



rev.research  
& advanced ML



# Understanding and Mitigating *Bias*


# Understanding and Mitigating *Bias*



rev.research  
& advanced ML

# 1. Foreword

Machine learning (ML) algorithms identify pattern in data. Its major strength is the desired capability to find and discriminate classes in training data, and to use those insights to make predictions for new, unseen data. In the era of “big data”, a lot of data is available with all sorts of variables. The general assumption is that the more data is used, the more precise becomes the algorithm and its predictions. When using a large amount of data, it clearly contains many correlations. However, not all correlations imply causality, because no matter how large the dataset is, it still only remains a snapshot of reality.



**In a training data** set on claims of a car insurance, red cars may have caused more accidents than cars of other colour. The ML algorithm detects this correlation. However, there is no scientific proof of causality between the colour of a car and the risk of accidents.

---


Beyond the incomprehension in terms of pricing for a customer, for the sake of the algorithm’s performance it is crucial to notice and eliminate this kind of unwanted correlations. Otherwise, the algorithm is biased and results on new data in production may be poor. In the previous example, a competitor with a better algorithm, which does not falsely attribute a higher risk to drivers of red cars, can offer a lower price to those customers and entice them away.

Besides the performance aspect, there is a second problem which appears when the predictions impact people, and when the algorithm is biased to favour privileged groups over unprivileged groups, this resulting in discrimination.

It is important to note that these unwanted discriminations may happen without explicitly providing sensitive personal data. In fact, other attributes can implicitly reveal this information serving as proxy. For example, a car model can hint at the owner's sex, or the zip code may correlate with a resident's race or religion. As of today, it is not clear how much a ML algorithm exploiting correlations can be seen as reconstructing a protected attribute and using it in a causal way.

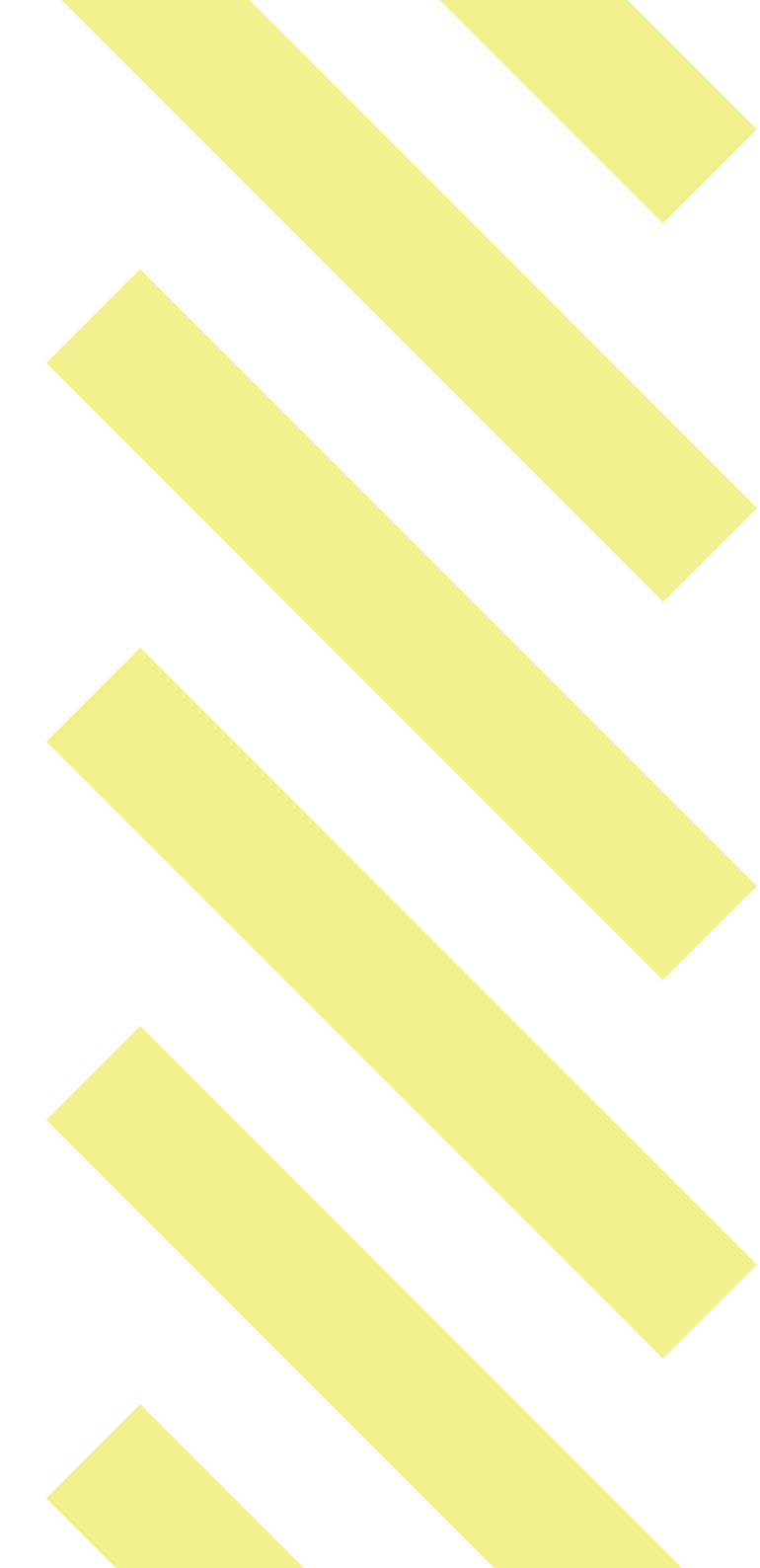
Although not everything is clear today, it is important to understand and to mitigate unwanted bias as much as possible, since it may not only result in low performance, but also cause unintended discrimination.

In this report, we want to encourage all stakeholders to understand the most fundamental sources of unwanted bias and the consequences it causes. More precisely, we seek to explain to CDOs, DPOs, data scientists, actuaries, and any other interested parties how bias in data and in data models can be identified and mitigated.

A handwritten signature in black ink, appearing to read 'Marcin Detyniecki', with a stylized initial 'M' and 'D'.

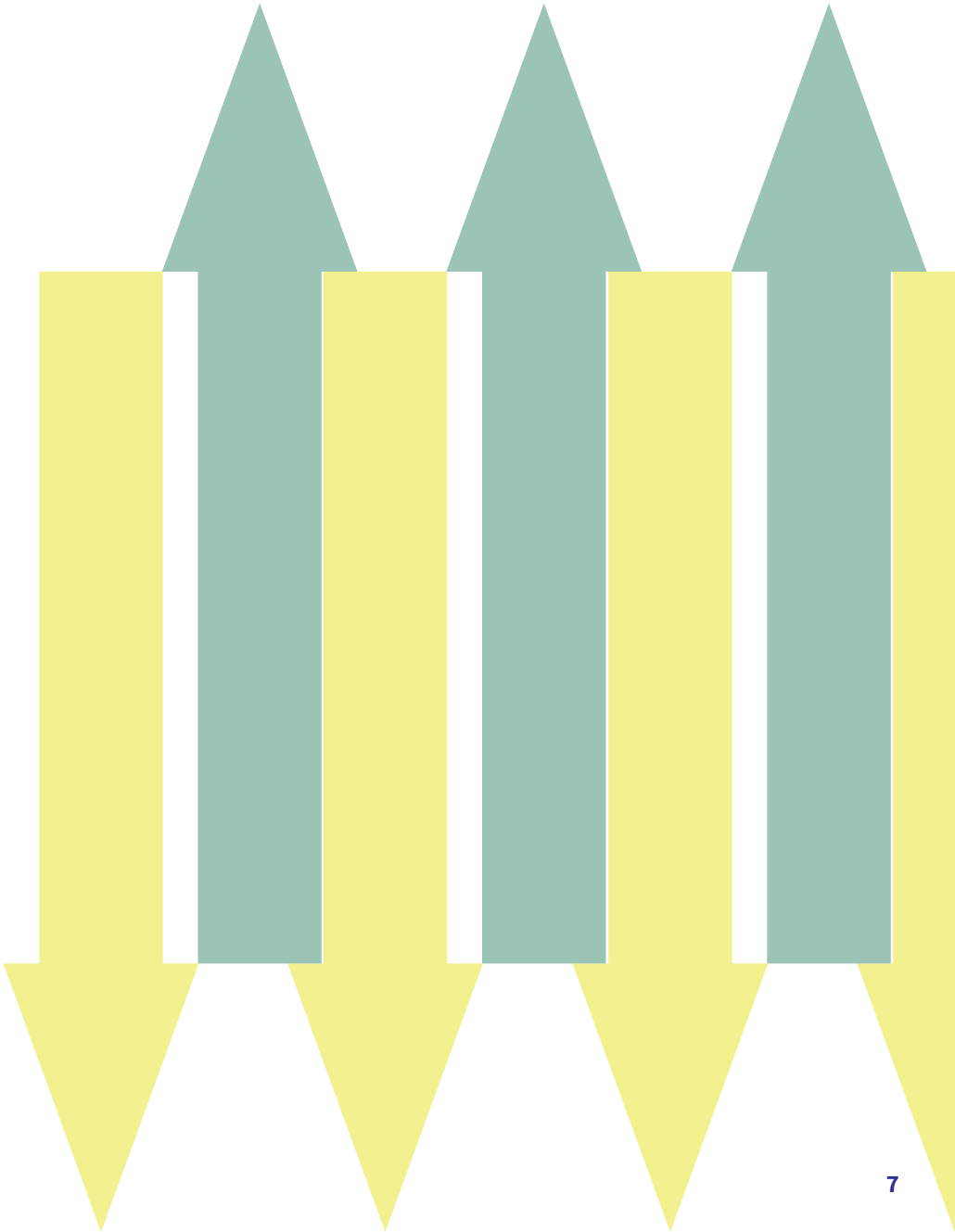
**Marcin DETYNIECKI**

Chief Data Scientist




# Table of Contents

<b>1.</b>	Foreword	3
<b>2.</b>	Introduction	8
<b>3.</b>	Distinctive features of machine learning	10
<b>4.</b>	Different sources of bias	13
<b>4.1</b>	Human bias	14
<b>4.2</b>	Selection bias	15
<b>5.</b>	What is fair?	17
<b>5.1</b>	Information sanitization	18
<b>5.2</b>	Statistical/Group fairness	18
<b>5.3</b>	Individual fairness	19
<b>6.</b>	Bias mitigation	20
<b>6.1</b>	Pre-processing	20
<b>6.2</b>	In-processing	21
<b>6.3</b>	Post-processing	22
<b>7.</b>	Legal context	23
<b>7.1</b>	State-of-the-art	23
<b>7.2</b>	Discussion	25
<b>8.</b>	Recommendations	26
<b>9.</b>	Acknowledgements	27
<b>10.</b>	Bibliography	28



## 2. Introduction



Machine learning models are increasingly used in decision making processes. In many fields of application, they generally deliver superior performance compared with conventional, deterministic algorithms. However, those models are mostly black boxes which are hard, if not impossible, to interpret. Since many applications of machine learning models have far-reaching consequences on people (credit approval, recidivism score etc.), there is growing concern about their potential to reproduce discrimination against a particular group of people based on protected characteristics such as gender, race, religion, or other. In particular, algorithms trained on biased data are prone to learn, perpetuate or even reinforce these biases [1]. In recent years, many incidents of this nature have been documented. For example, an algorithmic model used to generate predictions of criminal recidivism in the United States (COMPAS) discriminated against black defendants [2]. Also, discrimination based on gender and race could be demonstrated for targeted and automated online advertising on employment opportunities [3].

In this context, the EU introduced the General Data Protection Regulation (GDPR) in May 2018. This legislation represents one of the most important changes in the regulation of data privacy in more than 20 years. It strictly regulates the collection and use of protected personal data. With the aim of obtaining non-discriminatory algorithms, it rules in Article 9(1): “Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concern-



ing health or data concerning a natural person’s sex life or sexual orientation shall be prohibited.” [4]

Currently, one fairness method often used in practice is to remove protected attributes from the data set. This concept is known as “fairness through unawareness” [5]. While this approach may prove viable when using conventional, deterministic algorithms with a manageable quantity of data, it is insufficient for machine learning algorithms trained on “big data”. Here, complex correlations in the data may provide unexpected links to protected information. This way, presumably non-protected attributes can serve as substitutes or proxies for protected attributes.

For this reason, next to optimizing the performance of a machine learning model, the new challenge for data scientists is to determine whether the model output predictions are discriminatory, and how they can mitigate such unwanted bias as much as possible.

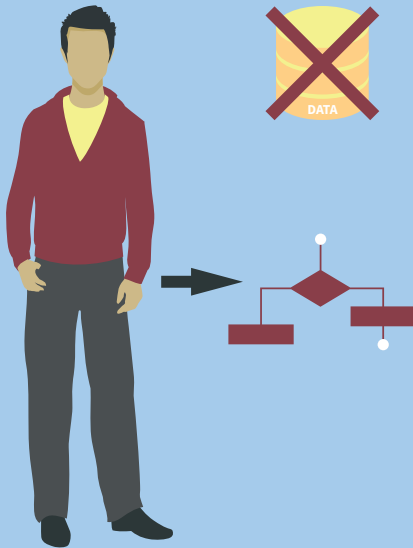


```
# set up the learners
learners = []
me_set = [0, 1, 5, 10]
for me in me_set:
    learners.append(learner_factory(me))
# load data, split into training and test set
data = dt.ExampleData(1000)
selection = dt.randomSubsetFromTraining(data)
train_data = dt.selectFromTraining(data, selection)
test_data = dt.selectFromTesting(data, selection)
# obtain results for each learner
result = dt.evaluateLearners(data(learners,
    train_data, test_data, numClassifiers = 1))
CA = dt.Stat.CA(result)
IS = dt.Stat.IS(result)
print "Example 9.1: Comparison of CA and IS"
for i in range(1, len(result)):
    print "Learner %d: CA = %f, IS = %f" %
```

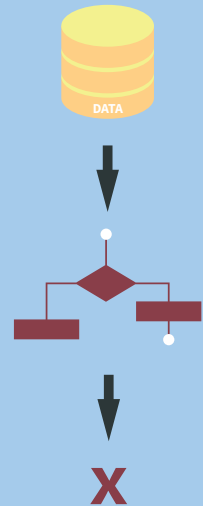
### 3. Distinctive features of machine learning

In order to understand the fundamental characteristic differences of machine learning (ML), which we believe correspond to a paradigm shift, we compare in this section the functioning of conventional algorithms with the new type of algorithms, ML. The two major interrelated differences are, first, a new type of relationship to data and, second, the nature of the algorithm used in the production phase (vs. development one). Through this prism, we propose to compare classical algorithms, which we call here deterministic algorithms (DA), with machine learning ones.

## Development



## Production

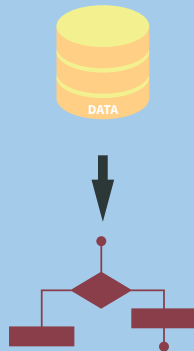
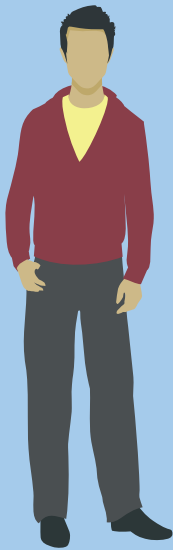


**Figure 1:** Deterministic algorithm (DA) in development and production phase

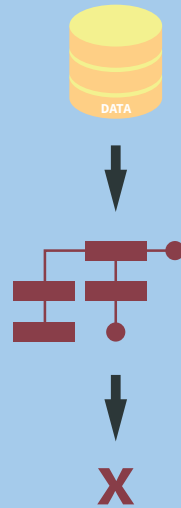
## Deterministic algorithm (DA)

Conventional algorithms usually are deterministic algorithms. Like a recipe, they consist of a hard-coded set of rules which always produce the same output. The software engineer explicitly programs the algorithm's logic without using any data. When the algorithm is put into production, data are fed to the algorithm in order to produce results. Data has no impact on the algorithm in itself.

## Development



## Production



**Figure 2:** Machine Learning (ML) algorithm in development and production phase

## Machine learning (ML)

In contrast to deterministic algorithms, when “programming” machine learning we have two different phases. The first one is programming the ML algorithm itself, which is de facto what we just described for the deterministic algorithms. In a second phase, usually called “training”, a data scientist (or data engineer) uses the ML algorithm together with data to produce a new algorithm: the production algorithm. Often, the ML algorithm and the production algorithm get confused. Data scientist call the latter a “trained algorithm” which contains thousands of parameters that were not explicitly programmed by a human, but rather automatically “learned”, i.e. estimated, using data samples. Here, data is **grown into** an algorithm.

ML algorithms are strongly dependent on the data they use to create the “production algorithm”. Because they are also prone to any hidden bias contained in the data, and due to the potential of getting deployed at scale, even minimal systematic errors in the algorithms can lead to reinforced discrimination.

## 4. Different sources of bias

There are plenty of different forms of bias which can cause unwanted and unexpected results [6] [7]. Automation bias for example is the phenomenon when people trust suggestions of automated systems over human reasoning. Several severe airplane accidents happened in the past because the pilots had trusted the autopilot more than their own judgment [8]. Another type of bias may occur when an algorithm is deployed in an environment for which it was not trained in the first place. For example, if it is applied in a different geographical region or on a different group of people.

While explicitly programmed rules in algorithms or the way they are used in practice may produce biased results, this is a long-known problem which already applies to conventional deterministic algorithms. In the following, we focus on a new source of bias resulting from data, which was introduced by the emergence of machine learning technologies. More specifically, we discuss human bias in data and selection bias.

## 4.1 Human bias

The first source of bias which comes naturally to mind is human bias. Different types of this kind of well-studied bias are outlined in Table 1.

Training data can consist of labels of objective observations, as for instance coming from a measuring device. However, training data may also involve human assessment. Data labels which include human judgment may have been labelled with prejudice. Since the labels serve as **ground truth**, the algorithm's performance directly depends on them, and any bias contained gets reproduced at scale in the model.

Type	Description	Example
In-group bias	Rather trust people of a group which you belong to.	Preferring candidates in a recruitment process you share a biographical similarity with.
Out-group homogeneity bias	Less trust in people of a group which you do not belong to.	Lack of language abilities may cause mistrust.
Implicite bias	False fundamental assumptions based on individual experience but not eligible for generalization.	Online streaming providers which tend to recommend movies about pink princesses to girls and movies about martial action heroes to boys.

**Table 1:** List of different sources of human bias

## 4.2 Selection bias

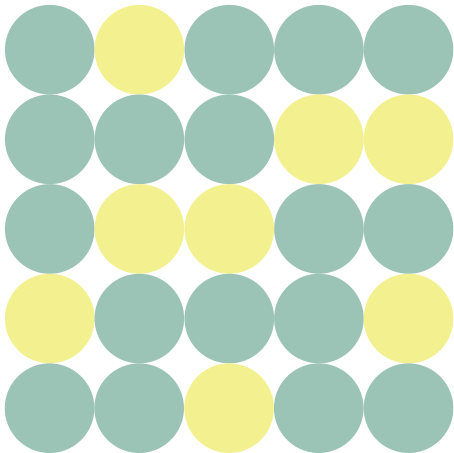
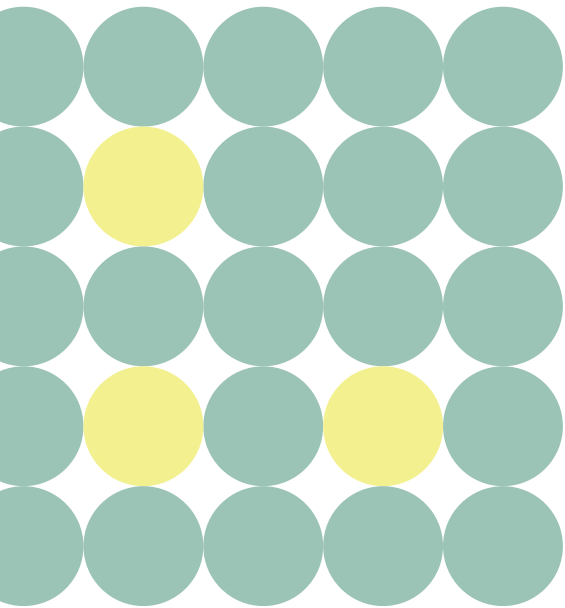
Another, less obvious source of bias is the process of how the data was collected. If data does not reflect the real distribution, a ML algorithm using these data for training will learn and enforce the bias.

In Table 2 we provide a list of different types of bias which may cause selection bias in data.

Type	Description	Example
Coverage bias	Data not selected in a representative manner.	Surveys conducted only on the Internet do not include households without Internet access.
Reporting bias	The frequency of events captured in the data does not correspond to the frequency in reality. This may happen when only extreme occurrences get registered while less outstanding events are omitted.	Movie, hotel or book reviews tend to be subject to reporting bias because only customers with extreme sentiments care to write a review.

Type	Description	Example
Participation bias	Some demographics may be underrepresented in the data.	Busy parents with young children are more likely to refuse a telephone survey than retired people with plenty of time.
Sampling bias	Using a non-randomized sub sample for model training can make it susceptible to bias.	Data from telephone surveys conducted on a working day between 10 and 11 a.m. may over-represent the non-working population.

**Table 2:** List of different sources of selection bias







## 5. What is fair?

Fairness is an ethical concept which refers to plural conceptions of justice between individuals. This concept is at the heart of social science research, and the difficulty to find a general definition is obvious: fairness is based on ethical value judgment, and its application will vary according to cultures, religions, political systems, etc.

In order to be able to measure and improve fairness in technical systems, we would first have to agree on a statistical definition of fairness as baseline. In current research, there exist plenty of different definitions which are mutually incompatible [29]. In the following, we outline the most popular approaches. No definition serves as silver bullet for all use cases, the right choice depends on the context of the use case and on the data available.

## 5.1 Information sanitization

This approach limits the data that are used for training the classifier. In its most straightforward version named “Fairness through unawareness”, an algorithm is considered fair if it does not make use of the protected attribute [5]. The notion is that omission of protected attributes when training the model prevents from unfair use.

## 5.2 Statistical/Group fairness

This type of fairness definition partitions the world into groups defined by one or several high-level protected attributes. It requires that a specific relevant statistic about the classifier is equal across those groups.

### 5.2.1 Demographic parity

An algorithm is considered fair if the prediction is independent of the protected attribute [9]. If the base ground truth outcome of the two demographic groups are totally different, this definition might not be appropriate, since for example a model which selects the best 5% women and randomly 5% of men would be perfectly fair according to this definition.

### 5.2.2 Equalized odds

An algorithm is considered fair if across both demographics for the positive outcome the predictor has equal true positive rates, and for negative outcomes, the predictor has equal false positive rates [10]. This constraint enforces that accuracy is equally high in all demographics. The rate of positive and negative classification is equal across the groups.

### 5.2.3 Equalized opportunities

An algorithm is considered fair if across both demographics, only for the positive outcome the predictor of has equal true positive rates [10]. The notion is that the chances of being correctly classified positive should be equal for every group.

## 5.3 Individual fairness

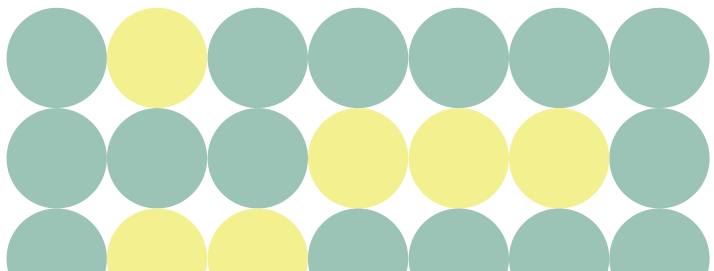
This family of definitions binds at the individual level. It suggests that fairness means that similar individuals should be treated similarly, specifying an adequate similarity metric.

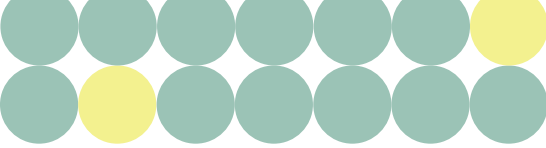
### 5.3.1 Fairness through awareness

An algorithm is considered fair if it gives similar predictions to similar individuals [9].

### 5.3.2 Calibration

The positive predictive value is equalized across the groups for a score [11]. For any demographic, an optimally calibrated classifier tries to match the percentage of individuals with a specific score with the probability score.





## 6. Bias mitigation

Many bias mitigation strategies for machine learning have been proposed in recent years. The different approaches can be divided in the following three distinct groups.

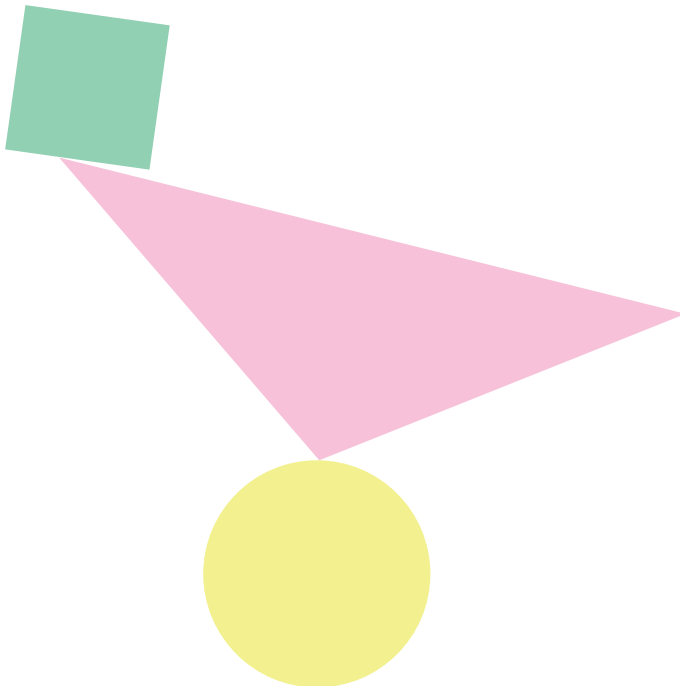
### 6.1 Pre-processing

Efficient bias mitigation starts at the data acquisition and processing phase since the source of the data and also the extraction methods can introduce unwanted bias. Therefore, a maximum of effort must be put into validating the integrity of the data source and in ensuring that the data collection process includes appropriate and reliable methods of measurement. Prior to the era of “big data”, most data were collected by questionnaires. This allowed the development of experimental designs to control possible biases by statistical analysis. Today, technology provides us with large amounts of data at low cost, however, information about the conditions under which the data were collected is often rare.

Hence, algorithms which belong to the pre-processing family ensure that the input data is balanced and fair. This can be achieved by suppressing the protected attributes, by changing class labels of the data set, and by reweighting or resampling the data [13] [14] [15]. In some cases, it is also necessary to reconstruct omitted or censored data in order to ensure the data sample is representative. There exist plenty of imputation methods to achieve this objective, and the hot deck procedures belong to the most efficient ones [14].

## 6.2 In-processing

The second type of mitigation strategies comprises the in-processing algorithms. Here, undesired bias is directly mitigated during the training phase. A straightforward approach to achieve this goal is to integrate a fairness penalty directly in the loss function. One such algorithm integrates a decision boundary covariance constraint for logistic regression or linear SVM [13]. In another approach, a meta algorithm takes a fairness metric as part of the input and returns a new classifier optimized towards that fairness metric [9]. Furthermore, the emergence of generative adversarial networks (GANs) provided the required underpinning for fair classification using adversarial debiasing [17]. In this field, a neural network classifier is trained as classical predictor, while simultaneously the ability of an adversarial neural network to predict a protected attribute is minimized [18] [19] [20].



## 6.3 Post-processing

The final group of mitigation algorithms follows a post-processing approach. In this case, only the output of a trained classifier is modified. A Bayes optimal equalized odds predictor can be used to change output labels with respect to an equalized odds objective [11]. A different paper presents a weighted estimator for demographic disparity which uses soft classification based on proxy model outputs [21]. The advantage of post-processing algorithms is that fair classifiers are derived without the necessity of retraining the original model which may be time consuming or difficult to implement in production environments. However, this approach may have a negative effect on accuracy or could compromise any generalization acquired by the original classifier [22].

## 7. Legal context<sup>1</sup>

In most situations, personal data will be used to train the ML algorithm. These data are subject of a special protection, at European level, mainly by the General Data Protection Regulation (GDPR). The purpose of this regulation, which entered into force on 25th May 2018, is to harmonize at European level the conditions for the processing of personal data and their use, particularly for decision-making. In the following, we provide details on the existing rules of present regulation in the context of fairness and bias. Further, we open the discussion by presenting limitations and gaps we identified with respect to the emergence of machine learning technologies.

### 7.1 State-of-the-art

At European level, several texts regulate the use of information on people in order to fight discrimination. This principle is stated in the Convention for the Protection of Human Rights and Fundamental Freedoms [23] in article 14 entitled “Prohibition of discrimination”. It is also contained in the Charter of Fundamental Rights of the European Union [24] which states in Article 21 that “[a]ny discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.”

These articles materialize the fact that society, via this regulation, expects that whatever is necessary will be done to avoid

---

1 B. Ruf, M. Hirot, M. Detyniecki and N. Shire, “Regulating Machine Learning: where do we stand?”, AXA, 2019.

any type of discrimination. Regulation goes one level further and gives some advice on how this could be achieved by proposing to forbid the use or the consideration of some type of data. For instance, in the field of insurance and financial services, it is forbidden to use sex as a factor in the calculation of premiums and benefits if it results in differences in individuals' premiums and benefits [25].

As a general principle, the GDPR prohibits the use of data which are considered protected and subject to special protection. For instance, data concerning health, a natural person's sex life, or sexual orientation (Art. 9) and data related to criminal convictions and offences (Art. 10) can only be processed under certain conditions (for example requiring consent of the data subject).

A processing for profiling may reveal some inferred protected data from correlations. In this case, the WP29 [23] recommends checking that:

- ✓ the processing is not incompatible with the original purpose;
- ✓ they have identified a lawful basis for the processing of the special category data; and
- ✓ they inform the data subject about the processing.

It is also important to note that in certain sectors, such as insurance law, specific rules exist that allow people's characteristics, such as age or health status, to be considered in order to offer them different products or services and therefore to process protected data.

For instance, in France, the regulation of life insurance allows the insurer to ask the subscriber to complete a medical questionnaire [27] that will determine if the insurer assures without special conditions, with exclusions, with a surcharge or even refuses to insure.



The French supervisory authority (Commission Nationale de l'Informatique et des Libertés) has issued simplified standards dedicated to the insurance sector [14], which determine for each purpose which data can be collected and processed. These standards specifically allow the collection of health data for contract subscription and contract management as these data will be needed to assess risk or harm. Even though these simplified standards are no longer in effect with the entry into force of GDPR, they may still be useful as guidelines.

## 7.2 Discussion

The idea of preventing algorithms from unfair use of protected attributes by forbidding to use them in the training process is also known as “fairness through unawareness” [26]. However, it falls short in the case of “big data” where other attributes or a complex combination of them may serve as proxy of a protected attribute. Seemingly insignificant attributes, or several attributes combined, may provide an unexpected link to protected information.

This risk is not totally solved but mitigated by the principle of data minimization according to which the data controller must collect and process only the personal data necessary for the intended purpose. By limiting the number of variables used, we theoretically limit the risks of finding proxies of a protected attribute. But market evolution and usage of new data, seeking for a more direct grasp of the risk, such as the one coming from connected objects (e.g. cars, home), will reveal the above-mentioned challenge.

Moreover, paradoxically, by forbidding to collect protected attributes, there is no possibility to measure for potential discrimination at a later point, which may even impede the pursuit of fairness.

## 8. Recommendations

In order to avoid negative impact from bias, such as low performance or unintended discrimination, it is absolutely necessary to minimize it. We have identified a few suggestions to tackle this challenge as of today.

- ✓ **Raise awareness and train stakeholders**

*Human bias is a major source of bias in AI. Therefore, we recommend producing educational material and conducting workshops and trainings adapted to different levels for employees. Creating spaces for interdisciplinary exchange drives the comprehension of the topic and leads also to less biased products or results..*

- ✓ **Identify context-specific fairness definition**

*To monitor and control bias, it is key to quantify fairness. Hence, for each application case, it is necessary to select the protected groups, decide on the best definition of fairness and identify a set of suitable metrics.*

- ✓ **Audit new products and monitor models in production**

*We need to detect and mitigate bias continuously. Open source libraries such as “AI Fairness 360” [28] should become an integral part of our development workflow. However, not all testing can be automated. Establishing operational procedures and practices can help to avoid unwanted bias systematically.*

- ✓ **Produce knowledge and share**

*The topic is complex, and the last months have given rise to significant public attention and debate. Therefore we believe that research effort is necessary. At AXA we have an internal team of experts who follow the latest developments and actively contribute together with the academic community. The generated knowledge should also be shared with regulators in order to push for stable and clear legal framework.*

# 9. Acknowledgements

Big thanks to Marie Hirot, Data Protection Specialist at AXA, who provided the legal assessment of how bias is covered in current regulation ([Section 7](#)).



## 10. Bibliography

- [1] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama and A. Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” CoRR, pp. 1--25, 2016.
- [2] J. Angwin, J. Larson, S. Mattu and L. Kirchner, “Machine Bias. There’s software used across the country to predict future criminals. And it’s biased against blacks.,” 23 5 2016. [Online]. Available: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [3] A. Lambrecht and C. Tucker, “Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads,” SSRN Electronic Journal, 2016.
- [4] European Union, “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC,” Official Journal of the European Union, pp. 1--88, 2016.
- [5] D. Pedreshi, S. Ruggieri and F. Turini, “Discrimination-aware Data Mining,” in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, 2008.
- [6] Wikipedia, “List of cognitive biases,” June 2019. [Online]. Available: [https://en.wikipedia.org/wiki/List\\_of\\_cognitive\\_biases](https://en.wikipedia.org/wiki/List_of_cognitive_biases).

- [7] Google, “Fairness: Identifying Bias,” June 2019. [Online]. Available: <https://developers.google.com/machine-learning/crash-course/fairness/identifying-bias>.
- [8] M. Konnikova, “The Hazards of Going on Autopilot,” June 2017. [Online]. Available: <https://www.newyorker.com/science/maria-konnikova/hazards-automation>.
- [9] S. Venkatasubramanian, A. Friedler and C. Scheidegger, “On the (im)possibility of fairness,” CoRR, vol. abs/1609.07236, 2016.
- [10] D. Pedreshi, S. Ruggieri and F. Turini, “Discrimination-aware data mining,” Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08, p. 560, 2008.
- [11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. Zemel, “Fairness Through Awareness,” 2011.
- [12] M. Hardt, E. Price and N. Srebro, “Equality of Opportunity in Supervised Learning,” pp. 1--22, 2016.
- [13] S. Barocas, M. Hardt and A. Narayanan, “Limitations and Opportunities,” 2018.
- [14] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” Knowledge and Information Systems, pp. 1--33, 2012.
- [15] T. Adel and A. Weller, “One-network Adversarial Fairness,” Aaai, 2019.
- [16] F. P. Calmon, D. Wei, K. N. Ramamurthy and K. R. Varshney, “Optimized Data Pre-Processing for Discrimination Prevention,” in Advances in Neural Information Processing Systems 30, 2017.
- [17] R. R. Andridge and R. J. Little, “A Review of Hot Deck Imputation for Survey Non-response,” International Statistical Review, no. 78, pp. 40-64, 2010.

- [18] M. B. Zafar, I. Valera, M. G. Rodriguez and K. P. Gummadi, “Fairness Constraints: Mechanisms for Fair Classification,” in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017.
- [19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative Adversarial Nets,” in Advances in Neural Information Processing Systems 27, 2014.
- [20] B. H. Zhang, B. Lemoine and M. Mitchell, “Mitigating Unwanted Biases with Adversarial Learning,” Association for the Advancement of Artificial Intelligence, 2018.
- [21] C. Wadsworth, F. Vera and C. Piech, “Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction,” CoRR, 2018.
- [22] G. Louppe, M. Kagan and K. Cranmer, “Learning to Pivot with Adversarial Networks,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.
- [23] J. Chen, N. Kallus, X. Mao, C. Tech and G. Svacha, “Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved,” 2019.
- [24] M. Donini, S. Ben-David, M. Pontil and J. Shawe-Taylor, “An efficient method to impose fairness in linear models,” in NIPS Workshop on Prioritising Online Content, 2017.
- [25] Council of Europe, Convention for the Protection of Human Rights and Fundamental Freedoms.
- [26] Council of Europe, Charter of fundamental rights of the European Union (2012/C 326/02).
- [27] Council of Europe, Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, Article 5.
- [28] WP29, Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679.

- [29] “French Insurance Code, Article L.113-2”.
- [30] R. K. E. Bellamy, K. Dey and H. Michael, “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” 2018.
- [31] L. E. Celis, L. Huang, V. Keswani and N. K. Vishnoi, “Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees,” CoRR, 2018.
- [32] M. Hardt, E. Price and N. Srebro, “Equality of Opportunity in Supervised Learning,” in Advances in Neural Information Processing Systems 29, 2016.
- [33] CNIL, Delib. n°2013-212 concerning automated processing of personal data relating to the execution, management and enforcement of contracts implemented by insurance, capitalization, reinsurance, insurance assistance and through their intermediaries.
- [34] J. Chen and J. Shao, “Nearest Neighbor Imputation for Survey Data,” Journal of Official Statistics, vol. 16, no. 2, pp. 113-131, 2000.

# Experts



**Boris RUF** [boris.ruf@axa.com](mailto:boris.ruf@axa.com)

---

Research Data Scientist



**Vincent GRARI** [vincent.grari@axa.com](mailto:vincent.grari@axa.com)

---

Research Data Scientist

# Sponsors



**Marcin DETYNiecki** [marcin.detyniecki@axa.com](mailto:marcin.detyniecki@axa.com)

---

Chief Data Scientist



**Roland SCHARRER** [roland.scharrer@axa.com](mailto:roland.scharrer@axa.com)

---

Chief Technology Innovation Officer



*Tous droits réservés – AXA GROUP OPERATION  
81 rue mstislav rostropovitch 75017 Paris - 2019*

*« Le Code de la propriété intellectuelle interdit les copies ou reproductions destinées à une utilisation collective. Toute représentation ou reproduction intégrale ou partielle faite par quelque procédé que ce soit, sans le consentement de l'auteur ou de ses ayant droit ou ayant cause, est illicite et constitue une contrefaçon, aux termes des articles L.335-2 et suivants du Code de la propriété intellectuelle. »*

*Achévé d'imprimer en 2019*

