A Study on Algorithmic Discrimination and its Effects on the Financial Sector

Thomas Graf*
thomas.graf@icloud.com
Hochschule Harz
Wernigerode, Sachsen-Anhalt, Germany

Abstract

In this study, we conducted a brief survey on the topic of algorithmic discrimination and explored potential methods and metrics for preventing and measuring this issue. We discovered several methods that have been developed through research over the past few decades. Despite the availability of effective methods, we found that algorithmic and data-based discrimination remains a highly topical problem, as demonstrated by a non-exhaustive set of examples. We also examined the issue of algorithmic discrimination in the financial industry and its potential impacts on society and the industry itself.

Keywords ai, machine learning, discrimination, prevention, banking, regulation

1 Introduction

The rise of AI is a cross industry wide phenomenon that will impact all areas and business processes of companies over the next decades. This will require a workforce which is able to understand, apply and implement AI and ML algorithms in an appropriate way. Based on the demand and hunger for these experts thousands of Computer Science and IT related students are deepening their knowledge in the field. Once these students finish their studies and begin their new careers as Data Scientist, they usually start applying AI and ML algorithms to real-world data. This data is often obtained from the massive datasets collected by their new employers, and may include sensitive client information. The purpose of using this data is to improve business operations and ensure they are more effective, efficient, and secure.

The issue with this situation is that applying algorithms in a careless manner to sensitive client data can cause several risks. These risks are intrinsic to highly sensitive industries like finance, banking, or insurance, as they involve clients' vulnerable areas such as their personal savings and health. One significant but often ignored risk is the discrimination of clients based on their personal information. Although extensively researched for years, our practical experience suggests that this issue is not adequately addressed in the industry's activities when actively applying ML and AI to their customer data.

For the sake of completeness, we want to remind the reader of the definition of unwanted discrimination in general, as proposed by Hajian and Domingo-Ferrer [2013]: *"In sociology, discrimination is the prejudicial treatment of an individual based on their membership in a certain group or category. It involves denying to members of one group opportunities that are available to other groups."*

Nevertheless we have to keep in mind that *"By definition, data mining is always a form of statistical (and therefore seemingly rational) discrimination."* but *"...data mining holds the potential to unduly discount members of legally protected classes and to place them at systematic relative disadvantage."* [3]

2 Algorithmic Discrimination in Real Life

Discrimination actually occurs in our day-to-day lives. In this section, we aim to present some real-world examples to emphasize the importance of the issue. The first case we discuss is not driven by AI, but rather by users who have access to sensitive data. This is a discrimination case observed on the online rental marketplace Airbnb by Edelman and Luca [2014]. The study found that non-black host are able to charge around 12% more in rental fees compared to black hosts who are offering a similar location, quality and other characteristics. Moreover black hosts are more severely penalized when offering lower-quality rental location. This case drastically illustrates the power of data when it comes to discrimination and the resulting economic impact on individuals affected by it.

Another case of discrimination which was driven by data and underlying algorithms was demonstrated by Lambrecht and Tucker [2018]. In this case, it was found that an algorithm for displaying job offers related to STEM fields significantly favored men over women. The study found that 20% more men than women saw these job offers, specifically in the age group of 25-54 years. The study identified higher targeting costs as a root cause of this situation. It turned out that women in the mentioned age group are a very demanded target group for product sales as they typically more likely click an ad and buy the advertised product. This in general raises the costs for showing ads to this group in general. As a result, the ad algorithm attempted to optimize the costs for the job offer ads, which led to fewer ad impressions for women. In addition to this, Datta et al. [2015] were able to demonstrate that females were distinctly discriminated against by showing them fewer instances of job ads related to coaching services for high-paying jobs. This situation was proven using an automated tool that can mimic specific user profiles, including the users' genders.

Furthermore, Buolamwini and Gebru [2018] tested a skin-type classification dataset using three commercial gender classification systems. The study showed that the trained classification systems had a dramatically higher misclassification rate of darker-skinned females at 34.7%, compared to a misclassification rate of just 0.8% for lighter-skinned males. Based on these and other observations, Bacchini and Lorusso [2019] argue that these systems, which are typically trained on biased datasets, reflect the society and its racist patterns. This has led to people of color being disproportionately stopped, investigated, and arrested by the police, which increasingly applies facial recognition software in many Western countries.

Finally, we would like to mention a well-known case from the banking sector, which was brought to light by the Danish developer David Heinemeier Hansson. The case was cited by Telford [2019]

in the Washington Post. However, no specific studies on the matter could be found, indicating that it was not further publicly researched. It relates to the partnership between Apple and Goldman Sachs, which made the so-called "Apple Card" available. This is a credit card offered by Apple to its customers, where Goldman Sachs acts as the issuing financial institution.

Hansson complained that a credit line increase for the Apple Card of his wife was rejected, and his credit line was 20 times higher than that of his wife, despite Jamie Hansson having a higher credit score than he does. Other customers and even Apple employees reported similar issues, indicating that the underlying algorithm made by Apple was systematically discriminating against women, which had a direct financial impact on them. This observation resulted in an investigation by the New York State Department of Financial Services.

3 Approaches for Detection and Prevention

The introduction provides insight into why discrimination is inherent in technical systems in general. In the following sections, we aim to delve deeper into this aspect and discuss studied approaches on detecting, preventing, and measuring discrimination.

3.1 Definition of Discrimination

Moving beyond the real-life cases, we first want to clarify what discrimination actually entails. To accomplish this, we need to start by considering the legal perspective and then supplement this viewpoint with researchers' approaches.

3.1.1 Legal Definitions

An early legal definition of discrimination based on sensitive personal data can be found in a 1971 judgment of the US Supreme Court. *"the Act makes it an unlawful employment practice for an employer to limit, segregate, or classify employees to deprive them of employment opportunities or adversely to affect their status because of race, color, religion, sex, or national origin..."* [41]

Newer judgments of the US Supreme Court have further enhanced this definition by making it clear that even if discrimination was not intended, it still constitutes discrimination when minorities face a disparate impact. *"...Title VII prohibits intentional acts of employment discrimination based on race, color, religion, sex, and national origin, 42 U. S. C. §2000e–2(a)(1) (disparate treatment), as well as policies or practices that are not intended to discriminate but in fact have a disproportionately adverse effect on minorities..."*[42]

In Germany, the *"Allgemeines Gleichbehandlungsgesetz"* (Anti-Discrimination Law) outlines similar parameters to help understand what discrimination entails. In summary, it states that any act that places individuals at a disadvantage based on their race, ethnic origin, gender, religion, disability, age, or sexual identity is an illegal disparate act. [19]

3.1.2 Research Approaches on Definitions

In addition to these legally committed definitions, researchers have proposed the terms *"Explainable Discrimination"* and *"Non-Explainable Discrimination"* [34]. Explainable discrimination is defined as a situation where non-personal attributes are driving the separation in final predictions. The authors cited a case where the higher income of male individuals can be explained by the fact that females work fewer hours, and on average, both groups still have the same income. However, this view can be criticized, as it may reflect hidden discriminatory social patterns that force females into situations with fewer working hours, resulting in lower absolute income. Non-explainable discrimination, on the other hand, is the opposite of the previous concept and refers to situations where discrimination cannot be explained and is therefore illegal.

Based on our current review of studies, we prefer to follow the earlier described legal definitions, along with some definitions that are more obvious and easier to explain, such as direct and indirect discrimination [49]. Direct discrimination can easily be explained as the case where protected attributes lead to *"...less favorable treatment.*. [49] and indirect discrimination can be explained as *"...situation where the treatment is based on apparently neutral nonprotected attributes but still results in unjustified distinctions against individuals from the protected group."* [49]. These concepts will reappear in subsequent sections when we begin discussing the reasons for AI discrimination in greater detail.

3.2 Reasons for AI Discrimination

Before exploring potential detection and prevention mechanisms, it is crucial to determine the underlying reasons for the unwanted behavior of applied ML and AI systems. The literature points to several directions from where discriminatory influences may arise. However, the main issue appears to be the underlying data used for algorithmic training and its handling. Zhou et al. [2021] identify five categories of potential data-related biases:

- *"Bias in historical data..."*
- *"Bias in data collection mechanisms..."*
- *"Bias in alternate source of data..."*
- *"Unobservable outcomes..." (labels)*
- *"Bias in unstructured data and feature engineering..."*

The most widely accepted source of biases exists in the area of (historical) data used to train the algorithms. Zhou et al. [2021] are pointing out that historical data is *"...often skewed towards, or against, particular groups."* and it can furthermore be *"...severely imbalanced with limited information on protected groups."*. Similarly Wang et al. [2021] states that *"Underrepresentation and misrepresentation of protected groups in the training data is a significant source of bias for Machine Learning (ML) algorithms, resulting in decreased confidence and trustworthiness of the generated ML models."*

One way to actively introduce bias into systems is by collecting data (un)intentionally from a biased group of people due to the mechanism used for data collection (e.g., data collection from smartphone users). The potential bias in *"...alternate source of data..."* can be proven by Berg et al., Bertrand and Kamenica [2019, 2018]. Metadata, often collected in a fully automated manner, can act as a proxy for a person's socioeconomic status. For instance, webpages can detect user information regarding devices and shopping behavior. Bertrand and Kamenica [2018] showed that more educated and higher-income groups tend to prefer technological products that are associated with innovation, such as iPads or iPhones. This leads to the conclusion that people using iOS devices are less likely to default on loan payments compared to Android users [6].

Similarly, [50] points to the issue that datasets intrinsically represent just the available observations. But these observations may be limited to a certain group due to the fact that other groups may just be underrepresented (e.g., as consumers for a certain product

Name	Definition	Constraints
Ballanced Error Rate [28]	$BER(f(Y),X) = \frac{Pr[f(Y)=0\|X=1]+Pr[f(Y)=1\|X=0]}{2}$	ε has to be defined; assumes exclusion of sensitive attributes
Equalized Odds [31]	$Pr\{f(X_{p,u})=1\|X_p=0,T=t\} = Pr\{f(X_{p,u})=1\|X_p=1,T=t\}$ $t \in \{0,1\}$	assumes inclusion of sensitive attributes; not considering distributions, incorrect sampling or misclassification
Equal Opportunity [31]	$Pr\{f(X_{p,u})=1\|X_p=0,T=1\} = Pr\{f(X_{p,u})=1\|X_p=1,T=1\}$	same as above
Average Odds Difference [11]	$AOD = [(FPR_U - FPR_P) + (TPR_U - TPR_P)]$	same as above + useful thresholds to be defined
Equal Opportunity Difference [11]	$EOD = TPR_U - TPR_P$	same as above
Decision Boundary Covariance [48]	$Cov(z, d_0(x)) \approx \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i)$	designed for convex margin-based classifiers; sensitive attributes to be excluded

Table 1. Overview Discrimination Detection Metrics

as this product may be too expensive or not interesting for other reasons).

Obviously, the mentioned items are all related to data-induced biases. Beyond that, we have to consider also an algorithmic bias. This kind of bias is introduced by model design choices made by data scientists or similar and represents preferences of these experts on e.g., choosing a preferred objective function or similar. *"A key reason why model design choices amplify algorithmic bias is because notions of fairness often coincide with how underrepresented protected features are treated by the model."* [32]

This overview of potential root causes for AI discrimination shows that collecting and handling the data is the major source of AI discrimination. As a next step, we want to evaluate methods and solutions that may help with detecting potential discrimination.

3.3 Discrimination Detection and Metrics

The detection of potential discrimination is driven by applying useful metrics to the output of the applied statistical prediction methods. This section gives a high-level overview of metrics provided by recent research, but it starts with a view provided by U.S. public authorities. To obtain an overview of all other metrics, see Table 1.

While the legislators and high courts in the U.S. as well as the EU have failed to define clear metrics to measure discrimination in general, the U.S. Equal Employment Opportunity Commission (EEOC) proposed the four-fifths rule (also 80 percent rule) in 1978. *"A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact."* [44]

3.3.1 Balanced Error Rate

Based on the previous work by Feldman et al. [2014], the 80% rule was transformed into a statistical measure that can be applied to all types of classifiers used in statistics and machine learning. The idea of a binary classification with the outcome C defined as "Yes" or "No" and a sensitive class feature X defined as "0" for the minority and "1" as the default class can be represented in the following confusion matrix shown in Table 2.

Outcome	X = 0	X = 1
C = No	a	b
C = Yes	c	d

Table 2. Confusion matrix [28]

Therefore Feldman et al. [2014] are defining the 80% rule as:

$$\frac{c/(a+c)}{d/(b+d)} \geq 0.8 \tag{1}$$

Alternatively to this rule Feldman et al. [2014] are proposing a metric called the *"Balanced Error Rate"* (BER). Given a dataset *D = (X,Y,C)* where X are the sensitive and to be protected class attributes, Y are the remaining non-sensitive attributes and C is class which is to be predicted. If $f : Y \rightarrow X$ is a function that can predict the protected attributes X from Y then the BER is defined as *"...the (unweighted) average class-conditioned error of f..."*:

$$BER(f(Y),X) = \frac{Pr[f(Y)=0|X=1] + Pr[f(Y)=1|X=0]}{2} \tag{2}$$

To determine predictability it is said that X is predictable from Y if there is a function $f : Y \rightarrow X$ with $BER(f(Y),X) \leq \varepsilon$. Vice versa X is not predictable from Y if there is a function $f : Y \rightarrow X$ with $BER(f(Y),X) > \varepsilon$.

The weakness of the BER approach is to find a useful threshold for ε. Based on the 80% rule Feldman et al. [2014] are proposing $\varepsilon = \frac{1}{2} - \frac{\beta}{8}$ based on $\beta \widehat{=} \frac{c}{a+c}$. It has to be mentioned that Feldman et al. [2014] have identified scenarios where β could be artificially small which would move ε to 0.5 and this *"...bound tends towards the trivial since any binary classifier has BER at most..."* 0.5. Nevertheless the overall advantage of the BER is that it is an intuitive and easy to understand approach.

3.3.2 Equalized Odds & Average Odds Difference

A more advanced approach compared to the BER is the concept of *"Equalized Odds"* (EOS) proposed by Hardt et al. [2016]. Given the definition EOS is achieved when the outcome of a prediction $f(X_{p,u})$ is independent to the protected attributes X_p conditional to the true value of the target T. The idea of the concept is to ensure that the protected attributes X_p can be included because a removal

of the sensitive attributes could cause lower accuracy or other issues which will be discussed later in this paper. The integration of the sensitive attributes is done under the condition of the actual true value which is typically trained on in machine learning. The final outcome is that the True-Positive Rate and False-Positive Rate should be equal for the minority and majority group:

A more advanced approach compared to the BER is the concept of *"Equalized Odds"* (EOS) proposed by Hardt et al. [2016]. Given the definition EOS is achieved when the outcome of a prediction $f(X_{p,u})$ is independent to the protected attributes X_p conditional to the true value of the target T. The idea of the concept is to ensure that the protected attributes X_p can be included because a removal of the sensitive attributes could cause lower accuracy or other issues which will be discussed later in this paper. The integration of the sensitive attributes is done under the condition of the actual true value which is typically trained on in machine learning. The final outcome is that the True-Positive Rate and False-Positive Rate should be equal for the minority and majority group:

$$\begin{aligned} Pr\{f(X_{p,u}) = 1 | X_p = 0, T = t\} = \\ Pr\{f(X_{p,u}) = 1 | X_p = 1, T = t\}, \\ t \in \{0, 1\} \end{aligned} \tag{3}$$

Leveraging the concept of the EOS Chakraborty et al. [2020] are proposing the *"Average Odds Difference"* (AOD) to establish a measure which can quickly be interpreted. The AOD is defined as

$$AOD = [(FPR_U - FPR_P) + (TPR_U - TPR_P)] \tag{4}$$

It also applies the True-Positive and False-Positive Rates of the protected and unprotected groups and creates the sum of the differences of both values. If the result is close to 0 it can be interpreted as equal treatment by the classifier for both groups. If it is higher than 0 the majority group is preferred and vice versa.

3.3.3 Equal Opportunity & Equal Opportunity Difference

The second concept proposed by Hardt et al. [2016] is the *"Equal Opportunity"* (EOP) which is a less strict version of the EOS as it is only focussing on the positive prediction class (e.g. 1). In that sense the positive prediction means that the outcome causes a positive impact on the related group or person (e.g. a loan is authorized). According to Hardt et al. [2016] that helps to make sure *"... that people who pay back their loan, have an equal opportunity of getting the loan in the first place (without specifying any requirement for those that will ultimately default)"*. EOP simply represents the equality of true positives rates for the minority and majority group:

$$Pr\{f(X_{p,u}) = 1 | X_p = 0, T = 1\} = Pr\{f(X_{p,u}) = 1 | X_p = 1, T = 1\} \tag{5}$$

As before on top of that idea Chakraborty et al. [2020] are proposing the *"Equal Opportunity Difference"* (EOD) which they define as the difference of the True-Positive Rates for unprivileged and privileged groups:

$$EOD = TPR_U - TPR_P \tag{6}$$

Its interpretation is similar to the AOD but one has to keep in mind this setting fully ignores the "negative" class and with that the False-Positive Rate. It sets the focus like the EOP only on the positive prediction class.

3.3.4 Decision Boundary Covariance

An approach which can be applied to margin-based classifiers is the *"Decision Boundary Covariance"* proposed by Zafar et al. [2015]. It is defined as the covariance between sensitive attributes $\{z_i\}_{i=1}^N$ and the signed distance $d_\theta * (x)$ from the training (non-sensitive) attributes $\{x_i\}_{i=1}^N$ to the decision boundary $\{d_\theta(x)\}_{i=1}^N$:

$$Cov(z, d_0(x)) \approx \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) d_\theta(x_i) \tag{7}$$

This is done under the constraint that sensitive attributes are not part of the dataset when the classifier is trained, given the fact that these attributes could be correlated with the class label or indirectly with other features. The DBC aims to be included in the formulation of margin-based classifiers to optimize the classifier with respect to fairness and accuracy.

3.4 Discrimination Prevention

The prevention of algorithmic discrimination can be divided into three major categories. Based on recent scientific literature, these methods can be grouped into pre-processing, in-processing, and post-processing approaches. In this section, we will provide an example of each of these categories to gain a deeper understanding of each area. To obtain a more complete overview of recent methods in science, see Table 3. We will also take a closer look at the idea of removing sensitive attributes from datasets to prevent discrimination.

3.4.1 Pre-Processing

The idea of pre-processing methods is to prevent algorithmic discrimination by removing the bias from the underlying training data. This class of methods acts preventively and aims to solve the problem before algorithms are applied to the problematic data. One specific pre-processing option is the data re-labeling approach proposed by Kamiran and Calders [2011]. The idea is to re-label the target variable for some candidates to decrease the persistent discrimination of the dataset. Parts of the majority group with positive target variable are re-labeled to the negative class (demotion) and part of the minority group with negative class label are re-labeled to the positive class (promotion). The candidates are selected via ranking. The ranker is trained according to the positive class probability. Based on that, the ranker provides a score which helps with identifying the relevant candidates. Promotion candidates are sorted in descending order based on the score and the demotion candidates are sorted vice versa. The first top elements of both lists are chosen and re-labeled. The procedure is repeated until the discrimination disappears. Kamiran and Calders [2011] calculate the number of necessary modification (M) steps as follows:

$$M = \frac{(b \times (\overline{b} \wedge +)) - (\overline{b} \times (b \wedge +))}{b + \overline{b}} \tag{8}$$

Where *"...b and $\overline{b}$ represent respectively the number of objects with $B = b$ and $B \neq b$ while $b \wedge +$ and $\overline{b} \wedge +$ are the number of object of class $+$ such that $B = b$ and $B \neq b$, respectively"*. Beside their re-labeling option Kamiran and Calders [2011] are proposing two options of re-sampling to remove the bias in the training data. These options are listed in our overview of methods as alternative approach.

Method	Processing Type	Method Type	Constraints
anti-classification (removal of sensitive attributes) [17]	pre-processing	bias removal	no usage of protected attributes risk of red-lining
changing distribution of data [10]	pre-processing	bias removal	protected attributes required intrusive on data level
relabeling of data [33]	pre-processing	bias removal	protected attributes required intrusive on data level may lead to wrong results
resampling of data [33]	pre-processing	bias removal	protected attributes required
re-weighting regression coefficents [46]	in-processing	model bias removal	protected attributes required only for regression problems works only for numerical variables applies to protected attributes
decoupled classifiers [20]	in-processing	classifier splitting	protected attributes required groups to be chosen carefully
decoupled classifiers [45]	in-processing	classifier splitting	same as above
deriving fair classifier from a discriminatory model [31]	post-processing	deriving classifier	protected attributes required data labels have to be accurate
debiasing word embeddings regarding gender discrimination by neutralizing gender neutral words and by equalizing words outside of gender subspace [31]	post-processing	model bias removal	intrusive on model level

Table 3. Overview Discrimination Prevention Methods

3.4.2 In-Processing

In contrast to pre-processing approaches, the class of in-processing methods is meant to address discrimination problems by adjusting the way algorithms are applied to biased data. Instead of fixing the problem on the data level, the idea is to enable the applied algorithms to handle problematic data. Ustun et al. [2019] developed the idea of classifier splitting for classification problems. The idea is to train a classifier per demographic group available in the dataset. The goal is to minimize both the training error (empirical risk) and the generalization error (true risk). This should be achieved by creating a classifier for each denoted group, as these classifiers should perform better than a pooled classifier that does not use sensitive information. The approach requires that the trained classifiers guarantee (1) rationality and (2) envy-freeness, which means *"that each group should prefer their assigned model to (i) a pooled model that ignores group membership (rationality) and (ii) the model assigned to any other group (envy-freeness)"* [45]. The challenge is to correctly select the groups on which the classifiers should be trained when more than one sensitive attribute is available. Some choices may be inappropriate. Therefore, it might be necessary to capture the groups at different levels of detail. A group selection could be, for example:

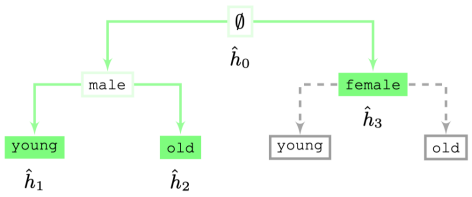

Figure 1. Group selection tree example. [45]

3.4.3 Post-Processing

The class of post-processing methods wants to avoid complex changes belonging to the data or the algorithm training. The main idea is to accept that both items are biased but based on that knowledge unbiased algorithms can be derived or the bias can be corrected. Hardt et al. [2016] proposed to derive an equal opportunity predictor from a classifier or score which is potentially discriminatory. The target is to find a fair classifier via a simple post-processing step without impacting the training process. *"A predictor $\widetilde{Y}$ is derived from a random variable R and the protected attribute A if it is a possible randomized function of the random variables (R, A) alone."* [31] So *"...$\widetilde{Y}$ is independent of X conditional on (R, A)"*. So $\widetilde{Y}$ depends finally on the joint distribution of (R, A, Y). And this joint distribution is required to construct $\widetilde{Y}$. The solution can derive $\widetilde{Y}$ from a binary predictor or a real valued score function. $\widetilde{Y}$ is found at the end by defining a convex hull of 4 vertices. It is derived if the false-positive rate of $\widetilde{Y}$ is element of the convex hull of $\widehat{Y}$.

3.4.4 Removing Sensitive Features and Dataset Rebalancing

The removal of sensitive features is often seen as an approach that can be easily applied to avoid the problem of algorithmic discrimination. However, recent research has shown that this may not be the case. Kamiran and Calders [2011] points to potential redlining effects that may arise when sensitive attributes are removed and remaining attributes have a hidden correlation with sensitive attributes. The result is a persisting indirect discrimination within the dataset. The term redlining comes from the practice of *"...denying inhabitants of certain racially determined areas from services such as loans. It describes the practices of marking a red line on a*

map..." [33]. Žliobaitė and Custers [2016] points in the same direction and shows that red-lining becomes a significant problem, especially for regression algorithms such as linear regression. They argue that the commonly suggested approach of avoiding the collection of sensitive attributes to sanitize regression models may not be feasible. This problem arises when sensitive data is correlated with non-sensitive attributes, which introduces the omitted variable bias. Similarly, Hardt et al. [2016] have shown that indirect discrimination can arise due to correlation with other features. They also argue that the idea of rebalancing the data to achieve demographic parity may lead to incorrect decisions, such as accepting unqualified individuals in A=1 but rejecting qualified applicants in A=0.

4 Algorithmic Decisions in the Financial Industry

This section explores the use of AI, ML, and other statistical algorithms in the financial industry, examining both the potential opportunities and risks associated with their use in banking.

4.1 Opportunities for the Financial Industry

As in most other industries the use of algorithms in the financial industry has a significant number of potential use cases. Often named areas are [12, 40]:

- risk management
- alpha generation
- customer interaction
- relationship manager augmentation
- fraud detection & AML applications
- hiring
- algorithmic trading and others

The basic idea is to leverage algorithms to automate processes to save costs on the one side and to increase the business volume on the other side. A key for financial institutions is the general area of risk management. Risk management in banking ranges from financial risks (e.g. credit defaults) over operational risks (e.g. cyber security) to compliance and regulatory risks (e.g. financial crime) and anything in between. As this area is strongly regulated, financial institutions hope that AI and algorithms in general can help improve their activities in the context of so-called RegTech applications [1]. The hope is to achieve better, faster, and less manual decisions and processes in this area.

Apart from the regulatory and cost side, another idea of the industry is to improve market chances and increase revenues by using new algorithms. One of the recent hot topics in the financial industry is to sell new "sustainable" products with a focus on ESG (Environment, Social, Governance) criteria [36]. The idea is to use new algorithms to analyze large datasets about companies to categorize them and identify their ESG potentials and risks to be marketed to investors [12]. The financial industry also sees a lot of potential in other areas that were mentioned already in different contexts in this paper. Classic banking business like credit risk scoring, for example, should be improved by algorithms to decrease credit defaults [5]. This increases the efficiency of the business while potential applications like improved customer interactions via chatbots [12] or automated investment advisors [27] are meant to improve the effectiveness.

4.2 Risks for the Financial Industry

In contrast to the opportunities, there are also several risks associated with the use of algorithms and AI in all industries, which must be carefully considered and mitigated. However, the risk topic is particularly significant for the financial industry. On the one hand, the industry heavily relies on the trust and confidence of customers in the financial system. On the other hand, financial institutions are crucial to the economic system in most countries around the world. Therefore, it is not surprising to see the following items that describe key risk areas for financial institutions when using AI and algorithms [37, 38]:

- data privacy violations and cross border issues
- insufficient data reliability and relevance
- algorithmic and data bias
- algorithmic discrimination
- lack of model explainability
- uncertainties in model robustness
- insufficient and outdated governance frameworks
- missing regulatory compatibility
- systemic risk of instability

As expected, the issue of algorithmic discrimination is on the list of potential risks. In the following sections, we will take a closer look at the implications of algorithmic discrimination when it occurs in the financial industry.

4.3 The Impact of Algorithmic Discrimination on the Financial Industry

Algorithmic discrimination is not a new topic to the industry *"...the financial sector has a checkered history of mixing risk-based discrimination at heart of their businesses with socially detrimental bias. However, the use of AI raises fresh concerns by establishing new vectors for the introduction of bias into decisions..."* [37]. There are a number of key areas in the industry which suffer most from the disparate impact of algorithms.

4.3.1 Inequality in access to credit

As already implied with some examples in section 2, algorithmic discrimination happens every day, also in the financial industry. One of the financial products with the most consistent growth is consumer loans and credit cards [8, 22]. In many countries, consumers rely on credit to finance a significant proportion of their consumption, such as cars, houses, education, or other products they cannot pay for immediately from their income or savings. If financial institutions apply algorithms to decide whether a customer gets access to a loan, potential algorithmic discrimination can remove certain groups from that access [13]. Possible consequences could be a reduced economic potential within a society due to reduced consumption. Or, in the worst case, it could inhibit the development of minority groups by reducing their access to goods that allow them to increase their income, such as cars for commuting or private education.

4.3.2 Reinforcement of wealth inequality

One of the worst consequences of customer discrimination in the financial sector is the exacerbation of existing wealth inequalities. Studies have shown discrimination by both face-to-face and algorithmic lenders. Although algorithmic lenders have been shown to reduce discrimination, it is still significant. Borrowers from black

and Latino communities pay 5.3 to 7.9 basis points higher interest rates than white and Asian ethnicity borrowers, depending on the type of loan. This increases the cost to discriminated borrowers by up to $765 million a year. This translates into a higher profit for lenders of up to 17 percent compared to the profit made by non-discriminated groups. This situation is not only more expensive for the minority groups, it also constantly shifts wealth from poorer to richer groups [4]. This kind of situation reinforces the bias that persists in society and in the data that is collected from people. It reinforces poverty and potentially self-destructive behavior in the disadvantaged groups.*"...If a poor student cant get a loan because a lending model deems him too risky (by virtue of his zip code), he's then cut off from the kind of education that could pull him out of poverty, and a vicious spiral ensues."* [39]

4.3.3 Reputational and legal risks

All of the aforementioned aspects create a significant risk of reputational and legal consequences for financial institutions. Recalling the example of Goldman Sachs [43] in section 2, it can be concluded that the reputation of financial institutions can be quickly damaged. This is a major concern for banks because, more than any other industry, they rely on the trust of their customers. Also legal risks are arising due to available regional regulations. US laws like the Equal Credit Opportunity Act or the Fair Housing Act have caused significant fines for local financial institutions like e.g. to the Newark's Park National Bank [29].

5 Regulation of Automated Decisions in Banking

The financial sector in general is a highly regulated industry. There are many regulations that directly or indirectly cover the area of automated and algorithmic decisions. We would like to summarise the most relevant regulations available in the United States and the European Union.

5.1 US Regulations

The US regulations related to discrimination in the financial industry goes back to laws from the mid of the 20th century. The focus was to generally avoid discrimination of minority in financial business. So there was still no specific focus on the impact of applied algorithms or any specific technology which might impact decisions of the financial institutions. For example the "Fair Housing Act" [14] is part of the American "Civil Rights Act" and was created to avoid discrimination when it comes to businesses like buying or renting houses or e.g. applying for a mortgage loan. The "Fair Credit Reporting Act" [15] from the 1970s is one of the first regulations taking care for data privacy and it gave consumers the right to e.g. correct data which influences credit scores. Another relevant regulation which was also initiated in the 1970s is the "Equal Credit Opportunity Act" [16]. It disallows any discrimination related to credit businesses and provides wider rights for consumers when it comes to applying for loans.

5.2 EU Regulations

The EU regulation in that area is more recent and tries to consider also more modern technological aspects. This is reflected in regulations that are not specifically made for the financial sector but rather for algorithmic decisions in general on potentially resulting discrimination. A financial sector specific regulation instead is e.g. MIFID II [23] with the add on RTS6 [25] which regulates the aspect of algorithmic trading with the intention to avoid instabilities in the financial markets. The currently most consumer relevant law regarding automated decision making in the EU is GDPR [24] which requires fair and transparent algorithms and the right to human intervention. As a next level of regulations specifically for the application of AI currently the European Union discusses the AI Act [26]. It is one of the first laws specifically relating to the risks of AI ever and wants to reduce inherent risks appearing from AI which may impact private persons in a disadvantageous way.

6 Conclusion

Algorithmic discrimination is a well-known and extensively researched topic, with regulations addressing data-based discrimination in place since the late 1960s in the US and more recently in the EU. Despite this, algorithmic discrimination occurs daily across various industries, as demonstrated by the examples provided. This is particularly concerning for the financial industry, as it plays a central role in every major economy. Discriminatory algorithms used by banks can have significant socioeconomic impacts, as well as posing reputational and legal risks for the industry. It is therefore essential to raise awareness among data scientists in the financial industry about the ongoing issue of algorithmic discrimination and its causes. Data scientists should also actively implement the methods proposed in recent studies and research papers to address the problem. Many of these methods require the use of sensitive data to prevent discrimination. While recent regulations do not explicitly prohibit the use of sensitive information, individuals must consent to the use of their data under the according to GDPR in the EU. Future regulations must take into account the findings of current research to avoid exacerbating the problem. Future research should also investigate why discrimination persists despite the availability of regulations and prevention methods. Is the problem due to a lack of knowledge among those who apply the algorithms, or are malicious intentions at play? Alternatively, is the issue simply ignored to avoid the additional effort required to develop accurate predictive algorithms?

References

[1] Saqib Aziz and Michael Dowling. 2018. AI and Machine Learning for Risk Management. *SSRN Electronic Journal* (01 2018). https://doi.org/10.2139/ssrn.3201337

[2] Fabio Bacchini and Ludovica Lorusso. 2019. Race, Again. How Face Recognition Technology Reinforces Racial Discrimination. *Journal of Information, Communication and Ethics in Society* 17 (2019), 321–335.

[3] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. California Law Review, Inc. https://doi.org/10.2139/ssrn.2477899

[4] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2019. Consumer-Lending Discrimination in the FinTech Era. http://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf

[5] Pradeep Bedi, S B Goyal, and Jugnesh Kumar. 2020. Basic Structure on Artificial Intelligence: A Revolution in Risk Management and Compliance. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. 570–576. https://doi.org/10.1109/ICISS49785.2020.9315986

[6] Tobias Berg, Valentin Burg, Ana Gombović, and Manju Puri. 2019. On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *The Review of Financial Studies* 33, 7 (09 2019), 2845–2897. https://doi.org/10.1093/rfs/hhz099

[7] Marianne Bertrand and Emir Kamenica. 2018. *Coming Apart? Cultural Distances in the United States over Time.* Working Paper 24771. National Bureau of Economic Research. https://doi.org/10.3386/w24771

[8] Board of Governors of the Federal Reserve System. 2023. Consumer Credit - G.19. (2023). https://www.federalreserve.gov/releases/g19/current/ Online available; accessed at 15/02/2023.

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research* 81 (2018), 1–15.

[10] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf

[11] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: a way to build fair ML software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM. https://doi.org/10.1145/3368089.3409697

[12] Chi Chan, Dr. Christine Chow, Janet Wong, Nikolaos Dimakis, David Nayler, Jano Bermudes, Jayant Raman, Rachel Lam, and Matthew Baker. 2019. Artificial Intelligence Application in Financial Services. (2019). https://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2019/dec/ai-app-in-fs.pdf Online available; accessed at 26/10/2022.

[13] Danielle Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Washington Law Review* 89 (03 2014), 1–33.

[14] Congress of the United States. 1968. Fair Housing Act. (1968). https://www.govinfo.gov/content/pkg/COMPS-343/pdf/COMPS-343.pdf Online available; accessed at 04/03/2023.

[15] Congress of the United States. 1970. Fair Credit Reporting Act. (1970). https://www.ftc.gov/system/files/ftc_gov/pdf/545A-FCRA-08-2022-508.pdf Online available; accessed at 04/03/2023.

[16] Congress of the United States. 1976. Equal Credit Opportunity Act. (1976). https://uscode.house.gov/view.xhtml?req=granuleid%3AUSC-prelim-title15-chapter41-subchapter4&edition=prelim Online available; accessed at 04/03/2023.

[17] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *CoRR* abs/1808.00023 (2018). arXiv:1808.00023 http://arxiv.org/abs/1808.00023

[18] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings. https://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdf Online available; accessed at 01/11/2022.

[19] Deutscher Bundestag. 2021. Allgemeines Gleichbehandlungsgesetz. (2021). https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/AGG/agg_gleichbehandlungsgesetz.pdf?__blob=publicationFile

[20] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Vol. 81. 119–133. https://proceedings.mlr.press/v81/dwork18a.html

[21] Benjamin Edelman and Michael Luca. 2014. Digital Discrimination: The Case of Airbnb.com. Harvard Business School, Boston, MA. https://www.hbs.edu/ris/Publication%20Files/Airbnb_92dd6086-6e46-4eaf-9cea-60fe5ba3c596.pdf Online available; accessed at 19/10/2022; Working Paper.

[22] European Central Bank. 2023. 9 Volumes of outstanding amounts of euro-denominated loans to, and deposits from, euro area residents. (2023). https://sdw.ecb.europa.eu/reports.do?node=1000005104 Online available; accessed at 15/02/2023.

[23] European Parliament. 2014. MIFID II. (2014). http://data.europa.eu/eli/dir/2014/65/oj Online available; accessed at 04/03/2023.

[24] European Parliament. 2016. General Data Protection Regulation. (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj Online available; accessed at 04/03/2023.

[25] European Parliament. 2017. General Data Protection Regulation. (2017). http://data.europa.eu/eli/reg_del/2017/589/oj Online available; accessed at 04/03/2023.

[26] European Parliament. 2021. Artificial Intelligence Act. (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 Online available; accessed at 04/03/2023.

[27] Omar H. Fares, Irfan Butt, and Seung Hwan Mark Lee. 2022. Utilization of artificial intelligence in the banking sector: a systematic literature review. *Journal of Financial Services Marketing* (11 Aug 2022). https://doi.org/10.1057/s41264-022-00176-7

[28] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and removing disparate impact. arXiv. https://doi.org/10.48550/ARXIV.1412.3756

[29] Gill, Peter. 2023. Newark's Park National Bank reaches $9 million racial discrimination settlement in Columbus. (2023). https://eu.dispatch.com/story/news/2023/02/28/park-national-pays-9m-settlement-for-columbus-discrimination-case/69954190007/ Online available; accessed at 04/03/2023.

[30] Sara Hajian and Josep Domingo-Ferrer. 2013. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 25 (2013), 1445–1459.

[31] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv. https://doi.org/10.48550/ARXIV.1610.02413

[32] Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns* 2 (04 2021), 100241. https://doi.org/10.1016/j.patter.2021.100241

[33] Faisal Kamiran and Toon Calders. 2011. Data Pre-Processing Techniques for Classification without Discrimination. *Knowledge and Information Systems* 33 (10 2011). https://doi.org/10.1007/s10115-011-0463-8

[34] Faisal Kamiran and Indrė Žliobaitė. 2012. *Explainable and Non-explainable Discrimination in Classification.*

[35] Anja Lambrecht and Catherine Tucker. 2018. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260 Online available; accessed at 01/11/2022.

[36] Pedro Matos. 2020. *ESG and Responsible Institutional Investing Around the World: A Critical Review.* CFA Institute Research Foundation.

[37] Jess R McWaters. 2019. Navigating Uncharted Waters: A roadmap to responsible innovation with AI in financial services. (October 2019). https://www3.weforum.org/docs/WEF_Navigating_Uncharted_Waters_Report.pdf

[38] OECD. 2021. Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers. (2021). https://www.oecd.org/finance/artificial-intelligence-machine-learning-big-data-in-finance.htm Online available; accessed at 02/02/2023.

[39] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group.

[40] Hicham Sadok, Fadi Sakka, and Mohammed El Hadi El Maknouzi. 2022. Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance* 10, 1 (2022), 2023262. https://doi.org/10.1080/23322039.2021.2023262 arXiv:https://doi.org/10.1080/23322039.2021.2023262

[41] Supreme Court of the United States. 1971. Griggs v. Duke Power Co., 401 U.S. 424 (1971). (March 1971). https://supreme.justia.com/cases/federal/us/401/424/

[42] Supreme Court of the United States. 2009. Ricci v. DeStefano, 557 U.S. 557 (2009). (June 2009). https://supreme.justia.com/cases/federal/us/557/557/

[43] Taylor Telford. 2019. Apple Card algorithm sparks gender bias allegations against Goldman Sachs. *The Washington Post* (2019). https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/

[44] U.S. Equal Employment Opportunity Commission. 1978. Uniform guidelines on employee selection procedures. (August 1978). https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml

[45] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without Harm: Decoupled Classifiers with Preference Guarantees. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Vol. 97. PMLR, 6373–6382. https://proceedings.mlr.press/v97/ustun19a.html

[46] Indrė Žliobaitė and Bart Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24 (06 2016). https://doi.org/10.1007/s10506-016-9182-5

[47] Hao Wang, Snehasis Mukhopadhyay, Yunyu Xiao, and Shiaofen Fang. 2021. An Interactive Approach to Bias Mitigation in Machine Learning. 2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), 199–205. https://doi.org/10.1109/ICCICC53683.2021.9811333

[48] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2015. Fairness Constraints: Mechanisms for Fair Classification. arXiv. https://doi.org/10.48550/ARXIV.1507.05259

[49] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17.* 3929–3935. https://doi.org/10.24963/ijcai.2017/549

[50] Nengfeng Zhou, Zach Zhang, Vijayan N. Nair, Harsh Singhal, and Jie Chen. 2021. Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms. *International Statistical Review* 90, 3 (2021), 468–480. https://doi.org/10.1111/insr.12492 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12492