International Statistical Review

International Statistical Review (2022), **90**, 3, 468–480 doi: 10.1111/insr.12492

Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms

Nengfeng Zhou, Zach Zhang, Vijayan N. Nair, Harsh Singhal and Jie Chen

Corporate Model Risk, Wells Fargo, San Francisco, California, USA
Email: nengfengzhou@gmail.com

Summary

The advent of artificial intelligence (AI) and machine learning algorithms has led to opportunities as well as challenges in their use. In this overview paper, we begin with a discussion of bias and fairness issues that arise with the use of AI techniques, with a focus on supervised machine learning algorithms. We then describe the types and sources of data bias and discuss the nature of algorithmic unfairness. In addition, we provide a review of fairness metrics in the literature, discuss their limitations, and describe de-biasing (or mitigation) techniques in the model life cycle.

Key words: AI; algorithm; bias; fairness; ML.

1 Introduction

Artificial intelligence (AI) techniques are used increasingly in many areas of applications, including banking and finance (Dixon *et al*. 2020). They have several advantages over traditional statistical methods: (i) ability to handle new data types such as text, audio, and images; (ii) flexible models that yield excellent predictive performance; and (iii) ability to automate many of the routine, and time-consuming, tasks in model development. However, the use of these algorithms also raises several challenges. A well-known problem is the opaqueness of machine learning (ML) models and the difficulties in understanding and interpreting the model results (Arrieta *et al*. 2020; Hu *et al*. 2021; Lundberg and Lee 2017; Ribeiro *et al*. 2016 and Vaughan *et al*. 2018). In this paper, we focus on a related and equally important challenge: potential for bias and lack of fairness when using AI/ML techniques. This is currently a 'hot topic' with increasing number of publications, conferences and discussions (refer to, e.g., Barocas *et al*. 2020; Caton and Haas 2020 and Mehrabi *et al*. 2021). While the definitions of discrimination and fairness depend on the application of interest, the most relevant ones for our purpose are those associated with protected groups (Dwork *et al*. 2012; Verma and Rubin 2018).

There are regulatory requirements in banking aimed at preventing discrimination. For example, the Fair Housing Act (FHA 1968) and Equal Credit Opportunity Act (ECOA 2017)

The views expressed in the paper are those of the authors and do not represent the views of Wells Fargo.

prohibit unfair and discriminatory practices based on protected attributes. ECOA explicitly mentions nine categories: race, colour, religion, sex, national origin, age, marital status, receipt of public assistance, or any right exercised under the Consumer Credit Protection Act. Financial institutions can be liable for the following:

a Disparate treatment: This refers to differential treatment of members of a protected group compared with others, after taking into account relevant factors in the decision process. Disparate treatment (Charles River Associates. n.d. 2016) usually happens in decisions based on judgement where the protected variable may be used explicitly or implicitly by the human decision makers; and
b Disparate impact: This deals with differential (or adverse) impact of a seemingly neutral decision or policy on members of a protected group compared with others. Disparate impact is mainly a concern in model-based decision-making. Even if protected variables are not used in a model directly, it is possible that other variables may serve as their proxies. Note, however, that the Consumer Financial Protection Bureau (Klein 2019) has an important qualification to this rule: disparate impact will not create a violation if the necessity can be justified and no alternative decision or policy can have comparable performance with less discriminatory effect.

The main difference between disparate treatment and disparate impact is the following: a protected variable is used directly during decision making in disparate treatment, while it is not used in disparate impact. Traditional statistical methods have been used to test for disparate treatment in redlining cases and for checking consistency in pricing and underwriting (Charles River Associates. n.d. 2016; Iacus *et al*. 2011; Seventh Circuit 1994). Testing for disparate impact, however, is a multistep process that involves determining if a protected class has been adversely affected, then looking for a justification for that specific policy or practice, and finally searching any alternatives that would result in less impact.

Fairness issues arise in banking and finance, criminal justice, social programmes, healthcare, recruiting, marketing and so on. One study has received national attention deals with recidivism of criminal defendants (Larson *et al*. 2016). An analysis of the popular COMPAS algorithm (Correctional Offender Management Profiling for Alternative Sanctions) found that there was a higher false positive rate in identifying black defendants to be at risk of recidivism compared with white defendants; conversely, there were lower false negatives for flagging whites as low risk compared with blacks. The article at the above link cites several other examples of discrimination in social/criminal justice. Refer also to the draft book (Barocas *et al*. 2020) for examples as well as state-of-the-art research discussion in this area. Hutchinson and Mitchell (2019) provide a good discussion of issues within education and hiring. One example that specifically documents fairness issue with ML algorithms (bias against women) is the Amazon's recruiting tool. Refer also to the recent book, *The Ethical Algorithm* (Kearns and Roth 2019), for other examples and easy-to-read discussion of research developments.

Within the banking industry, AI/ML techniques are being used in consumer lending (credit scoring and more recently marketing and collections), conduct analysis and compliance management. Fair-lending considerations are at the forefront in credit scoring and decisioning. Conduct analysis and compliance management rely on analysis of unstructured data (text, e-mails, etc.) using natural language processing and ML techniques. These are newer applications areas that also have significant potential for bias and unfairness.

The rest of the paper is organised as follows. Section 2 provides a brief overview on AI/ML algorithms and specifies those of interest to the discussion in this paper. This is followed by a description of the sources and types of bias and fairness issues (Section 3). Section 4 provides a review of fairness metrics in the literature and their limitations. Section 5 discusses de-biasing

(or mitigation) approaches and how they can be used at different stages of the model lifecycle. Section 6 outlines some general guidelines on ensuring fairness.

2 Scope of AI/ML Algorithms

While the phrases *Artificial Intelligence* and *Machine Learning* are often used interchangeably in popular discussion, they are quite different. As noted in Hu *et al*. (2021) and elsewhere, AI is much broader in scope and ML is just one of the pathways to accomplishing the goals of an AI project. AI has a very long history dating back to formal reasoning in logic, philosophy and other fields. The term itself was coined by John McCarthy only in 1959 as

> The study of "intelligent agents"—devices that perceive the environment and take actions that maximize its chance of success at some goal.

Artificial intelligence has had mixed successes in the past and has gone through periods of 'AI winters'. There has been a massive resurgence recently, due to the amounts of data available and exponential leaps in computing power.

The term *Machine Learning* was proposed by Arthur Samuel in 1959. An engineering-oriented definition was given by Tom Mitchell in 1997 as follows:

> A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

It was not until the last two to three decades that there has been wide usage of ML techniques. The main reasons again are increased capabilities in data storage, data transfer, data architecture and fast computing. Refer to Hu *et al*. (2021) for further discussion and references.

Some people use the term ML very broadly to include even traditional statistical methods such as linear regression and clustering, techniques that have been around for a long time. We restrict usage to modern approaches for supervised, unsupervised, and reinforcement learning, including support vector machines, ensemble algorithms (random forest & gradient boosting), and various types of neural networks (feedforward, CNN, RNN, LSTM, GAN, auto-encoders, isolation forest and so on). In particular, the present paper restricts attention to supervised machine learning techniques for regression, classification and prediction. Supervised machine learning models are based on very flexible nonparametric algorithms that can learn complex patterns in the data. As such, they are very vulnerable to inherent biases in the data. Moreover, because the fitted models are complex and hard to understand, it is difficult to appreciate biased decisions that arise in using the model.

3 Potential Sources of Bias and Discrimination

Fairness considerations have been around for a long time. For example, Courchane *et al*. (2000) describe the results of a case study on fair-lending examinations of national banks from 1994 to 1999 and note that despite 'years of intense scrutiny, lending discrimination still persists'. The arrival of flexible and automated AI/ML algorithms as well as the availability of alternative sources of data are leading to new challenges and exacerbating existing ones.

We group potential sources of bias and fairness issues into two broad areas, data bias and algorithmic bias, and discuss each of them below. Refer also to Mehrabi *et al*. (2021) for a comprehensive survey of sources of bias and types of discrimination.

3.1 Data Bias

a. Bias in historical data: Historical data are often skewed towards, or against, particular groups. Data can also be severely imbalanced with limited information on protected groups. The literature is full of examples where historical biases have resulted in discrimination. Recall the COMPAS study mentioned earlier. Biases in historical data can be exacerbated with the use of ML techniques for the following reasons. Users tend to include a large number of input variables in the model because ML algorithms can automate feature engineering. Further, time-consuming tasks such as checking for multi-collinearity and variable selection are not needed for prediction problems. As a result, one may not detect biases in the data nor the presence of surrogate variables for protected attributes. Recall the earlier example on Amazon's recruiting tool that resulted in bias against women because historical data were biased towards men who dominated the technology industry.
b. Bias in data collection mechanisms: Advances in data capture technologies have made it easier to collect different types of data. But insufficient attention is being paid to inherent biases in the data collection mechanisms and lack of representativeness. A well-known example is 'crowdsourcing' of pothole information where smart phones are used to automatically sense and transmit data when a vehicle goes over potholes (Crawford 2013). This mechanism is clearly biased towards the segment of the population that owns smart phones and the roads they use. There are many other examples in the literature, including development of marketing campaigns based on the usage of the internet, and use of video imaging and facial recognition systems that have better performance for people with lighter skin tones and have higher error rates in other cases.
c. Bias in alternate sources of data: Much of the excitement with the big data phenomenon is due to new sources of readily available data: worldwide web, social media, blogs and so on. A recent paper (Berg *et al*. 2018) demonstrates problems with such alternate sources of data in the context of credit scoring. They show that 'digital footprint' variables, such as borrower's computer (Mac vs. PC), type of device (phone, tablet & PC), whether name is part of the borrower's e-mail, have strong predictive performance in credit scoring. However, these predictors are highly correlated with socio-economic variables that are surrogates for protected groups. Another example discussed in Klein (2019) is data on divorce proceedings that are good predictors of potential bankruptcy. More often than not, these variables are highly related to protected classes, and the availability of such seemingly innocuous information, combined with flexible 'data snooping' ML algorithms, can easily lead to 'proxy discrimination'.
d. Unobservable outcomes: Corbett-Davies and Goel (2018) discuss measurement problems (bias) associated with labels/responses and predictors. Unobservable outcomes (or labels) are a huge issue, one that makes it difficult to even measure discrimination. Consider an example with mortgage applications where the lender uses a model to decide on approving or denying the loan. In term of 'true' outcomes, we observe loan defaults only for those who received a mortgage and we do not have any information for those who were denied mortgage—whether they would have defaulted or not had they been approved for a mortgage (a counterfactual). Thus, we have true outcomes (default or not) for only part of the population. (Reject inference is a technique used in banking to try to address this problem.) The same issue arises in job applications when we use a model or even interviews to decide on job offers. True outcomes (job performance that measure whether candidates are qualified or not) are available only for those who actually got the job. In the absence of information on all outcomes, one cannot even measure fairness.

e. Bias in unstructured data and feature engineering: Unstructured data, such as texts, audio, and images, are increasingly analysed through AI/ML techniques in banking and finance. In these applications, one converts data summaries and selected features, such as word embeddings, for analysis. However, the feature selection process can suffer from 'implicit biases'. Figure 1 shows word embeddings that suffer from gender bias. A commonly quoted example of a word embedding is 'Man is to Computer Programmer as Woman is to Homemaker' (Bolukbasi *et al*. 2016).

3.2 Algorithmic Bias

f. The automated nature of modern ML algorithms presents its own challenge. Model developers in banks are using data at account levels (millions of observations) with thousands of predictors. In automated use of ML, the algorithms artificially create hundreds of derived predictors (by transforming the original ones) with the hope of getting tiny improvements in predictive performance. However, such processes do not incorporate subject-matter knowledge to carefully review the selected variables and miss the potential for correlated surrogate variables causing proxy discrimination.

g. The flexible nature of the algorithms can lead to overfitting of training data. In fact, one could ask whether the increased predictive performance is in fact good model fitting or is its good memory. The algorithms can 'remember' patterns in the data and even individual observations, leading to concerns of privacy as well as fairness. There are techniques for addressing the overfitting problem including the use of test/validation data and regularisation methods. However, these approaches do not fully address the problem. An interesting anecdote relates to how *Baidu* tried to get a 'leg up' in a Kaggle competition by using several e-mail accounts to get multiple views of the validation dataset (https://www.technologyreview.com/s/538111/why-and-how-baidu-cheated-an-artificial-intelligence-test/).

h. Optimization of ML algorithms can also lead to biases. The primary objective in hyperparameter tuning and optimization is maximising predictive performance. In doing so, one ignores other important issues such as robustness and fairness. Unlike traditional parametric models, ML algorithms are so flexible that they can end up amplifying small biases in the data (refer to Hooker 2021).

i. Data bias together with poor optimization of algorithms can cause severe harm to protected groups, especially when information on such groups is under-represented in the training data

Figure 1. *Example of gender bias in word embedding (Buonocore 2019).*

(refer also to Hooker 2021). In such situations, the results from ML algorithms can be inaccurate or the model can be very unstable.

j. A related concern is the opaqueness and lack of interpretability of complex ML algorithms. For traditional models, which are typically global, one can measure fairness in a global fashion. Many tree-based ML algorithms, such as RF and GBM, are inherently local, and a different approach is needed to measure fairness locally. If one can identify the input–output relationships, including complex interactions and local behaviour, one can isolate potential algorithmic bias. There are techniques available for exploring such relationships in the context of ML algorithms relationships (refer to Goldstein *et al*. 2015). However, these tools are currently limited in their ability to tell the full story. Thus, one may not be able to identify the presence or reasons for discrimination.

4 Fairness Metrics

An important challenge in bias and fairness discussions is how to define fairness. There is no universally accepted definition, and researchers have tried to approach the problem from different angles. Figure 2 shows common and widely used metrics to assess model or algorithmic fairness. We discuss them in detail below (refer also to Mitchel *et al*. 2019) for a survey of different categories 'fairness' related to ML and statistical model-based predictions.

a Group fairness

We start with a definition of relevant notation and use the recidivism example for illustration.

- $Y \in \{0, 1\}$: binary response variable (whether a prisoner will recidivate or not after he is released);
- $\widehat{Y} \in \{0, 1\}$: binary prediction variable (result of an algorithm used by, say, a parole board, to predict recidivism);
- $X : p -$ dimensional feature associated with the defendant, for example, education, work experience, and past criminal history; and
- $A \in \{0, 1\}$: a binary protected attribute, for example, gender (in general, it can be multi-dimensional and take on multiple values).

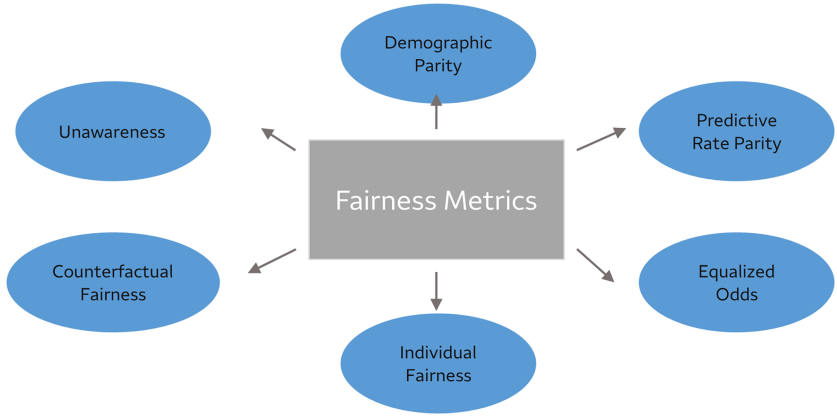

Figure 2. *Different fairness metrics.*

(Note: In this recidivism example, if the parole board used the prediction on recidivism to decide on parole, it might in fact end up affecting the true recidivism outcome. We assume that this is not the case here, but this can be a problem in practice.)

Here are some common group-fairness metrics:

i *Demographic parity (statistical parity)* aims to ensure that probability of recidivating is equal across the sensitive attribute (Dwork *et al*. 2012):

$$P\left[\widehat{Y} = j \middle| A = 0\right] = P\left[\widehat{Y} = j \middle| A = 1\right], j \in \{0, 1\}.$$

Demographic parity is the independence of the decision and the protected attribute. In our recidivism example, this means that the predicted recidivism rates for men and women must be equal.

There are variations that enforce it subject to constraints such as $P\left[\widehat{Y} = 1 \middle| A = 1\right] \leq \text{const} \times P\left[\widehat{Y} = 1 \middle| A = 0\right]$). Hardt *et al*. (2016) noted that, while this metric is 'simple and seemingly intuitive, it has many conceptual problems' (refer also to Dwork *et al*. 2012). For instance, women are known to have lower recidivism rates compared with similarly situated men. By enforcing demographic parity, we may be treating women unfairly.

ii *Predictive rate parity* means the predictive value should be equal for the protected and unprotected class (Verma and Rubin 2018):

$$P[Y = k|\widehat{Y} = j, A = 0] = P[Y = k|\widehat{Y} = j, A = 1], k, j \in \{0, 1\}$$

In other words, Y is independent of A conditional on $\widehat{Y}$. In the recidivism example, this metric implies that, conditional on predicted recidivism, the true recidivism rate must be equal for men and women.

iii *Equalised odds* require that the false positive rates and false negative rates should be equal for the protected and unprotected class (Hardt *et al*. 2016; Zhong 2018).

$$P\left[\widehat{Y} = j \middle| Y = k, A = 0\right] = P\left[\widehat{Y} = j \middle| Y = k, A = 1\right], k, j \in \{0, 1\}$$

Equalised odds mean that, conditional on the outcomes, the decision and the protected attribute are independent. In our example, for the algorithm to be fair, the predicted recidivism probabilities must be the same for men and women across all classes.

iv *Equal opportunity* is a special case of equalised odds where the equality is satisfied only for the special case = j, where j is the 'advantageous' outcome (Hardt *et al*. 2016).

$$P\left[\widehat{Y} = j \middle| Y = j, A = 0\right] = P\left[\widehat{Y} = j \middle| Y = j, A = 1\right]$$

This is a weaker condition than equalised odds where the equality is enforced only for the 'advantageous' outcome. Equalised opportunity implies independence of the decision and the protected attribute conditional on the 'advantageous' outcomes.

v *Conditional parity* measures fairness conditional on the values of some appropriate variables (Ritov *et al*. 2017). For example, in fair lending, it is reasonable to assess fairness conditionally on legitimate credit characteristics such as credit worthiness (FICO score). To define the metric, let U denote information that is being conditioned on. For illustrative purpose, assume U takes on discrete values. Then, we have

$$P\left[\widehat{Y} = j \,\middle|\, U = k, A = 0\right] = P\left[\widehat{Y} = j \,\middle|\, U = k, A = 1\right], k \in \{0, 1, ..K\}, j \in \{0, 1\}.$$

Refer also to Zhang *et al*. (2018) and Mehrabi *et al*. (2021) for a comprehensive survey of fairness metrics in machine learning.

b Individual fairness

This concept states that 'similarly situated individuals' should be treated similarly (Verma and Rubin 2018). One way of enforcing this is to (i) define a suitable distance metric $d(x, y)$ on features associated with any two individuals x, y and a distance on the decision space $M(x)$ and $M(y)$; and (ii) ensure that if the distance between two individuals is small, the corresponding distance between their decision distributions is small (within a distance of some multiple of $d(x, y)$). This is related to a Lipschitz condition on the decisions of a classifier (Dwork *et al*. 2012; Gupta and Kamble 2021). Of course, the challenge is to identify meaningful and non-controversial distance metrics.

Counterfactual fairness is a special individual fairness metric (Kusner *et al*. 2017), which is related to the area of counterfactual modelling that arises in causal inference. Counterfactual fairness was proposed in Russell *et al*. (2017). In this case, one would consider a counterfactual scenario where (in our recidivism example) the defendant had a different protected attribute—say female instead of male. If we would have made the same decision regardless of gender, then the decision is considered to be fair.

c Unawareness (anti-discrimination)

This refers to situations where the protected attribute(s) are explicitly removed from the data before the model is trained. In other words, the model is 'not aware' of the sensitive attribute(s) during training. This metric is easy to implement and use. But the obvious limitation is that information in sensitive attributes is often present in surrogate variables (proxy discrimination). For example, it is known that zip codes could serve as proxies for race. This problem gets particularly acute with large datasets where such information is present but hard to detect due to the use of automated algorithms for variable selection. For this and other reasons, some researchers in the literature have argued that it is better to collect and appropriately use information of protected attributes rather than not have access to them.

4.1 Limitations of Fairness Metrics

The concept of demographic parity has been around for a long time. This, together with equalised odds and predictive rate parity, has been discussed by many authors. Interestingly enough, these metrics are in conflict with each other, and there is an 'impossibility theorem of fairness' that states that any two of the three criteria are mutually exclusive (refer to Chouldechova 2017 and Kleinberg *et al*. 2016). Sam Corbett-Davies (2018) presents the limitations of three popular fairness metrics. The paper shows that requiring unawareness (anti-classification) or equalised odds (classification parity) can hurt the protected groups by

relabelling the model predictions to achieve the fairness, and predictive parity (calibration) is usually not strong enough to guarantee equity.

Hardt *et al*. (2016) describes a case study for FICO model where they measure the effectiveness of different fairness metrics. They found that equalised odds are very difficult to achieve in consumer lending even for the FICO model. It requires both the loan approval rate of non-defaulters and the loan approval rate of defaulters to be constant across groups. This cannot be achieved with a single threshold for each group but requires randomisation of prediction between two thresholds.

Fairness for some decisions is better decided at a group level than at individual levels. For example, in university admissions, many people have argued that a diverse student body will benefit the entire institution.

5 De-biasing and Mitigating Unfairness

Researchers have suggested a number of approaches and techniques for 'de-biasing' or finding ways to mitigate unfairness. Some of these are drastic approaches that suggest, for instance, actively manipulating the data. Their potential is limited due to legal and compliance considerations. Nevertheless, we provide a review of the methods in this section.

Figure 3 shows an overview of the model lifecycle together with corresponding points for bias detection and mitigation efforts. The main stages in a model lifecycle are as follows: (i) data sourcing and pre-processing, (ii) model development that includes variable selection, feature engineering, and model training; (iii) model testing and assessment that includes assessing model fit, use of model diagnostics including model interpretation and explainability; and finally, (iv) model deployment, usage, and monitoring. One should develop and implement strategies to detect and reduce bias at the different stages of the model life cycle. Further, this should be done before the model is deployed and used in production.

Obviously, any bias mitigation effort should start with the data sourcing and pre-processing stage. Some of the techniques suggested in the literature include suppression of sensitive variables and their proxies, data massaging (changing outcome labels to correct data bias), and reweighting/sampling to promote certain instances while demoting others (Kamiran and Calders 2012). A more formal approach, proposed in Calmon *et al*. (2017), involves creating new data points that preserve the distribution of the original data as much as possible and mapping the original data to the newly created data points to preserve demographic parity. This idea covers the model development stage as well because one cannot compute demographic parity without model results.

The above proposals are all fraught with danger and can lead to serious abuses. They are also likely to violate legal and compliance considerations. One should not engage in data massaging,

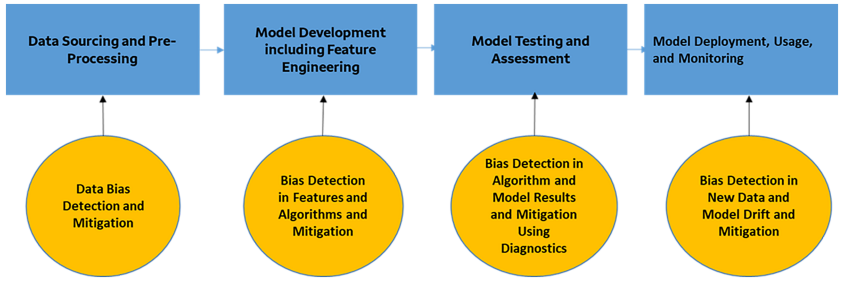

Figure 3. *An overview model life cycle, different types of biases and mitigation efforts.*

suppression or reweighting (as described above) without clearly understanding potential downstream consequences. Unfortunately, it is difficult to accomplish that with large datasets and complex machine learning algorithms. The proposal in Calmon *et al*. (2017) to create and use artificial data is even more subject to potential abuse. In our view, there is no good substitute for carefully reviewing the datasets for historical and measurement biases as well as checking for potential limitations and non-representativeness in data collection mechanisms. These challenges are magnified with large datasets and new alternate sources of data. Further, they go against the spirit of automation promised by AI/ML algorithms.

The second stage for bias detection and mitigation efforts is the 'in-processing' or 'model development/training' stage. One natural approach here is to incorporate the fairness metrics in the estimation process as constraints. This can be a hard constraint where one optimises the decision subject to a suitable fairness metric, say demographic parity. For example, we can optimise credit decisions subject to the given threshold:

$$P\left[\widehat{Y} = j \middle| A = 1\right] \geq \boldsymbol{C}\, P\left[\widehat{Y} = j \middle| A = 0\right],$$

for some constant $0 < C < 1$. Alternatively, we can incorporate it as a soft constraint through regularisation as follows:

$$Min\left\{\sum_i L(y_i, \widehat{y}_i) + \lambda\left(P\left[\widehat{Y} = 1 \middle| A = 1\right] - P\left[\widehat{Y} = 1 \middle| A = 0\right]\right)^2\right\},$$

for suitable loss function L and tuning constant λ. This is the approach taken in several papers in the literature. Kamishima *et al*. (2012) proposed ensuring demographic parity by adding a regulariser in the logistic regression setting (refer also to Zemel *et al*. 2013). Zhang *et al*. (2018) suggest a different approach based on an adversarial objective function along the lines of a GAN network. Celis *et al*. (2018) suggest a framework that uses a set of fairness metrics in a meta-algorithm. Kamishima *et al*. (2012) discuss ways to remove group prejudice using regularisation. Zemel *et al*. (2013) consider the concept of individual fairness and address it together with group fairness (refer also to Agarwal *et al*. 2018; Barrio *et al*. 2020; Calders and Verwer 2009; Woodworth *et al*. 2017 and Zafar *et al*. 2017).

Recall, however, the earlier discussion on limitations and conflicts of the fairness metrics. Further, no single metric of fairness is uniformly better than other metrics. Therefore, the choice of fairness metric has to be done carefully with the particular application as well as appropriate laws and regulations in mind. Another important issue is the trade-off between prediction accuracy and fairness. Pleiss *et al*. (2017) examined the trade-off between minimising error disparity across different population groups while also maintaining calibrated probability estimates.

The next point for bias detection and mitigation efforts is the model testing and assessment stage. In addition to the traditional metrics for model performance, one should also compute the relevant fairness metrics and assess disparate treatment and impact. This may require refitting the model by incorporating fairness thresholds/constraints and refining any existing constraints. At an extreme, this may also require going back to the original data to identify and mitigate data bias. Kamiran and Calders (2012) discuss discrimination-aware classification that does not require data modification or modifying the classification results. This belongs to the 'post-processing' stage. However, the use of such algorithms requires the protected attributes to be available in the model usage stage. Besse *et al*. (2021) provide another example of using post-processing method to mitigate demographic parity.

Another important dimension at this stage is identifying sources of algorithmic bias. This would require the ability to interpret and explain the results from complex ML algorithms.

Finally, once a model has been deployed, it must be continually monitored for disparate impact testing in order to ensure that its predictive performance does not decline. Degradation in model performance can be caused by several factors: changes in the datasets, availability of new data and changes in environment that violates the model's assumptions (model drift). Any new types of bias in the model during production have to be compared with the development data. Model retraining may be needed to ensure that the model continues to be fair and accurate.

The Institute of International Finance survey addressed specific responses and solutions taken by financial institutions to ensure fairness. In addition to technical measures, they propose looking closely at the governance process for ML.

One important issue in these discussions is the availability and usage of protected attributes. In our discussion of the fairness metrics calculations and de-biasing algorithm, we have assumed the protected attributes are available. However, except in mortgage loan applications, banks are not legally allowed to collect protected attribute data when a person is applying for a loan. It is illegal in many areas to use protected attribute for model training or decision. This may prevent the institutions in ensuring fairness because the de-biasing algorithm needs information on protected attributes. There are also situations where protected groups would get better treatment than non-protected groups if the protected attributes were used, such as the example of lower rates of recidivism for women.

6 Concluding Remarks

We close with a discussion of some key points that we have considered in developing AI/ML techniques in banking.

- Certain application areas, such as consumer lending, have potential for serious harm from use of black-box algorithms that are not well-understood. This view appears to be shared by regulators, and we expect application of AI/ML algorithms will be limited in these areas in the near future.
- Fairness concerns are heightened when alternative sources of data, such as social-media data, information on biometrics, speech or language, are used. In these cases, it is not easy to scrub the data of demographic proxies.
- There are multiple approaches to mitigating unfairness concerns. No single approach is universally best, and choosing the most appropriate one will require expert judgement as well as knowledge of relevant legal and compliance requirements.

ACKNOWLEDGEMENTS

We are grateful to the referees and editors for their helpful reviews and suggestions on additional references.

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J. & Wallach, H. (2018). A reductions approach to fair classification. *Proceedings of the 35th International Conference on Machine Learning, PMLR*, (**80**:60-59).

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. & Chatila, R. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion.*, **58**, 82–115.

Barocas, S., Hardt, M. & Narayanan, A. (2020). Fairness and machine learning. https://fairmlbook.org/

Barrio, E., Gordaliza, P. & Loubes, J. (2020). Review of mathematical frameworks for fairness in machine learning. *arXiv*, 2005.13755.

Berg, T., Burg, V., Gombović, A. & Puri, M. (2018). On the rise of FinTechs—credit scoring using digital footprint. NBER Working Paper No. 24551.
Besse, P., Barrio, E., Gordaliza, P., Loubes, J. & Risser, L. (2021). A survey of bias in machine learning through the prism of statistical parity. https://www.tandfonline.com/doi/abs/10.1080/00031305.2021.1952897
Bolukbasi, T., Chang, K., Zou, J., Saligrama, V. & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *arXiv*, 1607.06520.
Buonocore, T. (2019). Man is to doctor as woman is to nurse: the gender bias of word embeddings. https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17
Calders, T. & Verwer, S. (2009). Three naive Bayes approaches for discrimination-free classification. *Data. Min. Knowl. Discov.*, **21**, 277–292.
Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N. & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems 30*.
Caton, S. & Haas, C. (2020). Fairness in machine learning: a survey. *arXiv*, 2010.04053.
Celis, L., Huang, L., Keswani, V. & Vishnoi, N. K. (2018). Classification with fairness constraints: a meta-algorith with provable guarantees. *arXiv.org*.
Charles River Associates. n.d. (2016). Fair lending monitoring: where to focus statistical analysis. ABA Bank Compliance.
Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. https://arxiv.org/abs/1703.00056
Corbett-Davies, S. & Goel, S. (2018). The measure and mismeasure of fairness: a critical review of fair machine learning. https://arxiv.org/pdf/1808.00023.pdf
Courchane, M., Nebhut, D. & Nickerson, D. (2000). Lessons learned: statistical techniques and fair lending. *J. Hous. Res.*, 277–295.
Crawford, K. (2013). The hidden biases in big data. *Harv. Bus. Rev.* https://hbr.org/2013/04/the-hidden-biases-in-big-data
Dixon, M. F., Halperin, I. & Bilokon, P. (2020). *Machine Learning in Finance*. Cham, Switzerland: Springer.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. S. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.
ECOA. (2017). The equal credit opportunity act. The United States Department of Justice. https://www.justice.gov/crt/equal-credit-opportunity-act-3
FHA. (1968). Housing discrimination under the fair housing act. U.S. Department of Housing and Urban Development. https://www.hud.gov/program_offices/fair_housing_equal_opp/fair_housing_act_overview
Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. (2015). Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.*, **24**, 44–65.
Gupta, S. & Kamble, V. (2021). Individual fairness in hindsight. *J. Mach. Learn. Res.*, **22**(144), 1–35.
Hardt, M., Price, E. & Srebro, N. (2016). Equality of opportunity in supervised learning. *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.
Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, 2–4.
Hu, L., Chen, J., Vaughan, J., Aramideh, S., Yang, H., Wang, K., Sudjianto, A. & Nair, V. (2021). Supervised machine learning techniques: an overview with applications to banking. *Int. Stat. Rev.*, 573–604.
Hutchinson, B. & Mitchell, M. (2019). 50 Years of test (un)fairness: lessons for machine learning. Association for Computing Machinery. https://arxiv.org/abs/1811.10104
Iacus, S. M., King, G. & Porro, G. (2011). Causal inference without balance checking: coarsened exact matching. *Polit. Anal.* Advance Access.
Kamiran, F. & Calders, T. (2012). Data processing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 1–33.
Kamishima, T., Akaho, S., Asoh, H. & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50.
Kearns, M. & Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.
Klein, A. (2019). "*Credit Denial in the Age of AI*." A Blueprint for the Future of AI. A series from the Brookings Institution.
Kleinberg, J., Mullainathan, S. & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. https://arxiv.org/abs/1609.05807
Kusner, M., Loftus, J., Russell, C. & Silva, R. (2017). Counterfactual fairness. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (4069–4079).
Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, **30**, 4765-4774.
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, **54**(6), 1–35.
Mitchel, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. (2019). Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions. https://arxiv.org/pdf/1811.07867.pdf
Pleiss, G., Ragahavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. (2017). On fairness and calibration. *NIPS Proceedings*.
Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, (1135–1144).
Ritov, Y., Sun, Y. & Zhao, R. (2017). On conditional parity as a notion of non-discrimination in machine learning. *arXiv abs*, 1706.08519.
Russell, C., Kusner, M. J., Loftus, J. R. & Silva, R. (2017). When worlds collide: integrating different counterfactual assumptions in fairness. *Adv. Neural Inf. Process. Syst.* **30**, 6414–6423.
Sam Corbett-Davies, S. G. (2018). The measure and mismeasure of fairness: a critical review of fair machine learning. https://arxiv.org/abs/1808.00023
Seventh Circuit. (1994). EEOC v. O & G Spring & Wire Forms Specialty Co., **38 F.3d 872**, 874–875. https://casetext.com/case/eeoc-v-o-g-spring-wire-forms-specialty
Vaughan, J., Sudjianto, A., Brahimi, E., Chen, J. & Nair, V. (2018). Explainable neural networks based on additive index models *The RMA Journal*
Verma, S. & Rubin, J. (2018). Fairness definitions explained. *ACM/IEEE International Workshop on Software Fairness*.
Woodworth, B., Gunasekar, S., Ohannessian, M. I. & Srebro, N. (2017). Learning non-discriminatory predictors. *arXiv.org*.
Zafar, M., Valera, I., Rodriguez, M. & Grummadi, K. P. (2017). *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment*. Max Planck Institute for Software Systems (MPI-SWS).
Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. (2013). Learning fair representations. *Proceedings of Machine Learning Research*, 325–333.
Zhang, B. H., Lemoine, B. & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *arXiv.org*.
Zhong, Z. (2018). A tutorial on fairness in machine learning. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

[*Received July 2021; accepted February 2022*]