly repay) must be similar for both men and women, and similarly, for those predicted as high-risk (likely to default), the actual proportion of defaults should also be similar across sexes.

Supervised Machine learning: a subcategory of machine learning and artificial intelligence that uses labelled datasets to train algorithms to classify data or predict outcomes accurately.

12 Endeavour Square, London E20 1JN
Telephone: +44 (0)20 7066 1000 Website: www.fca.org.uk