PROBABILITY & PARTNERS

AI Fairness in Financial Services

How to quantify and improve fairness in machine learning and AI applications?

By Alexandru Giurca

January 2021

www.probability.nl

Summary

Artificial Intelligence and Machine Learning applications are increasingly used in financial services. However, they can exhibit unintentional bias against certain groups of clients, e.g., based on race, age or gender. So it is important that ML algorithms implemented and validated properly, before material decisions can be made with their aid. Failing to do that may lead to financial institutions to be exposed to regulatory risk, but also to reputation damage.

Bias in ML algorithms can arise due to several reasons. Algorithms can incorporate human biases that are reflected in the data that they are trained on – even if sensitive variables such as gender, race, or sexual orientation are removed. These human biases represented in society can infiltrate algorithms along the entire development pipeline – from the data collection and the choice of training data, the algorithm design to its deployment.

To achieve fairness in ML algorithms, first it must be measured. The measurement of fairness starts with recognizing which sensitive clients' attributes can be affected and which definition of fairness one should use. Once the protected attributes and fairness definition are chosen, the algorithm's fairness should be continuously measured through the entire development pipeline: from the data selection stage, its pre-processing to training and testing of the algorithm and its deployment.

If a bias against a protected attribute is found, it can be removed at three places in the development pipeline: debiasing the training data, using fairness constraints during algorithm training or adjusting the algorithm performance to make it fairer when it is applied. The choice of the debiasing method depends on whether one has the access to the training data and the algorithm itself or whether the model is delivered as a black box.

Finally, there is a trade-off between algorithm performance and fairness: mitigating the bias usually leads to some decline in the model's performance. However, with modern debiasing techniques (provided they are properly chosen for the specific use case and the available data), the model performance will not be sacrificed significantly.

Introduction

Artificial intelligence is rapidly adopted in many financial services. From advising on personal wealth management, monitoring user behaviour to underwriting loans, insurance decision making, anti-money laundering and fraud detection[1] – artificial intelligence algorithm's footprint in financial services can be seen everywhere.

However, there is growing evidence that artificial intelligence systems might be biased in ways that may discriminate certain consumers or employees[2]. As policymakers turn their attention to the impact of financial technology on consumers and markets, they guide towards a human centred[3], ethical and fair usage of "Trustworthy Artificial Intelligence" [4]. Firms that deploy artificial intelligence may be exposing themselves to unanticipated risks of discrimination – which not only lead to regulatory risk, but also to tremendous damage of the company's reputation.

Therefore, it is necessary to move beyond the traditional development of artificial intelligence algorithms optimized solely for performance and embed ethical principles in their design, training, and deployment to ensure social good, while still benefiting from the huge potential of the technology. We are now at a critical transition point in the governance of AI ethics, as the focus shifts from formulating ethical principles to setting quantitative metrics and implementing them[5].

In this paper we describe AI fairness from a quantitative perspective, on the example of credit decision making – which candidates should receive a loan and which not – but the principles we will describe hold more generally in financial services. We explore the roots of bias in AI systems and present popular definitions of fairness adopted by the industry. Furthermore, we show where and how to intervene in the AI development pipeline to mitigate bias and minimize the performance-fairness trade-off. Finally, we conclude by providing guidance on quantitative validation of fairness and ethics for AI applications.

1. Algorithmic bias, discrimination and fairness

AI models are not inherently objective. They operate by learning from historical data and generalizing them to unseen data. Since unlike humans, artificial intelligence does not have the gift of morality, the various unwanted consequences of AI algorithms arise from biased data and the way AI algorithms are designed. Data, however, is a product of many factors, from the historical context in which it was generated to the particular forms of measurement errors it contains. The AI development pipeline involves a series of human choices and practices, from the training methodology to model definition and deployment, any of which can lead to unwanted effects.

In this context a **Fairness Bias** in an AI model refers to the "inclination or prejudice of a decision made by an AI system, which is for or against one person or group, in a way considered to be unfair" [5].

"85% of AI projects will deliver erroneous outcomes due to bias in data, algorithms or the teams responsible for managing them."

Gartner Inc. [6]

Protected Attributes (NL)

Gender, Age, Race, Pregnancy, Religion, Political Opinion, Nationality, Citizenship, Sexual Orientation, Civil (marital) Status, Disability Status.

Fairness is defined in relation to the **protected attributes**, which are set by law based on the fundamental human rights and democratic values enshrined in the EU Charter[7] and Dutch Equal Treatment Act (AWGB) [8], e.g., gender or race.

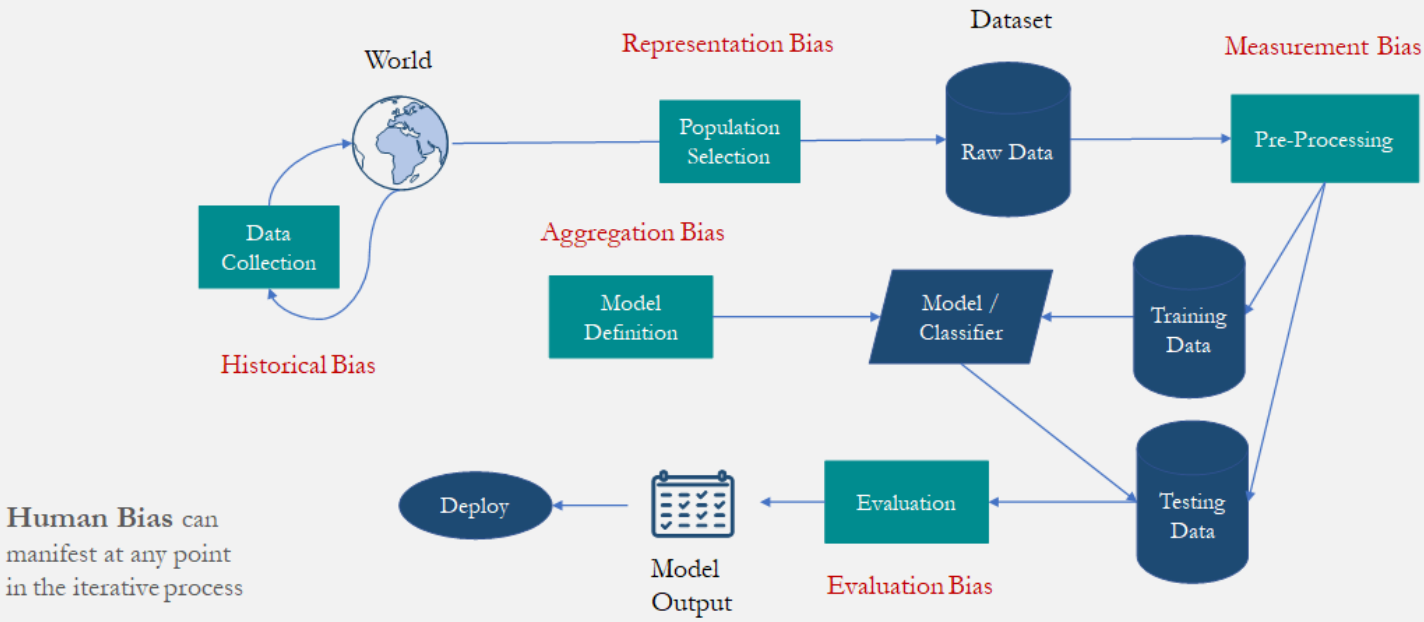
Additionally, these protected attributes should be extended – depending on the use case – with the company's own ethical principles that suit their brand values, e.g., not discriminating on the basis of level of education or whether the customer has children or not.

The process of defining and measuring fairness as well as mitigating bias should start with questioning what unintentional bias might exist in a particular use case and how it might manifest in the data. Identifying bias sources requires careful application-specific analysis.

2. How does bias creep into current models?

Before introducing ways of measurement and mitigation of bias, it is important to see where bias comes from and how it infiltrates in the AI model development pipeline. It turns out that the bias can arise at five different points in model development[9,10].

The major source of bias is the training data algorithm relies on. When AI applications are provided with data that is embedded with **human bias**, the model will replicate those possibly unfair judgements and inadvertently amplify these human biases.



Bias from Data

The first step in the development of an AI algorithm is the data collection. Existing societal prejudices, whereby certain social groups are disadvantaged, may be present in the data – the **historical bias**. The data reflects the reality, but whether or not these values or objectives should be encoded and propagated in the AI model is an issue worth considering. If the present reality puts certain individuals at a systematic disadvantage, then, without intervention, the AI system is likely to reproduce that disadvantage rather than reflecting a fairer future. Identifiable and discriminatory historical

Guidance for Representation Bias

- Ensure the population selected for algorithm's training has similar distributions and proportions for all subgroups (minority and majority groups), and for each protected attribute.
  → stratified sampling
  → oversampling/undersampling of the minority/majority population

bias should be removed in the data collection phase wherever possible.

Representation bias arises when defining and selecting a population – for example, when there is a lack of geographical or social diversity in the dataset. An underrepresentation of certain groups can happen in the data collection stage, where the sampling methods reach only a privileged part of the population. For example, data from two locations may be collected differently – e.g., a bank historically gave credit to families living in relatively wealthy areas. If a protected attribute (e.g., race) varies with location, this will induce bias. One of the key requirements for training data sets considered by the European Commission strictly addresses the desired representativeness – not historical.

"[…] Use data sets that are sufficiently representative, especially to ensure that all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination are appropriately reflected in those data sets."

European Commission[11]

Bias from Algorithm Design

In the next step of the AI model development, the data is pre-processed. This includes data cleaning and labelling, missing value imputation and feature engineering. **Measurement bias** may arise when choosing and measuring the particular features of individuals. Features considered to be relevant to the outcome are chosen, but these can be incomplete or less reliably collected for minority groups. So both the data labelling and imputation of missing values can induce biases.

Sometimes records are removed if they contain missing values, but these may be more prevalent in disadvantaged groups. For example, if missing values appear in attributes such as "native country" – which is correlated to the protected attribute "race" – then discarding or modifying the rows with missing values can significantly bias the sample.

If missing values are independent of protected attributes and occur entirely at random (MCAR), e.g., accidentally omitting a question in a questionnaire, the entire row can be deleted. If the data is missing not at random (MNAR), e.g., a certain question on a questionnaire tends to be skipped deliberately by participants of a certain gender, then removing these cases would cause bias.

This happened, for example, in the case of a pre-employment questionnaire where women did not want to disclose their number of children. If these records were deleted, this would lead to a very biased dataset in terms of gender representation.

Tools for Aggregation Bias

- **Multitask learning**[12] parameterizes different groups differently in the model definition and facilitates learning multiple simpler functions.
- **Fair representation learning**[13] transforms the data so that examples that are similar with respect to the prediction task are close to each other in the feature space.

Bias from Algorithm Deployment

Deployment of an AI algorithm can also lead to a number of unwanted biases. These are mainly aggregation bias and evaluation bias.

Aggregation bias may arise if a one-size-fits-all model is used for groups with different distributions. Aggregation bias can lead to a model that is fit to the dominant population and is less able to fit other groups.

Typically the data used for AI algorithm is split into training, validation and testing datasets. It can happen that the model is optimized on a fair training dataset, but the testing dataset might not represent the target population and the final model seems to perform well only on the majority groups. Such **Evaluation bias** can be exacerbated by the particular metrics that are used to report performance. For example, a single measure to optimize for the overall accuracy of the entire population hides minority group underperformance, but such metrics are used because they make it easy to compare models.

Guidance for Evaluation Bias

Use **Subgroup evaluation** that compares per-group metrics as well as aggregate measures that weight groups equally

3. How to define and quantify fairness?

A natural question is how to define fairness. Specifically, how fairness can be quantified so that it can be considered in an algorithm? There is a wide variety of fairness definitions that have been proposed; however only a few have been successfully adopted in practice[14]. The majority of definitions focus either on **individual fairness** or on **group fairness**[15].

Individual fairness is more fine-grained than any group-notion fairness since it imposes restriction on the treatment for each pair of individuals. However, it is hard to determine an appropriate metric to measure the similarity of two individuals. We observe that financial regulators guide us in favour of group fairness: AI applications "should not inadvertently disadvantage certain groups of customers" (DnB[1]).

Individual Fairness

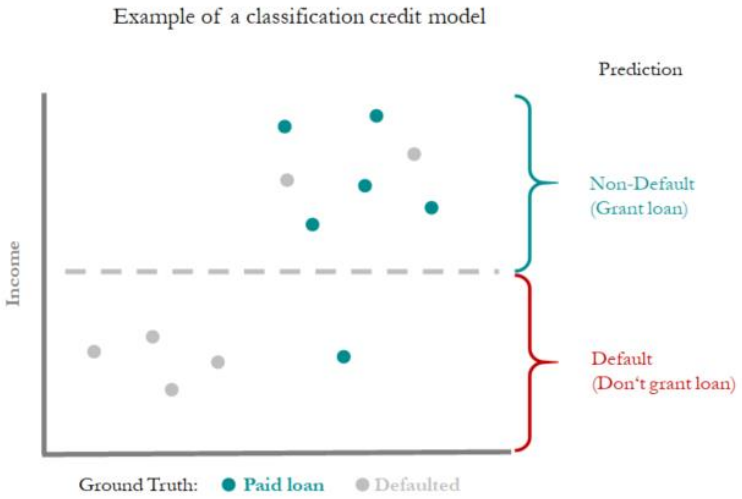
A requirement having the goal of **treating similar individuals in a similar way**. Individuals that are similar with regard to the task should be given similar decisions. Ensures that statistical measures of outcomes are equal for similar individuals.

Group Fairness

Partitions a population into groups defined by protected attributes. **Privileged groups should be treated similarly to the unprivileged group**. Ensures that statistical measures of outcomes are equal across groups.

"We're All Equal" (WAE) - all groups/individuals have similar abilities corresponding to the task

Consider an algorithm that predicts default for loan issuance decisions. We consider a data set with a protected attribute gender A and aim to facilitate non-discrimination regarding this protected attribute. Using an AI model, we compute a score that will be used to predict binary outcomes default/non-default $\hat{Y} \in \{0,1\}$. We refer to $\hat{Y} = 1$ as the favourable class of non-default (paid loan), since it represents the more desirable of the two possible results.



Further, we denote $A = a$ as the unprivileged group (e.g., women) and $A = b$ the privileged group (e.g., men). The actual outcomes – whether the applicant actually defaulted on the loan or repaid it – are $Y \in \{0,1\}$.

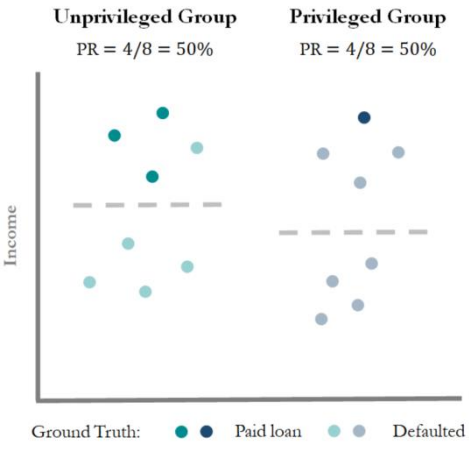
For illustration, consider a credit model based purely on "income". Therefore an income threshold, to decide who gets a loan in the future, is set. Those that are above the threshold will receive the loan (positive outcome). Those below the threshold are the ones who will not (negative outcome).

Fairness metrics, introduced below, are typically in the range between 0 and 100%. Although ideally an algorithm is fair when the metric is zero, in practice the 20-80% rule is often used, i.e., the difference in treatment between privileged and unprivileged group can be at most 20%.

Demographic Parity

Demographic Parity states that the proportion of each segment of a protected class should receive the positive outcome at nearly equal rates. The percentage of people getting a loan in the privileged group should be equal to the percentage of people getting a loan in the unprivileged group, within a margin of 20%. Formally,

$$P(\hat{Y} = 1 \mid A = b) - P(\hat{Y} = 1 \mid A = a) < 0.2$$



Demographic Parity is **suitable when**

- We want to change the state of the current world to improve it by supporting unprivileged groups (e.g. universities are aiming to improve diversity by admitting a fixed number of students from disadvantages backgrounds)

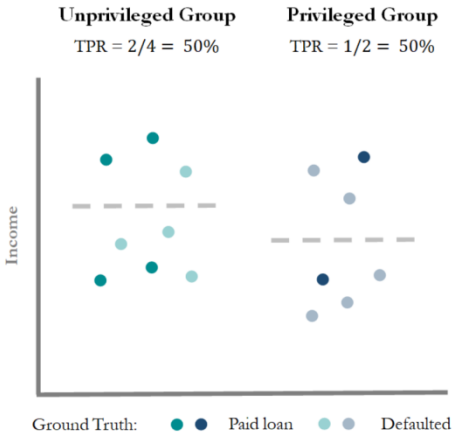
In our case, however, granting loans in equal proportions of men and women without taking into account their characteristics and risk profile is not meaningful. So the following two fairness metrics seem more suitable.

Equal Opportunity / Equalized Odds

Equal Opportunity states that each group should get the positive outcome at nearly equal rates, assuming that people in this group qualify for it. The same percentage of men and women who are likely to succeed at loans are given loans. This meets the lender's objective of identifying loan-worthy applicants, but avoids favouring one gender over another in terms of risk. Formally,

$$P(\hat{Y} = 1 \mid Y = 1, A = b) - P(\hat{Y} = 1 \mid Y = 1, A = a) < 0.2$$

In the figure the percentage of positives that were accurately predicted (True Positive Rate) is 50% for both groups.



Equal Opportunity is **suitable when**

- We want to predict the positive outcome correctly (e.g. we need to be very good at detecting a fraudulent transaction)
- Introducing false positives are not costly (e.g. wrongly notifying a customer about fraudulent activity will not be necessarily expensive to the customer nor the bank sending the alert)
- The target variable is not subjective (e.g.: labelling who is a 'good' employee is very subjective)

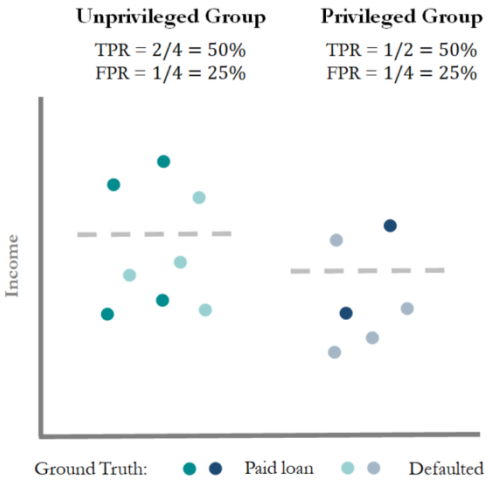
Equalized Odds extends Equal Opportunity stating that model should correctly identify the positive outcome at equal rates across groups (same as in Equal Opportunity), but also miss-classify the positive outcome at equal rates across groups. This additionally takes care of minimising costly False Positives:

$$P(\hat{Y} = 1 \mid Y = 1, A = b) - P(\hat{Y} = 1 \mid Y = 1, A = a) < 0.2$$

$$P(\hat{Y} = 1 \mid Y = 0, A = b) - P(\hat{Y} = 1 \mid Y = 0, A = a) < 0.2$$

Equalized Odds is **suitable when**

- We are interested in predicting the positive outcome correctly (e.g.: correctly identifying who should get a loan drives profits)
- Minimising costly false positives is central (e.g.: reducing giving loans to people who would not be able to pay back)
- The target variable is not subjective (e.g. labelling a loan as paid or defaulted is non-subjective)



These considerations make credit decision models a good use case for Equalized Odds, if the target variable is reliably labelled. However, both Equal Opportunity and Equalized Odds can have flaws. The system might be wrong about who it approved. If among the people the system has rejected, it is wrong about women twice as often as it is wrong about men, then more women than men who deserved loans are being denied.

Predictive Parity

Predictive Parity expands the considerations of Equalized Odds by analysing the ground truth:

$$P(Y = 1 \mid \hat{Y} = 1,\ A = b) - P(Y = 1 \mid \hat{Y} = 1,\ A = a) < 0.2$$

$$P(Y = 0 \mid \hat{Y} = 1,\ A = b) - P(Y = 0 \mid \hat{Y} = 1,\ A = a) < 0.2$$

The probability of actually being in each of the groups (default or non-default) is nearly equal for men and women given that they were predicted to default or repay. The idea is that the decision returned from a prediction algorithm (used to determine the candidate's likelihood to default) for a candidate should reflect its real likelihood of default. This solves certain flaws from Equalized Odds.

The suitable notion of fairness must be chosen in the context of the specific use case and data at hand. It requires understanding of the AI application's goals and the selected protected attributes as well as the definition of the privileged group. Some fairness definitions cannot be satisfied at the same time, e.g. demographic parity and equalized odds. Furthermore, while in financial services group fairness can be adopted, it would not be appropriate in medical applications where gender and race can play an important role in understanding a patient's symptoms.

4. How to mitigate bias?

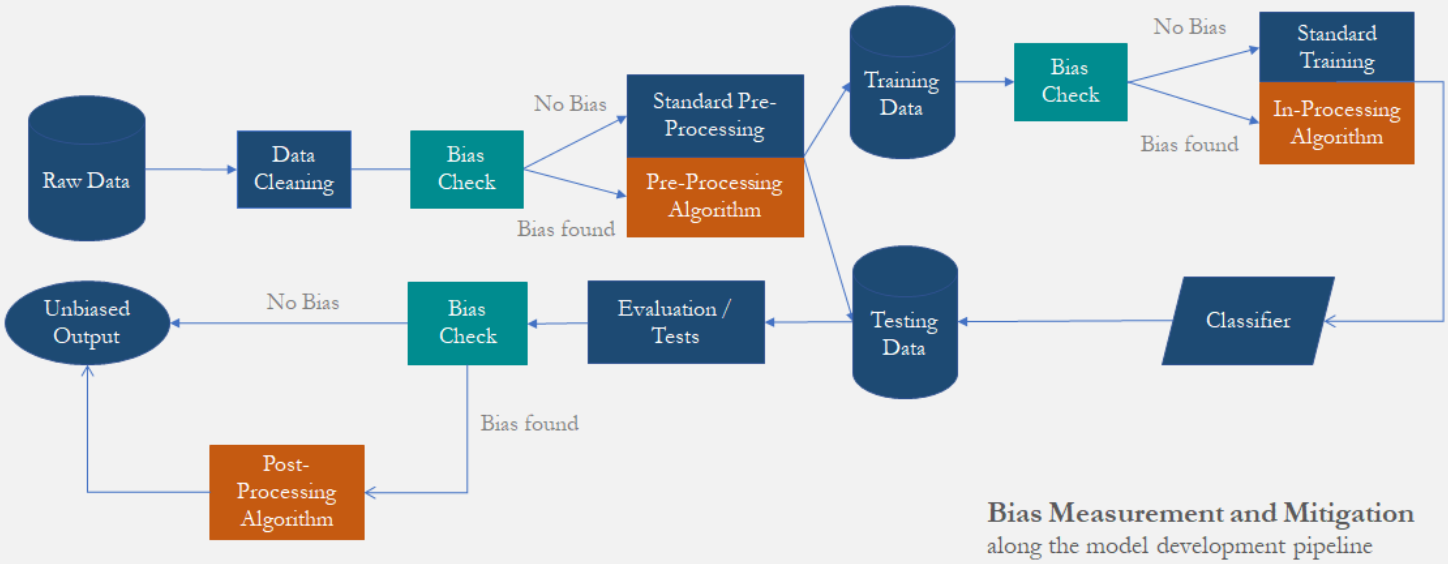
To address potential bias, we need to find out **which unintentional biases might exist** in the model and how they might manifest themselves in the data. A naive straightforward approach to eliminate bias would be to explicitly remove all protected attributes from the training dataset. However, this idea of **Fairness Through Unawareness**[14] rarely suffices due to the existence of redundant encodings – features correlated with the protected attributes. The model uses them as substitutes and causes unwanted discrimination. For example:

- Income level is often correlated with race, gender and age.
- Living area might be a proxy feature for race, since racial and ethnicity demographics often have spatial correlations.

If alternative data is used in model development, then many more extracted features can be correlated with protected attributes. For example, social media activity and features extracted from text and images can be related to gender, age and race. A check for correlations between all the features and the protected attributes can identify proxies and diagnose biases. The measurement of bias according to one of the fairness definitions and the chosen protected attributes should be done preferably at three points in the AI model development pipeline:

- after the raw data is cleaned
- after the training data is formed
- after the final model is evaluated on the testing dataset

Depending on the measurements, the bias can be mitigated at the same three points. We advise the mitigation of bias to be done at only one of the three points, since after such intervention fairness will likely be assured.



Bias Measurement and Mitigation
along the model development pipeline

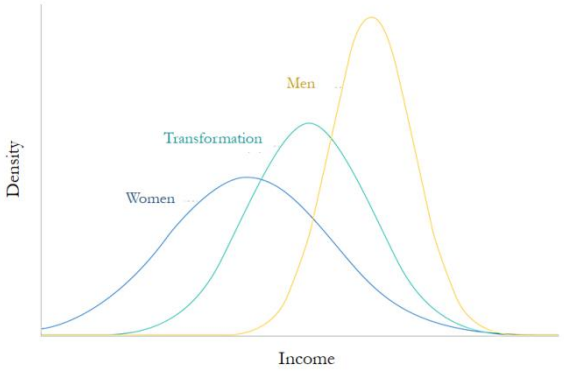
Methods that target bias fall under three categories, depending on where they intervene in the AI development pipeline: **Pre-Processing**, **In-Processing** and **Post-Processing**[16]. We describe a few commonly adopted algorithms below.

Pre-Processing: De-biasing the Data

Pre-processing algorithms are applied before the creation of the model and these transform the training data in order to reduce bias. After de-biasing the training data, the AI algorithm can be trained in an ordinary manner. Modifications of the dataset can be done to the labels, the observed data and the weighting of the data/label pairs.

The idea of **Reweighing**[17] is to apply appropriate weights to different tuples in the training dataset according to the protected attributes. Observations of the unprivileged class with a favourable label get higher weights and observations of the privileged class with a favourable class get lower weights. For example, non-defaulting women are weighted higher while non-defaulting men - lower.

Massaging[18] changes the class labels of some observations of the unprivileged class to the favourable class and vice versa, e.g. some defaulted females will be set as non-defaulted and some non-defaulted men to defaulted ones.



Lastly, **Disparate Impact Removal**[19] transforms features depending on the protected attributes. It aligns the cumulative distributions of features that are part of the privileged group with the ones that are part of the unprivileged group to a median cumulative distribution For example, the distributions of income of men and women are replaced by their median cumulative distribution (figure on the right).

In-Processing: De-Biasing the Algorithms

In-Processing Algorithms are modifications of the traditional learning algorithm itself for addressing unwanted bias during the model training phase. One possibility is to modify the cost function, to include an extra discrimination-aware regularization term (**Prejudice Remover**[20]). These take into account differences in how learning algorithms classify protected versus non-protected classes and penalize the total loss based on the extent of the difference. Besides, it is possible to minimize the standard loss function and add fairness constraints to the optimization problem (**Meta Fair Classifier**[21]).

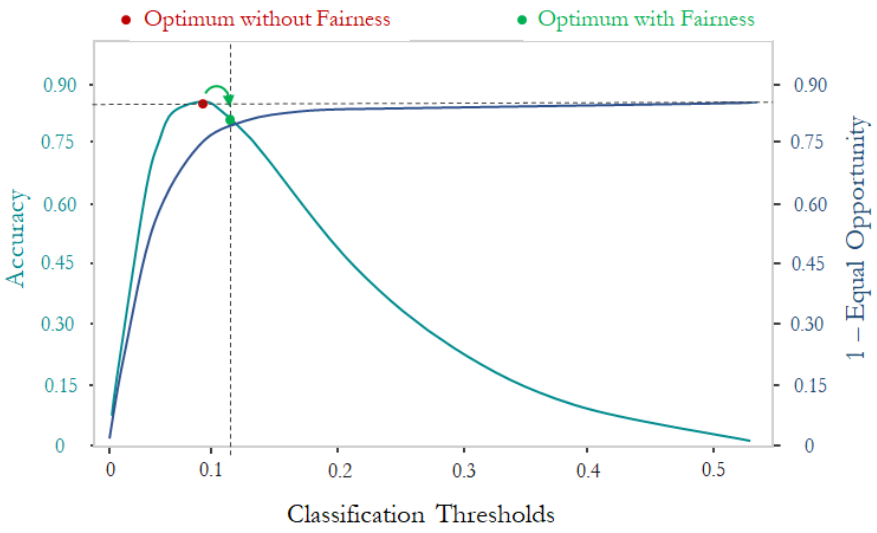
Adversarial Debiasing[22] uses generative adversarial networks. It learns a classifier to maximize prediction performance and simultaneously reduces an adversary's ability to determine the protected

attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any discrimination information that the adversary can exploit.

Post-Processing: De-Biasing the Outcomes

Post-Processing algorithms can be used to reduce bias by manipulating the output predictions in a way that minimizes a fairness metric, after training the classifier. Given the outputs, the idea is to find a proper threshold using the original score function for each group. For example, one can reweigh predictions to make the prediction distribution for privileged and unprivileged group equal and hence minimize the Equal Opportunity[23]. Besides, it is possible to intervene directly in the validation and to choose the suitable classification threshold that assures fairness. These approaches explicitly consider the **trade-off between algorithm performance and fairness measure**.

The figure below describes a situation, where the threshold is chosen so that Equal Opportunity is lower than 0.2. This results in a slightly lower accuracy than taking the pure performance driven outcome.



Trade-off between Model Performance and Fairness

Incorporating fairness in AI models usually comes at the cost of model performance. In general, fairness hurts model performance because it diverts the objective from performance only to both performance and fairness. This highly depends on the adopted definition of fairness, the AI and debiasing algorithms used as well as the data at hand.

We observe that simpler models such as logistic regression or decision trees are more heavily penalized by imposing fairness than more complex models such as neural networks or random forest. Through their complexity, the latter can handle fairness constraints more efficiently and fit data better. Furthermore, a complex fairness metric leads to a higher sacrifice of model performance. For example, choosing Equalized Odds – the most restrictive of the presented definitions – leads to a higher drop in performance than e.g., Demographic Parity. In general, using more representative data may increase algorithm

performance further, while also improving fairness. Additionally, the trade-off becomes narrower if bias is treated at its source – in data collection and training set creation.

When to intervene in the AI Development Pipeline

The suitable choice of intervention depends on the access to the AI model development pipeline and the use case. There is not a one-size-fits-all solution and the suitable method has to be chosen via experimenting and balancing model performance and fairness.

1. If the **model is available only as a "black box" from a third party**, we recommend de-biasing the training data by a suitable pre-processing algorithm. In this way, there is no need to modify the model. The de-biased data can be used further for any downstream task or other models. However, there can be legal issues involved: in some cases, training the decision algorithm on non-raw data can violate antidiscrimination laws.

2. If the **model is built in-house,** it is advisable to de-bias the AI model through an in-processing algorithm, since this offers the highest flexibility to choose the trade-off between performance and fairness. Typically, using in-processing algorithms at training has the lowest decline in performance. However, modifying the complex AI algorithm might be difficult or even impossible. Furthermore, the bias mitigation algorithms in this category depend on the AI model and hence, are task specific.

3. Lastly, if **access to the data and AI model is not given**, then only post-processing techniques are possible. They are easy to apply to any existing classifier without retraining it, however they require test-time access to the protected attributes. This may not be possible when people do not disclose their identities.

5. Fairness and Model Risk Management

Financial institutions are responsible for identifying possible negative ethical impacts of their AI systems. They need to establish a strategy and put in place well-defined processes to test and monitor for potential biases across AI model development, implementation, validation and use. Therefore, existing Model Risk Management (MRM) practices need to be modified for AI models and **concepts of bias and fairness have to be included**. Fairness has to be addressed especially for applications where a model's decisions are likely to impact individuals – and is highly dependent on the use case. Validation should be designed and performed by a group of people as diverse as possible, since multi-disciplinary perspectives help at mitigating inherent societal biases.

As bias is related mainly to the **underlying data** set used, particular focus has to be put on the input data – e.g., its quality, outliers and representativeness. Validators need to check whether model developers have taken steps to ensure fairness and whether an adequate working **definition of fairness** has been used. A **qualitative investigation** involves visualizations of predictions for privileged/unprivileged groups with regard to chosen protected attributes. A **quantitative assessment** needs to be done with respect to the fairness definition. In order to achieve fairness, it has to be tested and corrected at each stage of the model-development process. Additionally, performing data point inspections and exceptions testing (**adversarial examples**) with regard to minority groups that have historically been disadvantaged is advisable.

A strict assessment of the used **features** is crucial, especially if protected attributes are used in the model. Highly regulated models such as credit-decision models might require that every individual feature in the model is assessed and reasoned, while for low-risk models, financial institutions might choose to review the feature-engineering process only at a high level. A subsequent step might include evaluation whether the selected protected features are predictive for the particular decision making system as well as how impactful they are.

Addressing bias is a complex process closely associated with the topic of **model explainability**, as interpreting a model's decision is beneficial in detecting bias. AI algorithms are typically far more complex than their traditional statistical counterparts. The level at which explainability is needed, has to be assessed with reference to the type of decision aided by the algorithm, the potential impact on the customers and the level of organization's risk appetite. For example, an automated decision to refuse a mortgage application must be explained in simple and straightforward terms, but for a trading algorithm, this might be unnecessary (and quite difficult).

Certain types of AI models modify their parameters dynamically with new incoming data, requiring financial institutions to decide whether a **dynamic calibration** is appropriate and if yes, whether fairness is still assured. Often, vendor and third-party models are available as **black-box**. Typical approaches in model validation of such models consist of outcomes analysis, sensitivity analysis and benchmarking – especially in terms of fairness goals.

Finally, the agreed degree of removal of identified biases and the rationale behind trade-off decisions should be thoroughly described in the model documentation.

References

1. Joost van der Burgt (2019). General principles for the use of Artificial Intelligence in the financial sector. Amsterdam: De Nederlandsche Bank.
2. Vincent, J. (2019). Apple's credit card is being investigated for discriminating against women. The Verge. Available at https://bit.ly/33LxABD
3. OECD (2019). The OECD AI Principles. Available at https://www.oecd.org/going-digital/ai/principles
4. High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. European Commission.
5. FSB (2017). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. Basel: Financial Stability Board.
6. Gartner (2018). Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence. Available at https://gtnr.it/3op3yLS
7. European Parliament (2012). Charter of Fundamental Rights of the European Union 2012/C 326/02.
8. College voor de Rechten van de Mens (2005). Equal Treatment Act (AWGB). Utrecht.
9. Mehrabi, Ninareh et. al (2019). A Survey on Bias and Fairness in Machine Learning.
10. Suresh, Harini, and John V. Guttag. "A Framework for Understanding Unintended Consequences of Machine Learning." arXiv preprint arXiv:1901.10002 (2019).
11. European Comission (2020). White Paper: On Artificial Intelligence - A European approach to excellence and trust. Available at https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
12. Zhang, Y., & Yang, Q. (2017). A Survey on Multi-Task Learning. ArXiv, abs/1707.08114.
13. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C.. (2013). Learning Fair Representations. PMLR 28(3):325-333
14. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. ArXiv, abs/1104.3913.
15. Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. New York, NY, USA, 1–7
16. R. K. E. Bellamy et al., "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," in IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4:1-4:15, 1 July-Sept. 2019, doi: 10.1147/JRD.2019.2942287.
17. Faisal Kamiran; Toon Calders, Data preprocessing techniques for classification without discrimination. Retrieved 17 December 2019
18. F. Kamiran and T. Calders, "Classifying without discriminating," 2009 2nd International Conference on Computer, Control and Communication, Karachi, 2009, pp. 1-6,
19. Feldman, M., Friedler, S. et.al (2015). Certifying and removing disparate impact. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268, Sydney, Australia,
20. Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. Machine Learning and Knowledge Discovery in Databases, pp. 35–50, 2012.
21. L. Elisa Celis. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 319–328.
22. Brian Hu Zhang. 2018. Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, NY, USA, 335–340.
23. Michael P. Kim. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, USA.

How can Probability & Partners help?

Probability & Partners has a long-standing operational expertise in building, deploying and validating AI models. We provide guidance on ethical considerations in AI model development and advise you on the optimal balance between model performance and fairness. Further, we assist you at choosing the right fairness metric as well as debiasing techniques and help you to embrace responsible, trustworthy AI.

We can also help you with full stack validation of AI solutions irrespective of the area of application. We create a tailormade validation that best suits the materiality, importance for business and stage of deployment of your AI/ML model. We have experience with "black-box" validations that are performed without obtaining the full model documentation or source-code.

Dr. Svetlana Borovkova

Head Quantitative Modelling

svetlana.borovkova@probability.nl

Pim Poppe

Managing Partner

pim.poppe@probability.nl

PROBABILITY & PARTNERS