Towards Achieving Gender Equality in Automated Loan Approval Processes

By Amirali Khaleghi

Towards Achieving Gender Equality in Automated Loan Approval Processes

Master thesis submitted to Delft University of Technology

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in Management of Technology

Faculty of Technology, Policy and Management

by

Amirali Khaleghi

Student number: 4632036

To be defended in public on 03-18-2020

Graduation committee

Chairperson	: Dr. M.E. Warnier, Multi-Actor Systems
First Supervisor	: Dr. H. Asghari, Multi-Actor Systems
Second Supervisor	: Dr. J. M. Durán, Ethics/Philosophy of Technology
External Supervisor	: Mr. B. Timmermans, IBM, Lead Center for Advanced Studies

Acknowledgment

Dear reader,

This thesis is the culmination of my journey of Masters, one of the most challenging assignments I have faced so far. After 6 months of hardship, frustration, and yet excitement and fulfillment, I feel proud and honored to be given the opportunity to contribute to one of the most challenges that us humans face today: unjust access to opportunity in consumer lending. This research was generously supported by the Think Forward Initiative (TFI), a collaboration that includes IBM, ING, and others. The purpose of TFI is to empower people to make better financial decisions. This research supports this goal by exploring one key factor in these decisions – the perception of fairness of advice and decisions generated by algorithms or AI.

I would like to thank all of the members of my Graduate Committee starting with my first supervisor, Dr. Asghari, who provided excellent guidance and mentorship and inspired me in every step of the way. Next, I wish to thank Dr. Durán for teaching me, in the shortest span of time, how to approach, study, and apply philosophy and ethics to my work. This was truly one of the greatest challenges I faced throughout this project. Last but not least, I am particularly indebted to the committee's chairman, Dr. Warnier, whose deep knowledge, insights, and experience guided me through this work.

The support I received throughout this project extends beyond the university. I feel honored for being given the opportunity to work with some of IBM's most knowledgeable and experienced staff members. I wish to give my special thanks to Mr. Timmermans, Ms. Van Maare, Mr. Varshney, and Mr. Hind for providing me with day-to-day support in every technical and nontechnical aspect of the project.

Finally, I wish to express my very profound gratitude to my parents and to my partner, Alexandra, for blessing me with your boundless support, love, and continuous encouragement throughout this endeavor.

I now realize more than ever than pushing the boundaries would not have been possible without the support I received from all of you.

Amirali Khaleghi
Delft University of Technology
March 11, 2020

Executive Summary

The consumer lending domain has increasingly leveraged Artificial Intelligence (AI) to make loan approval processes more efficient and to make use of larger amount of information to predict their applicants' repayment ability. Over time, however, valid concerns have been raised about whether decisions made about individuals using these data-driven technologies can lead to bias against women. For instance, if algorithms are trained on datasets that reflect the disparities between men and women, the decision outcomes of the algorithms may become discriminatory. Even if datasets are free of biases, subjective decisions taken during the stages in which algorithms are developed and deployed can lead to discrimination. Therefore, it is important for financial institutions to have mechanisms in place to assess the fairness of their automated loan allocation models with respect to gender.

In an attempt to assess the fairness of an algorithm, 21 prominent definitions of fairness have been proposed by the computer science community over the years. However, what remains absent is consensus on which definitions are suitable for assessing gender equality in consumer lending. There is also a lack of knowledge on how to appropriately implement these metrics in practice. Apart from these issues, applying fairness metrics only reveals biases in the datasets and models without providing abundant information on what the root causes of bias of might be and where in the development and deployment machine learning process they can be encountered.

To tackle the problems mentioned above, this research has investigated *how automated loan approval processes can be assessed for gender equality.* Two essential elements for assessing predictive tools were identified and investigated through a separate research question:

- *Definitions of fairness.* What fairness metrics are suitable for assessing gender equality in consumer lending?
- *Detecting bias.* How can the metrics be applied to observe gender bias in lending history data?

Based on the questions above, the research was conducted in two stages: Stage 1 focused on analyzing the 21 prominent definitions of fairness, but before doing so, it conceptualizes gender equality in consumer lending by conducting an extensive literature review encompassing domains of philosophy, economics, gender studies, and history.

In investigating the first research question, it is found that group fairness metrics are a measure of distributive justice. These metrics are based on three different statistical criteria commonly known as *independence, sufficiency, and separation.* These criteria cannot be achieved at the same time except for a very narrow range of circumstances. This implies that some discrimination maybe *unavoidable*. The choice of fairness criteria should be based on the application scenario at hand and corresponding governing laws. *Independence* assumes that all individuals have the same effort-based utility. As such, the choice of independence can lead to distribution of the same rate of loans between the two demographic groups. By looking at *separation* through the above definition of equality of opportunity, it is found that separation assumes the predictions to be the harm/benefit applicants receive while true labels are what justifies inequality. By looking at *sufficiency* through the above definition of equality of opportunity, it is found that sufficiency assumes the true labels or ground truth to be the harm/benefit applicants receive while predictions are what justifies inequality.

Individual metrics are also a measure of distributive justice. Unlike group fairness they do not depend on any statistical criteria, however, they require the stakeholders to clearly formulate what similarity means and how it can be measured; such choices can lead to implicit bias. *Causal measures* can be thought of as a measure of one of the rules procedural justice; they check whether features that are selected to train the predictive models are discriminatory and/or illegitimate. These measures are beneficial for understanding the relationships between gender and other features in the dataset.

Moving on, in Stage 2 of this work the second research question was investigated by conducting an exploratory case study. The three key elements of the case study are:

- *German Credit Data:* sample containing information on 1000 loan applicants with their binary true label classes (ground truth).
- *AIF 360 toolkit*: IBM's extensive fairness toolkit for detecting and mitigating bias.
- *Taxonomy of bias:* A comprehensive list of possible causes of bias.

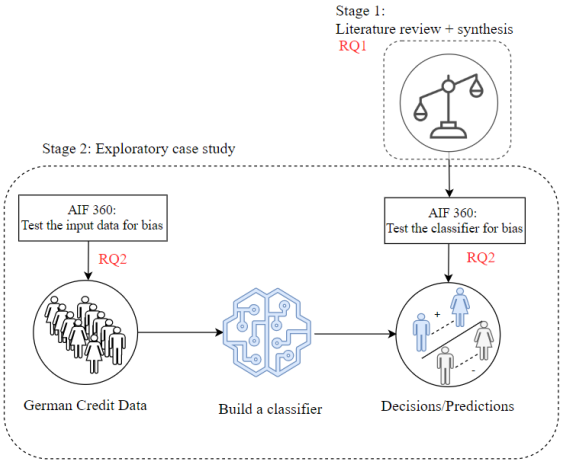


Figure A – Research paradigm

The case study started with an overview of the toolkit to determine what metrics are supported for testing bias in the German Credit Data. In order to test the classification model for bias, firstly, a logistic regression model was developed and optimized to attain the highest balanced accuracy. The model was then tested for group fairness. It was found that the choice of separation and sufficiency can have different repercussions for each demographic group in the German Credit Data. When *false distribution of utility* is under inspection, sufficiency advantages females more than males since it is more likely for male applicants who are assigned to negative class (what justifies inequality) to actually have good credit (utility). On the other hand, separation advantages males more than females since it is more likely for males who actually belong to the negative class (what justifies inequality) to receive a positive predicted score (utility). Such inconsistency highlights the importance of realizing how relevant distribution of harm/benefit depends on the choice of fairness criteria made by decision makers.

Finally, the taxonomy of bias was used as a framework to check for causes of bias which could be encountered during the model development process and results are demonstrated in the case study. To add practical value, the case study was conducted such that the reader can walk through common sequential stages of classifier development and observe *where* and *how* subjective choices made can lead to bias. Subjective choices often are taken as a trade-off between accuracy and fairness. Some of the key lessons are that, firstly, that enforcing fairness constraints reduces accuracy due to diverting the primary goal of the machine learning process from accuracy, only, to both fairness and accuracy.

Over the years, a great body of literature has been dedicated to tackling algorithmic bias issues with the goal to provide an all-encompassing solution, but only some investigate algorithmic bias issues in a context-targeted manner. Thus, one of the main contributions of this work lies in forming a cognitive process which can help with the interpretation of fairness metrics based on factors (gender equality in consumer lending) that lie outside of the domain of computer science. Furthermore, it conducts a study of fairness metrics and causes of bias in a manner specific to the context of lending. This work attempts to introduce a gateways/new line of thinking that incorporates knowledge from various fields that have contributed to achieving fair decisions for individuals.

This page is intentionally left blank.

Table of Contents

Table of Contents

List of tables

List of figures

This page is intentionally left blank.

1 Introduction

Defined by innovation, artificial intelligence (AI) has been increasingly adopted across various industries to increase labor productivity and efficiency in task performance. AI is believed to have the capacity to transform entire economies and rewrite the material conditions of human existence. Such significant influence on human lives fuels the motivation to investigate positive and negative impacts of this technology.

AI concentrates on understanding, development, and commercialization of intelligent computational processes. What makes it so compelling to examine is the fact that it is expected to bring about a direct quantifiable impact: changing and displacing jobs through elimination of certain monotonous and repetitive tasks while simultaneously creating millions of other new roles across industries.

Another important consequence of AI, which, in contrast, has proven to be difficult to predict and quantify, is that its adoption may trigger conditions that do not align with our norms and values. This impact is known as *algorithmic bias* and is used to describe unfair prediction of outcomes that favor a group in a protected class such as on gender, race, religion, or color (Friedman and Nissenbaum, 1996). Throughout this work, the term fairness is used to describe an algorithm that is unbiased.

Why is algorithmic bias an important topic to investigate? The answer to this question is that dependence and reliance on automated data-driven decision-making systems should not lead to unfair access to opportunities.

The general belief that algorithms do not discriminate since they are built upon mathematical concepts and sequence of instructions and that they are incapable of possessing mental states has been disputed due to numerous cases of unwanted bias and discrimination reported over the years. In 2015, for instance, the Correctional Offender Management Profiling for Alternative Sanctions (COMAPS) risk assessment tool which is used in criminal justice was found to discriminate against African Americans by scoring them as twice more likely to recommit a crime (Chouldechova, 2016).

In 2019, Apple's new credit card sparked an investigation for gender bias when it was found to approve men for higher credit limits than women. These examples highlight the criticality of investigating the moral costs associated with adoption of predictive tools, particularly in high-stake settings, where decisions can have high repercussions for individuals.

In view of the information presented above, this research project seeks to investigate how fairness in algorithmic decision-making systems may be achieved. Intuitively, one might seek to look for an *all-encompassing* solution. However, as the scientific and social community have come to agree over the years, it is very challenging to design a system that is deemed fair across different social contexts (i.e., how one thinks about fairness in criminal justice might be very different than case of recruitment or credit assessment) (Koene et al., 2017). Beyond the application domain, the protected class needs to be specified as well since implementing fairness requires understanding the inherent disparities between the members of a protected class.

With this in mind, the scope of this thesis project is narrowed down to tackling algorithmic fairness issues with respect to *gender* in consumer *loan* approval systems. It is critical to highlight why lending and gender are vital specifics to consider:

Lending: The influence, scope, and reliance of the lending domain on supervised machine learning (ML) tools has noticeably increased over the past few years. One of the key advantages of utilizing such tools is predicting the creditworthiness of loan applicants using larger datasets and alternative data (Walker, 2019). Having access to loans can increase the quality of life of individuals as it assists them to purchase homes, obtain financial stability, and protect themselves from unforeseen circumstances.

Gender: Discrimination against women in finance has a long history. Even though European laws that mandate equality of opportunity have effectively curtailed direct and hostile discrimination against women in access to financial products, discrimination still prevails due to the disparities and barriers that women face to financial inclusion. If information in the datasets reflects those disparities between men and women, the decision outcomes of the ML models may become discriminatory. Even if datasets are free of biases, subjective decisions taken during the model development and deployment processes can lead to discrimination. Not only this is morally objectionable and unlawful, but also detrimental to the overall health of the economy as economic development and gender equality are positively correlated (Quinones, 2016).

The information presented above highlights the necessity of having mechanisms in place that can help individuals and financial institutions assess the fairness of their automated loan approval systems. Therefore, as its main objective, the research will investigate the following question:

> *How can automated loan approval processes be assessed for gender equality?*

In order to assess fairness of an automated decision-making model, a definition of fairness needs to be selected and tested for. This requires the stakeholders to define certain key fairness metrics and to have access to appropriate tools for bias detection in their datasets and models. To summarize, two key elements required to ensure fairness are:

1. Selection of appropriate fairness metrics.
2. Detection of bias in the datasets and predictive models.

In light of the information presented above, this work is divided into two stages in the following manner: In the first stage, it studies what constitutes fairness in automated decision-making processes and explains the rationale behind different notions of fairness. The second stage involves an exploratory case study in which a binary ML model is developed to classify the loan applicants in a sample dataset, commonly known as the German Credit Data. Both, the dataset and the model, are tested for gender related bias using the Artificial Intelligence Fairness 360 (AIF 360) toolkit provided by International Business Machines (IBM).

The remainder of this chapter consists of two sections: Section 2 provides the theoretical background and in Section 3, the elements of the research including the problem statement, research questions and research methodology are provided.

1.2 Background

1.2.1 Adoption of artificial intelligence by financial institutions

One of the key services that financial institutions provide is lending (financing). Lending is a massive business extending to almost every part of an economy. In 2018, the credit card debt alone reached $1.08 trillion dollars in the United States (Holmes, 2019). The financing domain is necessary for smooth functioning and the overall health of the economy. Due to the interdependencies between financial institutions, there exists a system-wide risk of failure if one of the member institutions fails. For this reason, financial institutions are heavily regulated and are required to limit and manage risk.

Considering some of the risk management systems already in place, there have been several global and regional regulatory bodies coordinating rules and oversight to help manage the systemic risk of the financing domain. To illustrate, the five major banks in Canada can provide credit cards to the so-called newcomers with no credit history, but must limit the maximum amount of credit granted to them -- limit-setting is in accordance with the policy set by the Department of Finance in Canada (Option conformateurs, 2014). However, perhaps the most noticeable nontraditional approach of financial institutions is the adoption of the state-of-art machine learning by these institutions that help with risk management in a more prominent way.

Conventionally, the risk assessment process starts by collecting material information about an applicant's credit history, payment habits, savings, and current debt. Since traditional credit-scoring models weigh a relatively limited set of data points, they may not adequately predict the creditworthiness of many "thin-file" consumers (Hurley et al., 2016). Ideally, to lower risk through thorough evaluation of the applicant's repayment ability, the lender may want to, after obtaining a

consent from a borrower, incorporate more consequential factors, such as economic conditions, psychometric data, and digital footprint into his or her assessment.

However, as most of the financing applications are conducted on a case-by-case basis by a human decision-maker, the process can be time-consuming. The data used in the traditional credit-scoring mechanism can also be inaccurate. According to a study done by the Federal Trade Commission (FTC) in the United States in 2013, twenty-six percent of the consumers surveyed had errors in their credit reports and these mistakes were material for thirteen percent of consumers, potentially resulting in higher denials and rates of interests (Hurley et al., 2016).

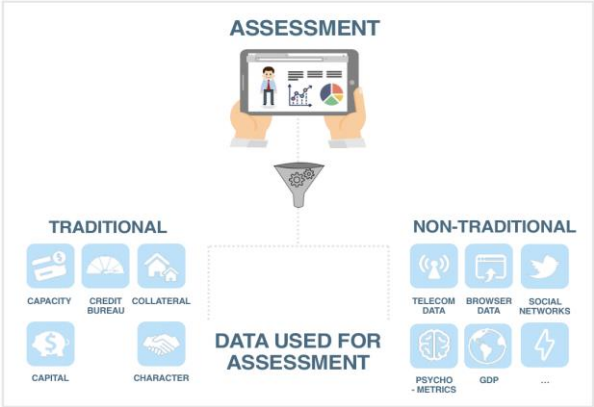


FIGURE 1: Traditional vs alternative data

To overcome the issues presented above, the financial institutions have been increasingly relying on ML tools to collect, verify, analyze, and process billions of data points, thus reaching a more accurate decision in a much shorter amount of time. ZestFinance and Lenddo are examples of fintech companies which use advanced machine learning tools to process vast number of features to determine an applicant's creditworthiness. Fair Isaac Corporation (FICO), a provider of one of the scoring models used by the majority of banks and credit unions, has recently announced the use of Lenddo's data analytics technology for assessing 350 million applicants with thin credit history in India (Srivats, 2016).

1.2.2 Concerns and legal framework

As discussed in the previous section, the lending market is an area of the economy where there is a large potential for efficiency gains from the use of machine learning. However, together with the widespread adoption of these new prediction technologies, some valid and indisputable concerns have been raised about whether their use would be beneficial for different groups. Before it is discussed how AI algorithms can actually lead to discrimination, it is important to elaborate on some of the current ethical and legal issues with deploying algorithmic decision-making models.

Firstly, financial institutions may cast doubt on whether they act with integrity, competence, diligence, and respect in an ethical manner towards their current and prospective clients if they ignore the social norm of equality.

Secondly, the financial institutions may find themselves in violation of laws that prevent creditors from discriminating against their applicants on the basis of gender, sex, color, religion, nationality, marital

status, and age. Banks, in particular, are expected to act with care when complying with such laws. For instance, after the introduction of Equal Credit Opportunity Act in the US in 1974 prohibiting sex-based classification, banks dramatically changed their policies towards women, many of which may have been based on outdated stereotypes about women's commitment to the labor market (Hurley et al., 2016).

In Europe, the Non-Discrimination Law (ECHR) is the main legal safeguard that protects people from AI-driven discrimination. As outlined in the article 14 of the European Convention on Human Rights (ECHR, 1953), this law states that:

> "*the enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, color, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status*."

ECHR classifies two categories of discrimination: direct and indirect. In the context of machine learning, direct and indirect discrimination are referred to as "*disparate treatment"* and "*disparate impact",* respectively (Borgesius, 2018). Under the acting laws, both types of discriminations are prohibited.

The European Court of Human Rights (ECtHR) broadly describes direct discrimination as follows: "there must be a difference in the treatment of persons in analogous, or relevantly similar situations, which is based on an identifiable characteristic" (ECtHR, 2016). The key idea here is that discrimination can only be classified as direct if it attests to have an identifiable discriminatory *intent*. For instance, an underwriter who makes his approval decisions based on gender of the borrower is said to imply direct discrimination.

On the contrary, a policy or practice is described as indirect discrimination if it is facially neutral and unintentional but results in a disproportionate adverse impact on a protected group (Corbett-Davies, Goel, 2018). To see how a lending policy can indirectly disadvantage women, consider the study done by the World Economic Forum (WEF):

In 2016, it was found that there was a 52% difference between the average salary of men and women in the Netherlands ($31,310 USD for women and $65,446 USD for men) (WEF, 2016). A hypothetical lending policy that only accepts applicants with income of *greater* than $55,000 USD in the Netherlands may have an adverse impact on women as it hinders their ability to access loans amid the differences between the average income earned by men and women in the country.

1.2.3 Classification models

In this section, a brief explanation of the fundamentals of classification is provided. This section also familiarizes the reader with the frequently used jargons which will be used throughout the report.

When categorical decisions such as approve/reject, or rich/poor/average is being made about individuals, a classification model can be used. Expressed in more technical terms, a *classifier* maps its data subjects into a *class* that is separated by a decision boundary from other classes. Since this work solely focuses on approve/reject decisions regarding loan applicants, a binary classifier will be used

which assigns the applicants into either the approved (positive) or rejected (negative) class. To illustrate how a classifier is built, the simple diagram below is provided:

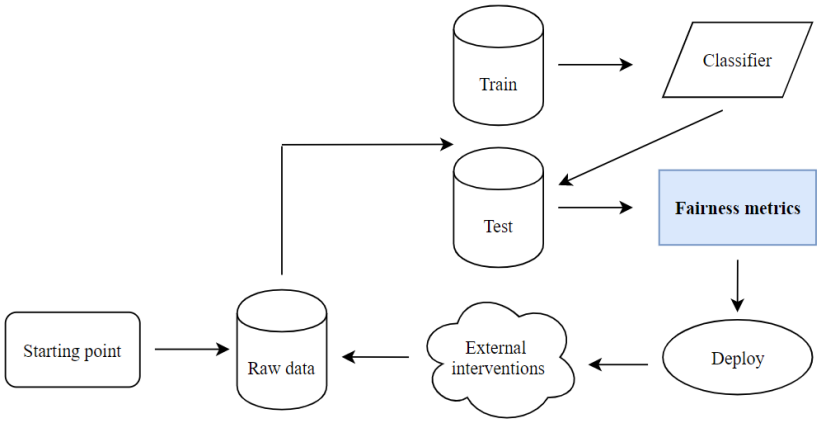


FIGURE 2: Machine learning classifier development pipeline

The classifier development process starts with the raw data. Once the data is processed and prepared, a cross-validation technique is applied in which the dataset is split into test and train partitions. Each partition contains *instances (applicants)* and for each instance, there are *features* and *labels* present in the dataset. Features may contain sensitive information such as race or gender in which case, they are referred to as *sensitive features* or *protected attributes*. *Target variables* (labels) are what the model predicts. As an example, features can be income and savings, and the true labels can be defined as risky/not risky. Once a chosen classification model is trained, it can be tested for accuracy using the test partition dataset.

On the figure, the point in which the classifier is tested for bias is also shown. As explained above, to test for a notion of fairness, a quantification of bias should be used which is commonly referred to as a *fairness metric*. According to the pipeline, the classifier is tested for bias, does this mean that the classifier discriminates?

The answer is that a classifier alone does not discriminate, however, the process in which it is developed can lead to bias (d'Alessandro et al., 2017). This is one of the reasons why bias is difficult to source: it can creep into the system at various stages. As part of the research, an extensive investigation on types of bias which can be encountered during the development stages of a classifier is conducted throughout the case study.

1.3 Research Description

1.3.1 Research problem and knowledge gap

In machine learning, fairness can be defined in numerous ways. Over 21 prominent definitions have been proposed over the years that can be used across different contexts (Narayanan, 2018, Verma and Rubin, 2018). These definitions are grouped into three categories as described below:

1. *Statistical measures* require equal treatment across groups in the protected class. As such, they are also called group fairness.
2. *Similarity-based* measures require similar individuals to be given the same decision outcome.
3. *Causal measures* require the features to be non-discriminatory and legitimate.

	Definition	Category
1	Disparate treatment	Statistical Measures
2	Group fairness or statistical parity	
3	Conditional statistical parity	
4	Predictive parity	
5	False positive error rate balance	
6	False negative error rate balance	
7	Equalized odds	
8	Conditional use accuracy equality	
9	Overall accuracy equality	
10	Treatment equality	
11	Test-fairness or calibration	
12	Well calibration	
13	Balance for positive class	
14	Balance for negative class	
15	Causal discrimination	Similarity -based measures
16	Fairness through unawareness	
17	Fairness through awareness	
18	Counterfactual fairness	Causal reasoning
19	No unresolved discrimination	
20	No proxy discrimination	
21	Fair inference	

TABLE 1: 21 prominent definitions of fairness. Adapted from Verma and Rubin (2018)

One of the challenges in using these notions of fairness is that there is no consensus and guideline on how they should be selected. For example, financial institutions and their applicants may have conflicting views on what metrics too choose; lenders may favor false positive errors rates (*to reduce risk of default to the financial institution)* while applicants may favor false negative error rates (*to reduce the risk of falsely getting rejected).* Another challenge in using these metrics is that some of them are incompatible and cannot be satisfied simultaneously, except for a very narrow range of circumstances and, therefore, some discrimination is *unavoidable*.

A closer look at the problem above reveals a more fundamental issue: these metrics measure fairness, an abstract concept rooted in the perception of individuals, has no formulation and cannot be concretely explained. This is perhaps the main reason why there is no consensus on which metric to choose in the first place. In response, one logical approach to tackling this issue might be to step away from these technical definitions and, as a preliminary step, conceptualize and ultimately formulate what it is gender equality in consumer lending. If we find a concrete way to define fairness, we can then more effectively analyze the suitability of these fairness metrics.

Another important issue to consider here is that there is a lack of knowledge on how to implement these metrics in practice. Often choosing an appropriate metric for a given dataset is chicken-and-egg problem: the types of bias present in the dataset determine which metric is appropriate, however, identifying which types of bias are present requires some way to measure bias (Hinnefeld et al., 2018). Beyond applying the metrics, there is lack of knowledge on how to appropriately interpret the results and make conclusive statements about fairness.

Apart from the above challenges, a fairness metric is only a quantitative measure of bias and it does not always indicate where the source of bias is. For instance, for a model which makes more inaccurate decisions for females than males, possible sources of bias could be that females are underrepresented in the dataset, or that the training dataset reflects some prejudice of the past decision makers.

In light of the information above, the problem this thesis investigates is as follows:

Without having an already agreed upon fairness metric(s) and knowledge of how to effectively apply them in practice, data-driven loan approval processes cannot be effectively assessed for gender equality. Furthermore, metrics are only a measure of bias and to carefully examine how decision-making models discriminates, possible root causes of gender-related bias need to be examined for the context of consumer lending.

1.3.2 Research questions

Given the aforementioned problem statement, the main objective of this research is investigating the following overarching question:

> *How can automated loan approval processes be assessed for gender equality?*

The thesis project investigates the main research question through answering the following two research questions:

RQ1. *What fairness metrics are suitable for assessing gender equality in consumer lending?*

- *Deliverables*: Conceptualize gender discrimination in finance. Construct a definition of fairness which can be used as a basis for evaluation of the proposed fairness metrics. Reveal the moral assumptions that are embedded in the metrics and provide intuition about how they can be viewed through the lens of the equality of opportunity.

- *Research strategy*: Literature review and synthesis.

RQ2. *How can the metrics be applied to observe gender bias in lending history data?*

- *Deliverables:* Identify how the choice of dataset used can affect the performance of fairness metrics. Explain where in the classifier development process the metrics can be used. Investigate how fairness toolkits can be applied in practice to detect bias. Provide intuition and guideline on how to interpret the results after the metrics are applied.

- *Research strategy:* Exploratory case studies.

The entire report is divided into chapters that follow the above three research questions. The diagram below can be used as a reading guide for the user.

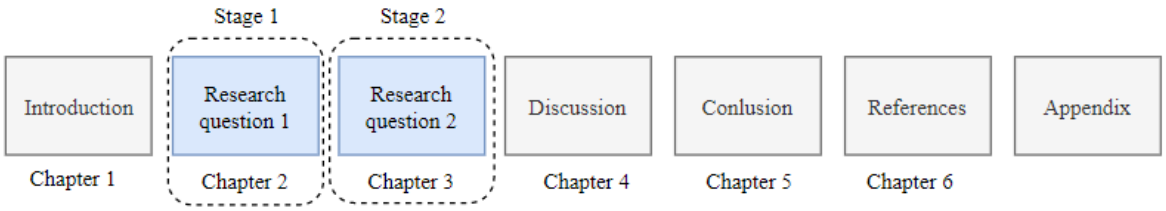


FIGURE 3: Thesis architecture (reading guide)

1.3.3 Research Methodology

The research is conducted in two stages that are interrelated as illustrated in Figure 2.

Stage 1

In Stage 1, the focus is on the fairness metrics proposed by Narayanan which are later used in Stage 2 to test the model for bias. To assess their suitability, the research takes a step away from these technical definitions and tries to conceptualize fairness by understanding the notion of *gender equality in consumer lending*. Ultimately, if we can find and construct a definition that incorporates the perspectives of key stakeholders while fitting well within the governing European laws, we can use it as a basis for evaluating the suitability of fairness metrics.

To conceptualize fairness, an extensive literature review encompassing philosophy, gender studies, economics, and computer science is conducted with the intention to form a cognitive process which can help with the interpretation of fairness metrics based on factors that lie outside of the domain of computer science. This requires understanding and synthesizing ideas that explain abstract concepts such as fairness, discrimination, and gender equality.

Stage 2

In the second stage of the research, a case study is designed in which a classifier is developed to predict the creditworthiness of applicants in a sample dataset. Both the dataset and the classifier are tested for bias. Furthermore, a detailed examination of sources of bias is conducted. Exploratory case study is chosen as the research method due to its suitability for scenarios where no single set of outcomes is expected. The elements of the case study are:

German Credit Data: This dataset is available on UCI Library. It contains information on 1000 loan applicants. Each applicant has 20 features (both numerical and categorical) and a true label representing whether he/she has actual good or bad credit. More information on the dataset is provided in chapter 3 of this work.

The AIF 360 toolkit: Over the past few years, some of the greatest contributions in addressing algorithmic bias have come from the development of open source libraries and toolkits that can be used by learners, practitioners and researchers. One of the latest additions is the AIF 360 toolkit developed

by IBM. AIF 360 is an extensive toolkit that offers bias detection, mitigation, and interpretation as an all-inclusive package. One of the main advantages of the toolkit is that its architecture has been constructed to conform to a standard paradigm used in data science, thereby further improving usability and integrability into common practices of machine learning (Bellamy et al., 2018).

Taxonomy of bias. A taxonomy containing an extensive set of practices that can lead to bias is used as a framework in the case study. The taxonomy also lists all the stages of the ML process where biases can be encountered.

The case study is conducted in sequential steps such that it resembles a simple process of classifier development which can be seen in practice. It starts with pre-processing the raw data into the desired format followed by training and testing of the model for accuracy. In every step of the way where a *subjective decision* is being made, a possible source of bias is discovered and demonstrated. Finally, fairness metrics, after being evaluated in Stage 1, are used to test the classifier for bias.

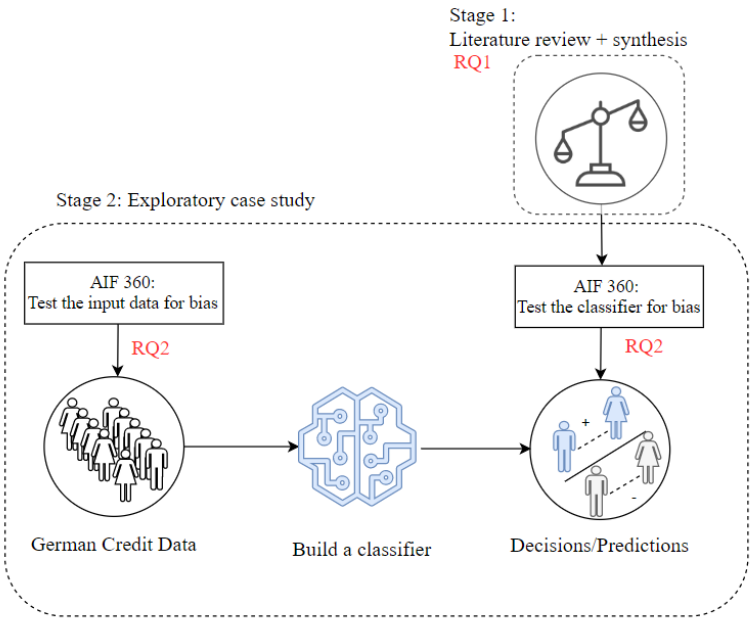


FIGURE 4: Research methodology paradigm

RQ1. What fairness metrics are suitable for assessing gender equality in consumer lending?

RQ2. How can the metrics be applied to observe gender bias in lending history data?

1.3.4 Relevance to the MoT program

The Management of Technology (MoT) program at Technical University of Delft explores how emerging technologies can be leveraged and managed to advance the processes in organizations. The program puts great emphasis on ethics and responsible innovation of technology. Therefore, the aim of this research, understanding, evaluation, and correction of an impact that artificial intelligence technology exerts on the credit assessment process, is aligned with the core theme of MoT. The research also encompasses some of the subjects taught in the program such as Research Methods, Inter- and Intra- organizational decision making, and Social and Scientific Values.

This page is intentionally left blank.

2

Conceptualizing fairness

in automated loan allocation systems

2.0 Introduction

Recent years have seen an increasing reliance on automated decision-making systems to make consequential decisions about individuals in high-impact areas such as in consumer lending, and criminal justice. People critically examine the fairness of the treatment they receive. It is therefore not surprising that concerns about whether these automated decisions are fair and unbiased have been raised as they may have long-lasting impact on society.

What, however, constitutes fairness? An essential step in addressing such concerns is formalizing what it means for a decision-making model to be 'fair'. To formalize fairness, the machine learning community has proposed 21 metrics that quantify unwanted bias in datasets and models. However, little information is provided about their suitability and their strength in measuring fairness across different settings under scrutiny. This comes with no surprise as the very concept of fairness extends beyond mathematical terms and is not a matter that provokes clear and shared understanding.

As views on the meaning of fairness are shaped around personal characteristics, power positions, individual experiences, and societal norms (Dobbie et al, 2010), it is critical to turn to the philosophical literature to a gain better understanding of the concept of fairness and to, ultimately, develop conceptual tools for idealizing a suitable and fair automated decision-making models in the context of consumer lending.

For this reason, this chapter of the thesis project is dedicated to investigating its first research question:

> *What fairness metrics are suitable for assessing gender equality in consumer lending?*

Given the specifics of the setting described here, a synthesis is provided as follows:

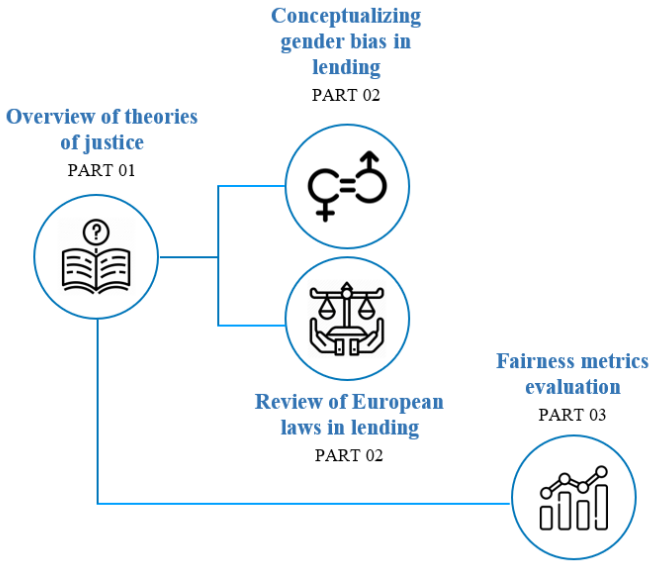


FIGURE 5: Chapter 2 architecture

As a first part of the synthesis, an overview of the philosophical theories of justice is conducted systematically to gain a better understanding of what constitutes fairness. The objective is to determine exactly what *good* is being distributed, what are its *guiding principles of distribution*, and finally, what are its fairest *distribution patterns*. In this part, some of the governing European laws surrounding consumer lending and algorithmic decision-making practices are also studied to examine which of the theories of justice are more suitable for conceptualizing fairness in practice.

Secondly, a literature review is conducted with the central goal to understand the position of female borrowers through conceptualizing the notion of *gender discrimination in consumer lending*. Over the past few decades, the social science community has contributed a great deal of work to measure and promote gender equality across various contexts. The benefit of reviewing social science literature on the topic lies in the fact that discrimination from the lens of public and, most importantly, victims of discrimination is primarily taken into consideration.

Finally, the last section examines various notions of fairness (Table-1) that have been proposed over the past few years. These notions can be grouped into three main categories: group-level measures, individual-level or similarity-based measures, and causal measures (Narayanan, 2017; Verma & Rubin, 2018). The ultimate goal of this section and this chapter is to reveal the moral assumptions that are rooted in the technical notions of fairness and to pave the ground for more in-depth analysis of the fairness metrics using the emerging theories in economics and in philosophy.

2.1 Overview of theories of justice

Above all, academics in social sciences and philosophy often describe fairness perception as multi-dimensional. It is defined in multitude ways, for instance, in terms of maximizing utility for the greater number of people in a society or considering different societal rules such as individual rights and freedoms (Binns, 2018; Mil, 2016; Rawls, 1999).

To determine how material and non-material goods should be distributed in a society, political philosophers for many years have proposed many different normative theories in attempt to understand what constitutes fairness and how to achieve it. In contexts such as consumer lending, understanding what constitutes fairness and how to measure it is still subject to a lot of debate with no clear definition.

To conceptualize fairness for the context at hand, this synthesis systematically investigates some of the most relevant and applicable theories of justice by answering the following three interrelated questions (Pereira et al., 2016):

- *What good is being distributed?*
- *What is the guiding principle of distribution?*
- *What is the fairest distribution pattern?*

2.1.1 What good is being distributed?

In the case of loan allocations, there are numerous ways in which the good can be thought of. For instance, some instinctive choices are the *principle amount* of the loan, or interest rates, or even time to maturity of the loan. In these examples, all goods are tangible and take a numerical form.

One can also define the good to be the opportunity or likelihood to access to loans for which, drawing numerical values becomes an obsolete task. However, when machine learning models are used to allocate loans to applicants, opportunity to access loans *can* be measured since these predictive models assign *probabilities* to instances in a dataset.

If the good is taken to be opportunity, from a legal point of view, fairness of an algorithm can then be broadly defined as providing *equality of opportunity to access loans*. While this broad definition is not a conclusion and rather a value-judgment, it serves great benefit to this work since only approve/reject decisions are being made about individuals, and not loan amount or interest rates.

Good = Opportunity
Descriptive choice? No, but it fits well within the governing laws.

Equality of opportunity is a widely supported ideal of fairness in philosophy. As opposed to equality of outcome, equality of opportunity contrasts morally acceptable and morally unacceptable inequality. The idea at the heart of equality of opportunity is to emphasize the important factors for which people can be held morally accountable and minimize the impact on circumstances or arbitrary factors (Arneson, 2015). Arbitrary circumstances are those factors for which individuals cannot be held accountable, while voluntary factors, or effort, are factors which justify the inequality in distribution of benefits or harms.

There are three prominent interpretations of equality of opportunity that can be applied to the algorithmic decision-making processes in the context of lending: libertarian, formal, and substantive equality of opportunity.

Libertarian equality of opportunity

Under the libertarian interpretation, it is morally acceptable for individuals to take whatever actions or make whatever decisions with what they own as long as they do not harm others in ways that violate their rights. Ways of harm are considered to be fraud, theft, breach of contract, extortion, and infliction of physical damage on persons and their properties (Arneson, 2015; Heidari et al., 2019). Outside of these violations of rights, outcomes are morally acceptable under libertarian equality of opportunity.

In the loan allocation scenario, operating under the libertarian ideal would imply that financial institutions have the freedom to implement any algorithm with the inclusion and exclusion of any individuals or any input features, including sensitive feature gender, as long as they do not breach data privacy rights. According to libertarian, the financial institution has the moral right to refuse loans to females even though this practice is nonetheless morally wrong.

The two theories of justice under which libertarian equality of opportunity can be conceptualized are utilitarianism and libertarianism (see Section 2.1.3).

Formal equality of opportunity

The formal interpretation of equality of opportunity brings the wrongness of direct discrimination based on irrelevant and arbitrary traits such as gender and race into light. In the loan allocation scenario, under formal equality of opportunity, the financial institution is required to allow anyone to apply for loans and is prohibited to refuse an applicant based on his or her gender. Formal equality of opportunity in algorithmic decision-making practices requires that sensitive features are not directly used in training algorithms. But do all individuals have a genuine opportunity to compete for the loans?

This is the weakness of formal equality of opportunity: it is merely formal, meaning that it does not account for policy or practices that have indirect impact on individuals due to the differences in their circumstances. Assume a bank assesses its applicants based on whether they have taken parental leave in the past or not. A decision based on parental leave may significantly impact women more than men as women tend to take more responsibility at home for providing care for children. This impact, however, is morally right under formal equality of opportunity.

Substantive equality of opportunity

Substantive ideal moves the starting point of the competition farther back and requires not only open competition (formal equality of opportunity) for accessing financial products, but also fair access to the necessary qualifications to be competitive. Substantive equality of opportunity holds, with respect to financial inclusion, if all candidates are eligible to compete for loans by being given a sufficient opportunity to develop the qualifications needed to be indeed successful.

Substantive theory implies that access to qualifications should not depend irrelevant factors and involuntary choices, such gender or social class. It is closely tied to the concept of indirect discrimination: A decision-making model which indirectly discriminates against people with a certain protected class membership (e.g. women or Hispanic) is often considered to be in violation with substantive equality of opportunity (Heidari, 2019). There are two widely adopted ideals that belong to substantive equality of opportunity: Rawlsian and luck egalitarian.

To better understand this concept in the context of lending, let's consider the following scenario. A financial institution favors applicants who have a mobile device registered under their name as

compared to applicants who do not because it may provide an insight about applicants' bill repayment habits.

The problem is that on average women seem to own fewer telephones which constitutes one of the (potential) involuntary gender-related constraints that undermines their access to financial products. In algorithmic decision-making practices, unfair competition is usually a result of using features that reflect inherent disparities between the two genders. Therefore, an essential step in adhering to substantive equality of opportunity is to investigate the relationships between the sensitive feature (gender) and other features in the dataset. Reverting to the example above, cell phone ownership serves as a proxy for gender.

What features, however, constitute applicant's fair access? The shortcoming of substantive equality is that it does not provide a clear formulation of differences, how they should be measured and accounted for. The concept of fair access or competition may be interpreted with wide range of meanings, it is difficult to measure, and thus its implementation raises various problems. Continuing with the example above, there is no mutual agreement in whether using the ownership of a phone as a feature for evaluation is *actually* indirectly discriminatory. Although statistical findings showing lower cellphone ownership may indicate some disadvantage amongst women, it cannot be reliably said what percentage of this disparity reflects involuntary choices. Section 2.2.2 of this work is dedicated to investigating such differences between men and women in consumer lending that may lead to discrimination against women.

The table below summarizes the key points discussed above to help evaluate the similarities and differences of the three ideals of equality of opportunity.

<table>
<tr><th></th><th>Libertarian</th><th>Formal</th><th colspan="2">Substantive</th></tr>
<tr><td>Description</td><td>People are at the liberty to do what they please with what they legitimately own, as long as they do not harm or infringe on other people's rights and belongings.</td><td>All individuals who possess the attributes relevant for selection must be included in the pool of eligible candidates.</td><td colspan="2">Candidates should obtain fair access to necessary qualifications for competition and the practice or policy should incorporate difference amongst groups. It has two refinements</td></tr>
<tr><td rowspan="2"></td><td rowspan="2" colspan="2"></td><td>Rawlsian</td><td>Luck egalitarian</td></tr>
<tr><td>Individuals with the same effort level and ambition should obtain the same prospects of an outcome regardless of their involuntary circumstances such as gender.</td><td>Narrower than the Rawlsian ideal as it requires efforts of individuals to be distinguished from brute luck and voluntary choices.</td></tr>
<tr><td>Restricts</td><td>Use of force, theft, fraud, or damage, infringement of rights</td><td>Direct discrimination</td><td colspan="2">Direct and indirect discrimination</td></tr>
<tr><td>Restricts (in ML)</td><td>Use of features that breach data privacy laws</td><td>Exclusion of unprivileged group</td><td colspan="2">Practices that apply to all candidates in the same way, but disadvantage a group with certain circumstances more than others</td></tr>
</table>

TABLE 2: Three prominent interpretations of equality of opportunity

2.1.2 What is the guiding principle of distribution?

The second question aims to identify the guiding principle for distribution of the chosen good. The political philosophy differentiates two guiding principles of fairness that are rooted in human

perception: *distributive justice and procedural justice* (Colquitt & Rodell, 2015; Blader & Tyler, 2003).

There are three primary reasons for adopting these two guiding principles: Firstly, they both specify what is meant fair/ equal opportunity. Secondly, they have been widely used for evaluating gender inequality. Finally, as it will be explained later, fairness metrics are measures of distributive and procedural justice.

Next, starting with the principle distributive justice, the two guiding principles are explained in greater detail.

2.1.2.1 Distributive Justice

> *"The economic, political, and social frameworks that each society has—its laws, institutions, policies, etc.—result in different distributions of benefits and burdens across members of the society. Arguments about which frameworks and/or resulting distributions are morally preferable constitute the topic of distributive justice. Principles of distributive justice are therefore best thought of as providing moral guidance for the political processes and structures that affect the distribution of benefits and burdens in societies, and any principles which do offer this kind of moral guidance on distribution, regardless of the terminology they employ, should be considered principles of distributive justice."(Lamont & Favor, 2017).*

Distributive justice refers to the fairness of the outcomes (ends; fair share) of decision making. Said otherwise, it translates to the degree to which one perceives the distributions of the goods to be fair. The outcomes may be material, such as salaries and promotions or loan approvals rates across different group memberships but may also consist of intangible outcomes such as praise. Often perception of fairness is determined by individuals' initial expectations of how the outcomes should be distributed, their knowledge of the situation, their prior experience with similar situations, or even external factors such as media. (Gilliland, 1993; Dobbie et al, 2010).

Examining distributive justice in many contexts (such as in machine learning where processes are hidden) is sometimes as the only way to detect or evaluate the existence of unjust treatment. As such, it has become widely applied in decision making models, particularly those concerned with making fair allocation of resources or removing bias from the classification process (Salles, 2017; Lee 2018). Practices that would strip machine learning algorithms of bias and discrimination has become extremely crucial to investigate due to emerging evidence that improperly formed machine learning process can lead to unfair decisions (Sweeney, 2013).

This pressure has led experts to provide various conceptualizations and definitions of fairness along with their corresponding quantifiable measures (metrics). For instance, definitions based on individual fairness allow machine learning algorithms to learn to categorize similar individuals, similarly, based on one's definition of similar individuals (Dwork et al., 2017). Group fairness offers statistical techniques that allow for treatment of protected groups similarly to the entire population (Corbett-Davies & Goel, 2018). Another research introduced validation technique that allows third parties to

investigate the machine learning process and evaluate whether the algorithms make decisions in accordance with individual or group principles of justice (Datta & Zick, 2016).

Group fairness and individual fairness metrics can be thought of as quantifiable measures of distributive justice. Group fairness metrics require the probability of receiving similar treatments (for males and females in the same class) to be the same (see Section 2.3.2.1). Individual fairness metrics require two male and female applicants that are identical/similar to receive similar treatment (see Section 2.3.3).

2.1.2.2 Procedural justice

> *"Outcomes need to arrive from somewhere. Procedural justice refers to the decision-making process or the set of policies that are used to make allocation decisions. The process should treat all parties consistently, be free from bias, use accurate information in rendering decisions, take into account the views of all (something like voice), be correctable in the event of an error, and remain consistent with prevailing ethical norms." (Cropanzano, & Molina, A, 2015).*

Procedural justice (process fairness) refers to the fairness of the decision-making processes (means; fair play) leading to particular outcomes. The inception of the study of procedural justice started with psychologists' realization that people's judgement of decision as fair was dependent not only the outcome itself, but also by the means in which the outcome was determined (Thibaut & Walker, 1975; Leventhal 1980).

To put it differently, people may identify identical outcomes as fair or unfair based on the procedure taken to arrive at the outcomes. Surprisingly, even negative outcomes can be viewed as fair provided that the corresponding decision-making process is justifiable (Thibaut & Walker, 1975). These findings on procedural justice have played an important role in social sciences, negotiations, and organizational justice where it has been proven that fair processes do, in fact, lead to higher acceptance of and adherence to negotiated arrangements as well as higher satisfaction with decision-making authorities (Lind et al., 1997; Pruit et al., 1993).

What criteria, however, constitute a fair process? Several studies mention the application of following procedural rules required to increase the perceived fairness of a decision-making process (Chan, 2011:

- Firstly, ensuring that processes utilize accurate and well-founded information which is also knowns as *accuracy*.

- Second is *absence of bias*. Decisions about individuals should be based on information that is free from personal biases, prejudice, or discrimination. For the context of lending, absence of bias could mean making decisions using non-discriminatory and relevant information that corresponds to creditworthiness and repayment ability.

- Third is *consistency* or consistently applying processes across individuals in analogous situations. Consistency includes equal, consistent application of rules across differing applicants, clients or prospect clients.

- Fourth comes *voice* or designing a process so that individuals may express their concerns like allowing appeals and grievances that could influence the final outcome. Voice is the opportunity to provide input or feedback in the decision-making process.

While all these four key elements influence how people perceive fairness of decision-making processes in consumer lending, moving on, this work solely focuses on absence of bias. That is, the term procedural justice here will represent only one of its dimensions. As it will be described in (see *Section 2.3.4), causal measures can be thought of as measures of procedural justice.*

Key takeaways so far:

In the case of automated decision making, equality of opportunity can be based on the two guiding principles of distributive and procedural justice. To visually express these two notions, the figure below is provided.

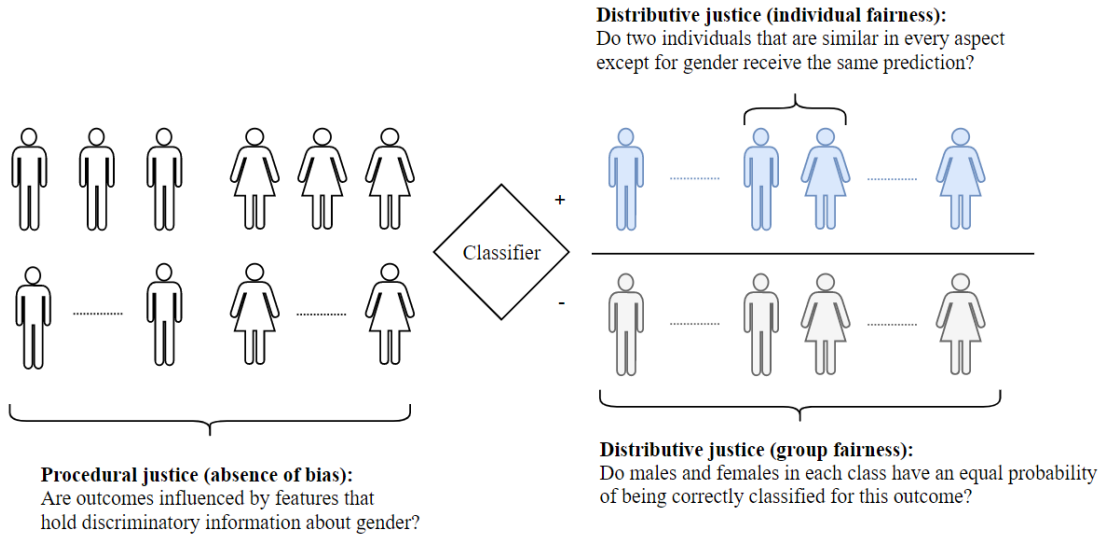


FIGURE 6: Possible interpretations of procedural and distributive justice in classification

Firstly, before any predictions or decision are made about the individuals, *process fairness* requires the information used for assessment (features selected) be free of bias. If features that disproportionately disadvantage a particular group based on their irrelevant circumstances are used in the data, algorithms can learn such disparities and enforce them when predicting other instances.

Fairness also depends on whether distribution of outcomes adhere to *distributive justice*. Fairness of outcome requires that individuals in the same true (or predicted) class be assigned the same likelihood (probability distributions, or opportunity) of being either correctly/incorrectly classified. This is not a straightforward concept to grasp on , as such, a great deal of this work is dedicated to explaining and providing examples that are tailored to context of consumer lending (see Section 2.3.3)

2.1.3 What is the fairest distribution pattern?

After having indicated the good and its guiding principles of distribution, the last and, perhaps, the most normative step taken to conceptualize fairness in lending requires determining the fairest *pattern* of distribution. To remain descriptive this work doesn't advocate for any particular solution.

Putting the previous two steps together should, however, help identify whether the good in question should be distributed, for example, based on talent and hard work of individuals (as dictated by meritocracy) or distributed equally among everyone (as dictated by needs principle).

To explain this stage in better detail, the following sections lay out what each normative theory has to say about the fair patterns of distribution.

Utilitarianism

Utilitarianism considers utility or human well-being to have the greatest importance and that the moral judgement of an action, procedure, or policy should be based solely on the consequences. In other words, utilitarianism holds that the fairest and most efficient alternative is the one that maximizes the utility or welfare for the greatest number of people. Due to this characteristic, it is often referred to as *strict consequentialism.* One of the major shortcomings of the utilitarian approach is that it puts emphasis on maximizing utility without considering the legitimacy of the process in which the outcomes come about. Violation of human rights and disregarding liberty of others, for instance, under this moral theory, is justified as long as the outcome benefits aggregate welfare (Pereira et al., 2106).

Without diving into the philosophical debates on strengths and weaknesses of utilitarianism, this work presents to main reasons why this moral theory might not be an appropriate approach for assessing algorithmic fairness in the context of lending. Firstly, the governing laws that apply to consumer lending (see Article 21 of the Charter of Fundamental rights of European Union) require that processes in which decisions are made be nondiscriminatory and to not violate individual human rights, irrespective of whether the decisions are for the greater good of the society. Secondly, even in the absence of laws that prevent the override of individuals 'rights, it is hard to technically measure the total utility and to make inference about whether some practices truly promote or maximize the overall welfare.

Libertarianism

This moral theory states that all individuals are at the liberty to do what they please with what they rightfully own, providing that they do not commit fraud, theft, or damage other individuals' belongings. Accordingly, libertarians claim that free markets are best means of maximizing social wealth. Furthermore, they see governmental intervention in markets, such taxation or subsidization policies, as mechanisms that distort market functioning.

In some ways, libertarianism is similar to utilitarianism in that it gives liberty to individuals and institutions to take any actions they want and therefore puts little emphasis on the legitimacy of the processes. However, the main difference between the two is that utilitarianism gives priority to an aggregate well-being over individual rights while libertarianism gives priority to individuals' liberties, even if they come at the expense of human welfare (Hausman & McPherson, 2006).

One of the main criticisms of libertarianism is that it fails to realize that self-regulated markets cannot provide sufficient solutions to unfair outcomes emerging from collective action. As such, they are accused of ignoring market failures that may be result of large power imbalance.

Another criticism of this moral theory is that it fails to address direct and indirect discrimination. For instance, a bank under libertarianism may refuse to allow minorities compete for its financial products, and thus, refuse to provide equality of opportunity to everyone. Similar to utilitarianism, it can be observed that the right to nondiscrimination which is outlined in Article 21 of the Charter of Fundamental rights of European Union, in Article 14 of the European Convention on Human Rights, and in Articles 18–25 of the Treaty on the Functioning of the European Union does not give the liberty to the financial institutions to deny applicants based on their gender (Goodman et al., 2017). Therefore,

it can be concluded that the libertarian ideal alone does not guarantee equality of opportunity to the loan applicants.

Rawlsian justice

In his very influential theory of fairness presented in 20th century, philosopher John Rawls equated fairness and justice, theorizing that fairness is "a demand for impartiality" (Rawls, 1999). While justice is, in a broader sense, often regarded as transcendental, justice as fairness is more context-bound. Fairness ought to be typically implemented through negotiating something agreeable and establishment of procedures that resemble the rules of the game with the ultimate goal that individuals receive their "fair share" in exchange for their efforts, thus adhering to a system of "fair play" (Rescher, 2002).

Rawls' theory contains two central principles organized by priority. The first one has absolute priority and appertains specifically to fundamental rights and liberties. The principle argues that the rules which define individuals' fundamental rights and liberties should apply equally to all and that individuals should have as much freedom as possible provided that it does not infringe others' freedom.

The second principle revolves around the distribution of primary goods, which are various social positions and all-purpose means required to allow citizens to pursue their life goals (anything it might be); they contain, in broad categories, wealth and income, opportunities, powers and privileges of authority, and the social basis of self-respect (Pereira, 2016). It also attests that social and economic inequalities can only be justified if they simultaneously arise from a scenario of fair equality of opportunity and work to the benefit of the least advantaged members of society.

According to Rawls, individuals with the same talent, ability, and willingness must have the same prospect of obtaining desirable social positions and the prospect should be irrelevant to their circumstances and arbitrary factors (Rawls, 1999). Rawlsian equality of opportunity assumes that effort is interpersonally comparable and, in particular, assumes that level effort is not affected by circumstance and by policy implemented. It then requires that for individuals with similar effort level, the distribution of utility is the same, regardless of circumstance.

In explaining equality of opportunity, Rawls recognizes that some degree of inequality is inevitable and that true equality of opportunity is unattainable because individuals' innate or educated skills, freedom of choice, or even effort may not be fully isolated from their social conditions. It is unjust for individuals to be punished or rewarded for arbitrary circumstances such as innate compacities, nor the original position in society in which they were born as having so does not represent a voluntary choice.

Luck egalitarianism

Unlike Rawlsian equality of opportunity, luck egalitarian equality of opportunity offers a relative view of effort and allows for the possibility of circumstance and implemented policy in acting the distribution of effort. This ideal requires that in comparing the individuals' efforts with different circumstances, we should somehow adjust for the fact that those efforts are drawn from the distribution of efforts that are fundamentally different. Luck egalitarians propose that measuring a person's effort level by his rank in the effort distribution of his type, rather than by the absolute level of effort he exerts.

Luck egalitarian equality of opportunity requires that people sitting at the same rank of the effort distribution for their corresponding type, all have the same distribution of utility, regardless of circumstances. Finally, affirmative action usually falls under the luck egalitarian substantive category with goal to help disadvantaged group return to a fair starting point after a long period of discrimination. This is done by, for instance, involving government action or transferring resources from advantaged to disadvantaged groups (Roemer, 2002).

One of the criticisms of this theory is that it lacks cogency on dichotomy it attempts to differentiate: luck and choice. (Kibe, 2011). This makes it challenging for luck egalitarianism to properly examine the inequalities ingrained in social relations, like involuntary associations, in which voluntariness and contingency are interwoven.

To help understand the difference between luck egalitarian and Rawlsian equality of opportunity, consider a situation in which four candidates wish to apply for mortgages. For simplicity, assume that each candidate is applying for an equal amount of loan and that they are similar in any determining factor except for income: income of person A (female) = $50,000, income of person B (female) = $45,000, income of person C (male) = $50,000, income of person D (male) = $70,000. In this scenario, the Rawlsian ideal would require individuals A and C to receive the same prospects of receiving the loan, since they have the same income or effort level, irrespective of their gender.

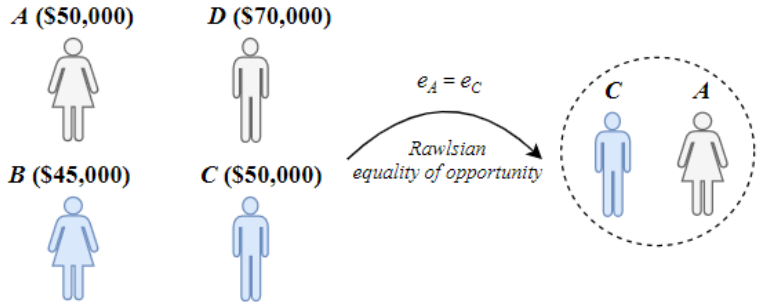


FIGURE 7: Substantive equality of opportunity: Rawlsian ideal.

On the other hand, luck egalitarian goes a step further and equates prospects based on rank and luck. This is done by first ranking individuals within their gender group, then equalizing prospects based on ranks: within females, A is ranked 1st and B is 2nd, while within males, D is ranked 1st and C is 2nd. Consequently, the individuals A and D should receive the same prospects of receiving the loan. As it can be seen in this scenario, the luck egalitarian assigns person C into a less desirable position even though he has the same effort-level as individual A.

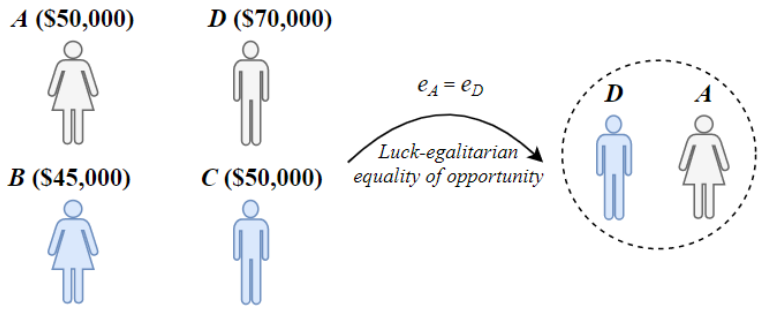


FIGURE 8: Substantive equality of opportunity: Luck-egalitarian ideal.

Following this example, it can be observed that when luck egalitarian requires 'neutralizing luck', it is sometimes no different that requesting affirmative action in favor of one group over another (Arneson, 2015).

Intuitionism

Intuitionism does not constitute a unified theory of justice, but instead reflects the viewpoints shared by scholars for whom moral premises are obvious and basic moral knowledge is intuitive (Rawls, 1999). Intuitionists hold the view that moral dilemmas in real-life contexts are so complicated and eclectic that they can solely be answered by a pluralistic definition of justice (Pereira, 2016). As opposed to utilitarianism and libertarianism, which present general theories about justice, intuitionism calls for a more context-dependent and pluralistic treatment. The judgement on what is the proper thing to do depends on the characteristics of every moral premise and may call for the consideration of various moral principles, such as formal equality, fundamental needs, rights, compensation, non-discrimination, or process fairness.

Intuitionism, from the view of universalists, is unsatisfying as it presents a series of arbitrary moral principles not defined by any coherent logical argument (Rawls, 1999). In practice, intuitionism is of little benefit to this work because it is unclear whether every principle would be the right one for financial institutions, or how one should proceed to reconcile or prioritize conflicting moral principles (Rawls, 1999).

Summary of the key points mentioned above are outlined in the table below.

Theories of Justice	**Description**	**Distribution of**	**Guiding principle**	**Distribution pattern (outcomes)**
Utilitarianism	The most ethical choice is the one that will produce the greatest good for the greatest number.	Welfare, well-being, utility	Distributive	One that maximizes aggregate welfare (Greatest good for the greatest number)
Libertarianism	Recognizes that all individuals equally share some fundamental right and the freedom to choose how to lead one's life according to one's values and goals.	Basic rights and liberties	Self-ownership	Absolute equality
Rawls' Egalitarianism	Those who have the same level of ability and are equally willing to use them must have the same prospect of obtaining desirable social positions.	• Basic liberties • Opportunity • Primary goods such as income, wealth, etc.	•Distributive (Deontological justification) • Equality of opportunity as pure procedural justice • Distributive (Difference principle)	• Equal distribution • Equal distribution • Maximin criteria

Luck egalitarianism	Demands that variations in how well-off people are should be wholly determined by the responsible choices people make and not to differences in their unchosen circumstances.	• Opportunity • Primary goods such as income, wealth, etc.	•Distributive (Deontological justification) • Equality of opportunity as pure procedural justice	• Equal distribution • Equal distribution only after removing factors that constitute luck
Intuitionism	Argues for context-dependent and pluralistic approach.	Different resources such as food, financial assets, health, education, etc.	Context-dependent: • Distributive • Procedural	No unified distribution pattern: (Equity, equality, power, etc.)

TABLE 3: Overview of justice theories: each theory is explained via three factors. 1) what good is being distributed. 2) Through which guiding principle are goods being distributed. 3) What is the fairest distribution pattern. Source: adapted from Pereira et al. (2016).

Key takeaways so far:

It is useful to summarize the key and concluding ideas from the extensive overview provided above. In conceptualizing fairness, a good that is to be distributed should be defined. For the case of automated loan allocations, a good can be thought of as opportunity (likelihood) to access loans. In assessing fairness, two guiding principles are prevalent here: fairness of the process in which applicants compete for the loans (procedural justice), and fairness of distribution of the probabilities which are assigned to applicants (distributive justice). Finally, in conceptualizing fairness, this work does not advocate for any particular distribution pattern (needs principle, principle of equity etc.) which can justify unequal distribution of opportunity amongst the loan applicants.

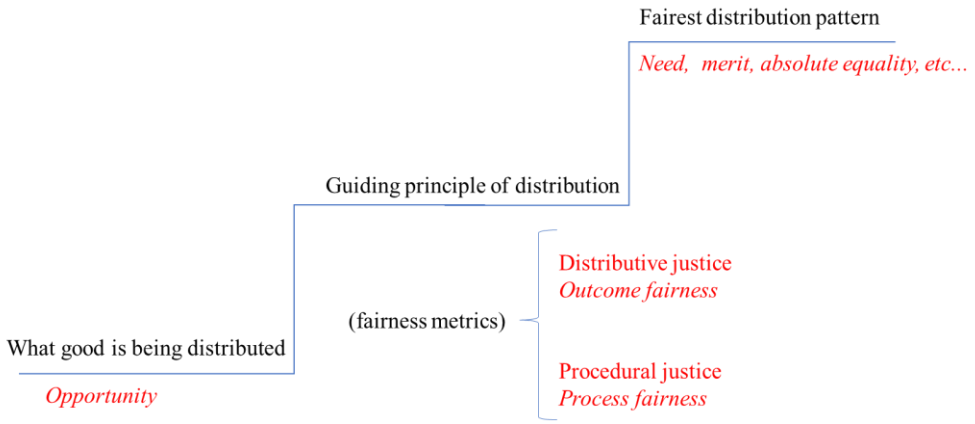


FIGURE 9: Conceptualizing fairness in three steps: First, the good that is being distributed is explained. Secondly, the appropriate guiding principle of distribution can be chosen. Finally, the fairest distribution pattern as dictated by its prevailing normative theory of justice can be used as guideline to evaluate fairness.

2.1.4 European laws (bottom line)

As mentioned earlier, European policies surrounding gender discrimination as well as non-discrimination laws on artificial-intelligence decision making in consumer lending (Article 21 of

Charter of Fundamental Rights of European Union) require that processes in which decisions are made be nondiscriminatory and to not violate individual human rights. It is important to note that Paragraph 71 and Article 22, Paragraph 4 specifically address discrimination from profiling which utilizes sensitive data. In unpacking this mandate, two potential interpretations should be distinguished:

The first minimal interpretation is that this requirement only pertains to cases where an algorithm is making direct use of data that is explicitly sensitive. This would include, given the case at hand, a feature column that represents gender. Therefore, under this interpretation, if a loan allocation model is not trained explicitly on the sensitive feature (gender), it meets the criteria for formal equality of opportunity. As described earlier, formal equality of opportunity only restricts direct discrimination and it is blind to practices that can have adverse impact on a particular group. However, as European laws prohibit both direct and indirect discrimination, ensuring formal equality of opportunity alone would not suffice as a nondiscriminatory practice (Goodman et al., 2017).

The second maximal interpretation takes a broader view of sensitive data, including not only the variables which are explicitly named, specifically adhering to formal equality of opportunity, but also any variables which they are correlated with. As stated in Article 9, this would require the decision-making model to ensure that algorithms are not fed with datasets containing features that are correlated with the "special categories of personal data" (Goodman et al., 2017). While the equivocal nature of Article 9 makes it difficult to make concrete conclusion about what features to use, one possible interpretation is that the machine learning algorithm should use variables that represent legitimate factors which, firstly, merely reflect *default risk or credit worthiness* and, secondly, can be traced back to individual's *voluntary choices*.

Under this second interpretation, the use of features correlated to sensitive attributes is deemed justifiable and fair. Substantive equality of opportunity, unlike the libertarian and formal explanations, highlights individuals' efforts as the basis for justified and morally acceptable inequality, seeking to eliminate the impact of irrelevant circumstances on outcomes. What remains unclear, however, is how to collectively identify what constitutes the three factors: effort, arbitrary circumstances, and luck.

The bottom line:

Substantive equality of opportunity dictates fair competition and elimination of *direct* and *indirect* discrimination in the selection process and in distribution of outcome, as such, it is closely tied to both procedural fairness and outcome fairness (Blond & Milbank, 2010; Stancil 2016). As it was already mentioned, equality of opportunity is a widely supported ideal of fairness amongst political philosophers and economists and it is the more emphasized way for dealing with unfairness for people with different race or gender (Loi et al., 2018). Moving on, this work gives particular attention to the *Rawlsian* and *luck egalitarian* ideals since these moral theories stress the importance of avoiding indirect discrimination in achieving fair decisions. As it will be explained later, many of the fairness metrics serve as a measure of equality of opportunity; they require the decision-making model to equate prospects of harm/benefits amongst individuals, regardless of their irrelevant characteristics such race or sex. *To summarize, fairness will be defined as*

Fairness:
Individuals who have the same effort-based utility (what justifies inequality), should be given equal prospect (likelihood) of receiving the same treatment, regardless of their irrelevant circumstances such as gender or race.

2.1.5 Normative thoughts

In the setting of automated allocation of loans, the decision outcomes primarily affect the distribution of scarce economic resources. When examining the fairness of distributions of these financial products, perhaps the most putative principle belonging to distributive justice is the principle of *equity* (the principle of meritocracy) (Wright & Boese, 2015). Therefore, it is the utmost belief of the writer that in consumer lending, what justifies fair inequality is *equity* as it achieves greater fairness of treatment and outcome.

Equity, paradoxically, calls for just, but unequal distribution of financial resources. Guided by meritocratic principles, it calls for equalizing opportunities instead of outcomes and defines justice as distribution of resources to the extent of an individual's *relative contribution and efforts*. According to the broad meritocracy ideal, if there is to be inequality in the rewards and remuneration and status dispensed by social arrangements, the fulfillment of equality of opportunity is necessary in order to ensure that individuals genuinely get what they deserve. Some theorists believed that in capitalist and most modern cultures, equity is understood to be the prevailing principle of justice (Walster et al., 1973).

2.2 Gender discrimination in consumer lending

While desirable, it is challenging to uniquely construct the notion of fairness in the context of lending as the ways of thinking of the stakeholders about fairness may have conflicting implications. For instance, a borrower may deem a lending experience just if he or she received the amount requested while overlooking the legitimacy of the process. Or even among borrowers themselves, one might find differences in what treatment they consider fair. For example, in the case of gender discrimination, a male borrower may only care about the outcome while a female borrower may perceive the process unjust, regardless of the outcome, if she felt that she was discriminated against during the decision-making process.

Consider another scenario in which a loan officer granting a loan to a male borrower instead of a pregnant woman that are similar in all material factors other than pregnancy. The lender may consider his decision procedurally fair (or in this case non-discriminatory) if he/ she believes that temporary discontinuation in employment due to maternity imposes great risk to the repayment capability of the borrower. In this example, the loan officer is considering continuation in employment to be a legitimate business necessity. Would the rejected female agree with him? A rejected female borrower may argue that the decision is unjust and discriminatory as the discontinuity in work due to maternity leave is involuntary and that it should not be accounted for as an evaluation criterion for credit worthiness. What about the male borrower who received the loan? A male borrower may very well be indifferent about whether the assessment criteria are discriminatory against him as historically men have not been victims of discrimination in the financial domain as compared to women.

These conflicting views indicate that achieving a fair decision-making system in consumer lending requires careful stakeholder analysis, study of value tensions, and compromise between actors. More importantly, this work emphasizes, as an essential step, the importance of identifying and listing factors that may lead to unequal treatment (with leaving out the argument on whether they are discriminatory or non-discriminatory in the case of loan allocations) of *women* in the context of lending. Knowing these factors can be especially useful for:

- Selecting features that have low dependencies to protected attribute gender.
- Assessing whether decisions adhere to distributive and procedural justice under Rawlsian equality of opportunity.

The next two sections of this chapter investigate what factors may lead to unequal treatment of women in consumer lending.

2.2.1 History of discrimination

To understand the factors that may shape perceptions of fairness in female borrowers, it is vital to look into the history and carefully examine how women were treated when applying for financial product over the years. Without a doubt, women have often been, and still are, victims of discrimination in various aspects of life. Great body of literature in social sciences, especially in gender studies, has described that discrimination stems from historically stereotyped social contracts and gender roles that were shaped in the private sphere (realm) of life (Denmark & Paludi, 2008; Ridgeway 2011; Lomazzi et al., 2018). Traditionally, the gender contract charged men with work outside the house to bring wealth and economic stability to the house, while women were in charge of domestic affairs, most importantly, of raising children. This traditional gender contract together with contemporary legal limitations has inevitably led to different treatment of groups in the labour market in the form of, for instance, wage (income disparity), employment (i.e., labour participation ratio), and segregation (i.e., degree of disparity in specific occupations and industries) differentials (Kulik et al., 1996, Malhotra & Schuler, 2005; Lomazzi et al., 2018).

Even today in most countries, women appear to have a lesser share of benefits and a greater share of burdens than men. Even in countries where women now enjoy the same legal rights as men, there are large disparities among these two genders in regard to wealth, income, living below the poverty line, and the participation in attractive positions such as high-income/high-status jobs, political functions - all of which are later reflected in women's capacity to apply for financial products. Such disparities seem to be especially hurtful and called upon in societies where women have started to even outperform men in work and academia (UN, 2010; UNDP, 2014). What are the underlying factors of the disparities between the two genders that seemingly result in unequal treatment of women in finance?

Unequal treatment of women in finance can be either direct or indirect. Direct discrimination, or 'disparate treatment', stems from prejudice or the belief that women are inferior to men in some aspects of life. Cases of direct discrimination had been evident the most during 60s and 70s, before anti-discrimination laws went into effect in Europe and in the United States. Some scholars refer to the ideology of direct, unequal treatment against women as hostile sexism (Denmark & Paludi, 2008). Hostile sexism resembles other forms of prejudice, typically directed towards groups who are seen as a threat to the dominant gender's status and power. Prejudice is also considered to be the starting point of the vicious cycle of discrimination, victimizing women against stereotype threats (Romei & Ruggieri, 2013).

Various cases of prejudice and direct discrimination have been gathered and documented throughout the past few decades. For example, a case of direct discrimination was the requirement to cosign loans by men or discounting a woman's income or by requiring them to provide baby letters; an affidavit that woman would not bear a child over the course of the loan (Bowdish, 2010). These types of discrimination that women suffered from those creditors undermined their ability to participate in credit

market (Bowdish, 2010). Another form of direct discrimination that has happened against women is called rational discrimination.

Rational racism, albeit illegal, stems from rational thinking. A form of rational racism is statistical discrimination, occurring when the lack of knowledge about the skills of an individual is compensated by a prior knowledge of the average performances of the group or category the individual belongs to (Romei, Ruggieri, 2013). Some cases of rational racism have been documented throughout the 1970s: Mortgage lending traditionally goes through three consecutive approval stages starting with the broker, then loan officers and bank's lending institution's loan committee. In many cases, the broker would not refer the female applicant to the next stage of the approval process, or the loan officer wouldn't pass her on to the loan committee even though they felt that she was a creditworthy borrower (Kendig, 1973). Rational racism and prejudice are more tied to direct forms of discrimination (Yamagishi et al., 1999; Barocas, Selbst, 2016). Initially, many people believed that if direct discrimination was legally prohibited, gender inequalities would entirely wither away, but clearly the currently implemented legal frameworks have not been successful in achieving this result. Lacking a holistic solution, the not so obvious indirect discrimination has entered into the picture (Gheaus, 2016).

Nowadays, although on occasion direct discrimination against women is still practiced openly, especially in less developed countries, it has become increasingly socially and lawfully undesirable to do so. Therefore, such discrimination as exists today is more likely to take on the more subtle and complex form of 'indirect discrimination'.

Indirect (implicit) discrimination, or 'disparate impact', is implementing a policy or procedure that seemingly treats everyone equally, but in practice disadvantages a protected group in comparison to the other. This usually occurs not because of malevolent decisions, but due to the lack of awareness on the effects of a decision (Romei & Ruggieri, 2013). For instance, a bank's policy to not offer loans to nail salons could be considered indirectly discriminatory as these businesses are disproportionately owned by women (Hertz, 2011). This would be, of course, unless the management can prove that not offering loans to nail salons constitutes a business necessity. It is important to note, however, that such unequal treatment is typically unintended and represents a case of indifference, incorrect design of procedures or practices, lack of proper stakeholder analysis, and understanding of the decision outcomes (Bertrand et al., 2005; Greenwald & Krieger, 2006; Kang & Banaji, 2010). Nevertheless, such treatment is inherently unfair or, as James W. Nickel writes, "indirect discrimination is morally wrong because its implication that one group is more worthy than another is insulting to its victims, because it harms its victims by reducing their self-esteem and opportunities" (Nickel, 2000).

While direct and hostile forms of discrimination against women still persist, they are less frequent than indirect discrimination in societies that have effectively adopted the anti-discrimination laws. Moving forward, indirect discrimination will become the main focus of this work.

Looking at the history, it can be inferred that directly discriminatory behavior towards women in consumer credit has become more easily identifiable and thus minimized due to new legislative measures and societal norms in most of the developed countries. Indirectly discriminative behavior due to its inadvertent nature and difficult claim process still remains a large hidden barrier to fairness which has to be addressed. This is why we have to look at the theories of justice to understand how this treatment can creep into our context at hand. In the next section, the selected two guiding principles and their potential applications in consumer credit are discussed.

2.2.2 Gender inequality and principles of justice

Without advocating for one guiding principle of justice over another, this section examines how each may be applied in the context of loan allocations. It is important to mention that the two principles are not independent and cannot be taken as orthogonal (i.e., one or the other) and in fact, fairness heuristic theory which is concerned with the relationship between perceived fairness of outcomes and procedures, proposes that people initially form a judgement on whether they have received a fair treatment or not with little distinction between procedures and outcomes (Van den Bos et al., 2001). The distinction comes to light when looking at what information people use to form judgments about the fairness of the treatment they receive from authorities.

To be able to reasonably judge whether an *outcome* is just, people need to compare their outcome of the decision they received with others. This emphasizes that distributive justice theories (e.g., equity or equality theory) rely on social comparison information in the process of evaluating the fairness of outcomes (Messick & Sentis, 1983). In loan approval processes that are heavily regulated by privacy laws, the social comparison information is not available, and as such, sound outcome judgement may not be possible. If information for comparing outcomes is not is available, how do applicants form judgement?

The answer is that in most traditional (i.e., non-algorithmic decision-making processes) information about the *procedure* is available. In other words, in situations where a person only knows his or her own outcome (and is not informed about the outcome of another person), the person will react more positively toward his or her outcome following a fair procedure than following an unfair procedure. This is commonly known as the fair process effect (Van den Bos et al., 2001).

On the other hand, when a person does have information about the outcome of a comparable other person, he or she will use this social comparison information to assess how to react to his or her outcome. Therefore, we expected less strong fair process effects in situations where a person does know what the other referent receives. An evident case of the former idea happened after a borrower condemned Apple Card for providing him a credit limit that is 20 times higher than his wife, even though the couple file joint tax returns and his wife has a higher credit score. In this case, two people that consider themselves 'creditworthy similar' have received different outcome without having any knowledge about the process with which the decision came about.

Let's suppose, for the sake of argument, that the loan process was not so restrictive and hence the information about the outcome and characteristics of referents was available so that reasonable inference about the fairness of the outcome could be made. Then the question becomes to what extent does the disproportionate distribution of outcome among groups can be assessed to be unjust?

According to John Rawls, in general, the key to achieving outcome fairness, or distributive justice, would be setting up societal rules that reflect the fair sharing of the burdens and benefits of social cooperation (Rawls, 1999). In other words, it signifies fair division of primary goods: wealth, income, and opportunities to obtain desirable social positions. Outcome comparison can be made amongst individuals or groups. It is much easier and more reasonable to infer whether a decision is just if comparison is done among two individuals. For example, two individuals that are deemed similar in creditworthy characteristic besides gender, must receive the same outcome, otherwise the decision-making mechanism is discriminating based on gender. On the other hand, uneven distribution of outcome amongst group is much more difficult to make inferences about the fairness of the decision-

making process. Said differently, disproportionate reward for equal contributions (inequalities of outcome) between women and men could then be *suggestive* of gender inequality, especially if gender ends up being a variable negatively influencing women's access to obtain attractive compensation or secure desirable social positions.

A potential source of easily detectable gender discrimination is, for example, a policy that outright discriminates against women. As described in the previous section, historically, this was the case in many societies. Nowadays, however, at least in developed countries, laws do not formally limit women's access to economic, political or social opportunities (Gheaus, 2016). Explanations, therefore, must be sought elsewhere, in the content and working of informal gender norms. Even legislation that does not openly participate in direct discrimination can only justify to a small extent the discord in economic, social, and political outcomes between the two genders. Many of the current obstacles faced by women still arise directly from policies or practices (legal-domain), which, while itself gender-neutral, affects women and men differently partly because of discriminatory treatment through biases faced in the other domains of life: demand-side (women), supply-side (financial institutions).

As another theory seems to confirm Rawls' idea of organizing society as a 'fair system of cooperation' (based on rationality, reciprocity, impartiality and mutual advantage), it also stresses the concept of an authentic equality of chances while taking into account the impact of *involuntary differences in social background.* When accessing financial goods and services, women and men can face similar constraints (Vossenberg et al., 2018). Many people lack affordable and accessible financial products and services due to barriers like erroneous government regulations with which the sector needs to comply and which underpin the cost structures of banks. But women, because of their gender and the assigned rules, behaviors, activities, and roles that apply to them in the home, in the law, the community, and the marketplace, can experience these constraints differently and to a greater degree (Vossenberg et al., 2018).

Demand Side Barriers	**Supply Side Barriers**	**Legal & Regulatory Barriers**
• Lack of bargaining power within the household • Concentration in lower-paying economic activities • Competing demands on women's time related to unpaid domestic work • Lack of assets for collateral • Lack of formal identification • Reduced mobility due to time constraints or social norms • Lower rates of cell phone ownership among women, needed to access many digital products	• Inappropriate product offerings • practices for product design and marketing • Inappropriate distribution channels	• Account opening requirements that disadvantage women • Barriers to obtaining formal identification • Legal barriers to owning and inheriting property and other collateral • Lack of gender-inclusive credit reporting systems

TABLE 4: The table above summarizes some of the potential involuntary gender-related demand and supply-side constraints women may experience when accessing financial services and products (Holloway, 2017).

As it can be seen from the table below, on the demand side, for example, inequalities in land-ownership regulations can limit women's options to present collateral needed for credit. Or, on the supply side, adverse societal norms about women's ability and right to handle finances, can limit bank's marketing

and outreach strategies to women. As indicated in the table above, women as a group experience more disadvantage compared to men in finance, therefore it would not be surprising to see that the distribution of financial assets such as loans would be skewed towards men. A practice, however, would be considered discriminatory if it increases the inequality even further as opposed to a scenario without such policy. This, however, would be very difficult to evaluate in real life.

Is this work advocating for procedural justice? No… but it's important to highlight some of the difficulties of using distributive justice as a guiding principle when it comes to gender equality.

The evidence of inequality could not be conclusive as we cannot reliably say what proportion stems from voluntary choices (e.g., traditional gender contracts) that lead to such disadvantage. Most theories of justice dictate that outcomes are not unjust if they reflect non-coerced and *voluntary choices*. Hence, just distributions are sensitive to individual choice (i.e., individual responsibility), the causes of inequalities of outcome between women and men will bear on whether or not they are unjust. Therefore, one might argue that instead look at the *processes* to most likely to explain the disparities in outcomes between women and men, in order to see if these disparities are indeed morally arbitrary or whether they can be traced to individual responsibility (Gheaus, 2016).

2.3 Fairness metrics evaluation

2.3.1 Set up

To start off, it is essential to explain the decision-making scenario and the important terminologies that will be used throughout this section. Suppose that a supervised machine learning algorithm is being used to make accept/reject decisions about applicants for personal loans. As the first step, a past (observational) dataset from loan applicants (instances) is collected. It includes sensitive input features representing gender of applicants, assuming binary gender of female or male, and non-sensitive features.

The dataset also includes the actual decision about each instance in the dataset, called true labels. It is then split into a training set which is to be used to train the model, and a test set with hidden true labels, which are known, but hidden from it to estimate the accuracy of the model. Using a classification model, the task of the algorithm is to classify instances as accept (low credit risk) or reject (high credit risk) based on the features used in the training set.

Let**:**

- $X \in \mathbb{R}^d$ denote a non-sensitive feature vector describing an individual.
- $g \in \{0,1\}$ denote sensitive binary feature describing the gender of the individual. g is assumed to be a binary variable representing the sex (male or female) of the credit seeker.
- $Y \in \{0,1\}$ represent the *actual* or true labels. Y is either an approved or rejected decision (or good vs. bad credit). As such, it must be denoted as a binary variable in the dataset.
- $S \in [0,1]$ is the predicted probability. For a logistic regression model, it is defined as:

$$S = Pr\,(Y = 0, 1 \mid X, g) =$$
$$exp\,[\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_g g] \,/\, (1 + exp\,[\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_g g\,])$$

where α, β_i , and β_g are the estimated regression coefficients.
- $\hat{y} \in \{0,1\}$ is the prediction decision of the algorithm. It depends on the predicted probability (S) of an applicant having good or bad credit. For a threshold value of S^*, $\hat{y} = 1$ when $S > S^*$, and $\hat{y} = 0$ when $S < S^*$. The table below illustrates the datasets used in the decision-making process.

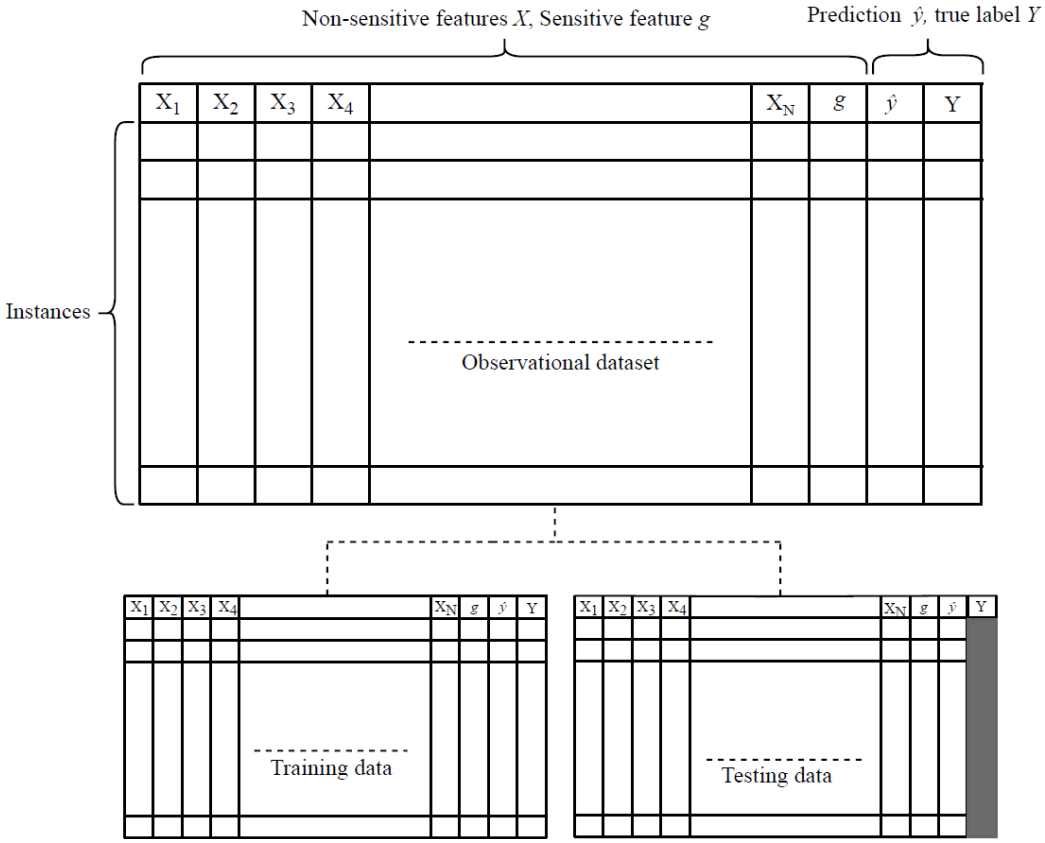


FIGURE 10: Dataset structure: Original data on top divided onto training (shown on left) and testing set (shown on right).

It is important to highlight two key assumptions here about the scenario mentioned above:

- *Both, the accepted and rejected true labels are present in the dataset.* There is bias due to the fact that only applicants who were granted a loan can be observed. That is to say, information about an applicant's actual repayment behavior *after* being rejected by the loan officer is absent.

- *The decisions are in accordance with predictions $\hat{y}$.* Algorithmic decision-making is a two-step process: at first, predictions are made by the algorithm, then decisions are made based on the predictions. In some cases, the relation between predictions and decisions made based on those prediction may not be straightforward. Suppose that output of the machine learning model is the probability score S (or risk score 1- S) and not the binary classification $\hat{y}$. Two applicants with the same risk score may be perceived differently by the loan officer if say, one of them has established a longer relationship with the financial institution, given the model is not trained on such factor. In the automated application scenario described above, the decision is made in a one step process, that is, value of $\hat{y}$ will dictate the faith of decision.

Moving on, the reader is encouraged to keep in mind what (X, g, $\hat{y}$, Y, S) each represent as they will be frequently used throughout the rest of the report. Next, each category of metric is explained and viewed through the lens of theories of guiding principles of justice.

2.3.2 Group-level metrics

Group-level notions of fairness require just (equal) distributions of benefit or harm across *groups*. Metrics belonging to this category are based on observational criteria as they depend on one of the following joint distributions. Throughout this report, they are referred to as subcategories of group-level notions (Verma & Rubin, 2018):

- Subcategory 1: distribution between predicted outcome $\hat{y}$ and sensitive attribute g (marginalizing features X and the actual outcome Y).

- Subcategory 2: distribution between predicted outcome $\hat{y}$, sensitive attribute g, and the actual outcome Y (marginalizing features X).

- Subcategory 3: distribution between predicted probability S, sensitive attribute g, and the actual outcome Y (marginalizing features X and predicted outcome $\hat{y}$).

In the remainder of this section, each subcategory is explained by, firstly, providing the mathematical notions that are used to measure fairness and, secondly, highlighting the statistical fairness criteria rooted in each subcategory. Group-level metrics are often regarded as statistical measures of fairness.

Subcategory 1

There are two metrics that belong to this category commonly known as demographic parity and conditional demographic parity. Demographic parity requires that the positive prediction between the two groups of males and females be equal. Conditional demographic parity extends the definition of demographic parity by conditioning the outcome to be based on certain legitimate factors (L). While there is no unified agreement on what legitimate factors are, it is useful to mention that conventionally, legitimacy of loan approval decisions is based on evaluating 5 key factors, commonly known as 5 C's of credit: credit history, character, collateral, and capacity.

	Metric	**Mathematical notation**	**Requires**
Independence	Demographic parity	$Pr(\hat{y} = 1 \mid g = 0) = Pr(\hat{y} = 1 \mid g = 1)$	Equal probability of positive predictions for both males and females.
	Conditional demographic parity	$Pr(\hat{y} = 1 \mid g = 0, L) = Pr(\hat{y} = 1 \mid g = 1, L)$	Equal probability of positive predictions for both males and females, but condition on a subset of legitimate input features.

TABLE 5:Demographic parity (a.k.a statistical parity, benchmarking) metrics.

The statistical fairness criteria here is that the sensitive features be statistically *independent* of the prediction outcomes or, in other words, the two metrics require that the decision outcomes $\hat{y}$ be uncorrelated with sensitive features g: $\hat{y} \perp g$.

Reverting back to the loan allocation scenario, it can be seen that conditioning fairness on parity measures can be problematic in two ways. First off, demographic parity metrics do not ensure fairness. Consider a bank is required by some law to distribute loans in equal percentages to females and males. For simplicity, assume that the decisions are only based on savings of the applicants. Among the pool

of applicants requesting loans, females happen (randomly) to be those with lower average savings, yet male applicants, on average, have higher savings than the females. Parity metrics will deem the decision-making model fair, despite how the applications are processed based on gender. In this example, the subpopulations are inherently different, meaning that they may not deserve the same outcome. Secondly, the loss in utility of imposing demographic parity can be substantial for no good reason. In other words, demographic parity measures are misaligned with the fundamental goal of achieving higher prediction accuracy. This is due to the fact that demographic parity ignores any possible correlation between decision and the protected attribute.

Subcategory 2

In contrast to demographic parity measures, metrics belonging to this subcategory look beyond the absolute distribution of predictions $\hat{y}$ by conditioning fairness on accuracy or imperfections of the model. To visually investigate this notion the plot below is provided. Let's assume a loan application model looks at three independent features as illustrated below. In an ideal case, a logistic regression model may find a decision boundary such that the two binary classes are linearly separable and that output decisions are 100% accurate.

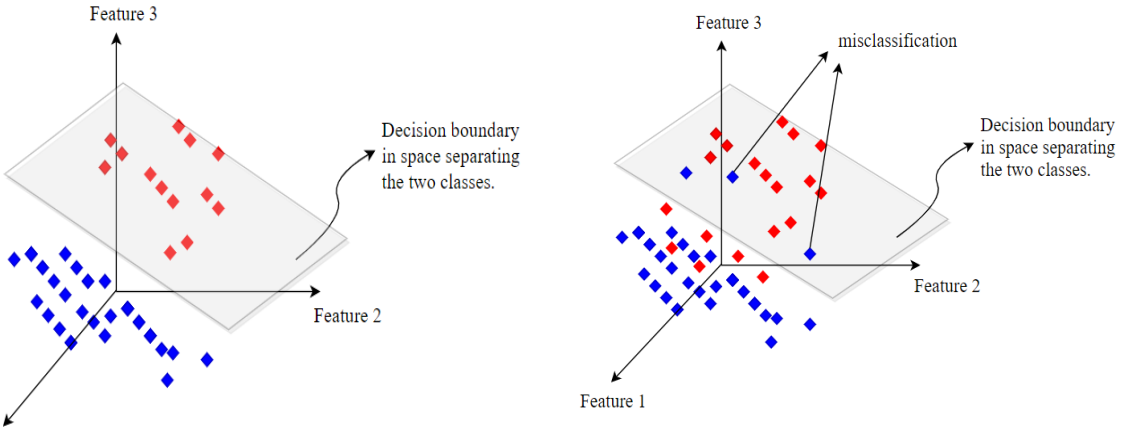


FIGURE 11:Liner regression: The left model is 100% accurate in classifying instances (i.e. Y= ŷ at for all instances). The model shown on the right hand-side misclassifies certain data points.

In another more realistic scenario, if the items in the training data are not perfectly, linearly separable, as shown in the figure above, the linear boundary found misrepresents some of the data points. All the metrics in this subcategory, total of seven, rely on a table commonly known as a *confusion matrix* shown below. Two of the elements in the matrix, row 1 & column 1 (R1C1), and row 2 & column 2 (R2C2), represent the accuracy of predictions with respect to true (actual) label. Two other elements, row 1 & column 2 (R1C2,) and row 2 & column 1 (R2C1), represent inaccurate predictions with respect to true label.

The fractions shown in the matrix below can be expressed as conditional probabilities between the actual and the prediction decisions. For example, FPR represents the probability of erroneously assigning a positive score (recommending credit approval) given that the actual outcome is negative (credit denial): *false positive rate (FPR)* = $Pr\,(\hat{y} = 1 \mid Y = 0) = \frac{FP}{FP+TN}$.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

TABLE 6: Confusion matrix: source: Fairness Definitions Explained by Verma and Rubin (2018).

R1C1: TP: true positive, PPV: positive predictive, TPR: true positive rate
R1C2: FP: false positive, FDR: false discovery rate, FPR: false positive rate
R2C1: FN: false negative, FOR false omission rate, FNR: false negative rate
R2C2: TN: true negative, NPV: negative predictive value, TNR: true negative rate

So how can erroneous predictions of the model lead to disparate mistreatment? In the paper "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment", Zafar and Valera describe a new notion of discrimination called disparate mistreatment (Zafar et al., 2017). Broadly, this notion can be understood as (un)fairness caused by imperfections in the output decisions of the model. Disparate mistreatment can also be realized as the harm or benefit of inaccurate misclassification on some of the instances. Existence of erroneous predictions do not necessarily lead to disparate mistreatment unless misclassification (error) rates for sensitive groups (gender, for the purpose of this thesis) are calculated to be different. In other words, the rate at which the model makes false predictions for females and males is calculated to be different.

Based on the confusion matrix shown above, seven fairness metrics have been proposed, three of which require accuracy and the other four require harm or benefits from false classifications to be equal across protected and unprotected groups. In the case of loan allocation, for instance, equal false positive rate would require that fraction of loans distributed to individuals who truly should belong to negative class (rejection) to be equal for both males and females. In contrast, false negative rates would require false rejection to those who truly deserve the loan (not default) to be equal across the protected and unprotected groups. In reality, it would be impossible to know the true outcome if an individual is predicted $\hat{y} = 0$ and if the first assumption stated above holds, that is, the decisions are based on the prediction $\hat{y}$.

From the confusion matrix several fairness formulations can be derived that are mathematically equivalent of each other. For instance, if a classifier satisfies predictive parity, that is:
$Pr(Y = 1 \mid \hat{y} = 1, g = 0) = Pr(Y = 1 \mid \hat{y} = 1, g = 1)$, then automatically, false discovery rate which is $Pr(Y = 0 \mid \hat{y} = 1, g = 0) = Pr(Y = 0 \mid \hat{y} = 1, g = 1)$ will hold.

At this point, two statistical fairness criteria, separation and sufficiency, underlying the metrics in this subcategory can be introduced. These fairness criteria help with conceptual categorization of metrics belonging to this subcategory. Later in this section, these two criteria are used as a basis for conceptual mapping of the metrics satisfying them onto their equivalent principles of justice

In the table below, the list of metrics with their mathematical counterparts are illustrated.

	Metric	Mathematical notation/Alias	Requires
Accuracy/Sufficiency	Predictive parity	$Pr(Y = 1 \mid \hat{y} = 1, g = 0) = Pr(Y = 1 \mid \hat{y} = 1, g = 1)$ *False discovery rate:* $Pr(Y = 0 \mid \hat{y} = 1, g = 0) = Pr(Y = 0 \mid \hat{y} = 1, g = 1)$	Equal fraction of correct positive predictions for both male and females.
	Overall accuracy equality	$Pr(\hat{y} = Y, g = 0) = Pr(\hat{y} = Y, g = 1)$ *Overall misclasification rate:* $Pr(Y \neq \hat{y} \mid g = 0) = Pr(Y \neq \hat{y} \mid g = 1)$	Equality of positive and negative predictive values for both males and females.
	Conditional accuracy equality	$Pr(Y = 1 \mid \hat{y} = 1, g = 0) = Pr(Y = 1 \mid \hat{y} = 1, g = 1) \cap$ $Pr(Y = 0 \mid \hat{y} = 0, g = 0) = Pr(Y = 0 \mid \hat{y} = 0, g = 1)$	Equal correct positive predictions and equal correct negative predictions for both males and females.
Misrepresentation/Separation	False positive error rate	$Pr(\hat{y} = 1 \mid Y = 0, g = 0) = Pr(\hat{y} = 1 \mid Y = 0, g = 1)$ *True negative rate:* $Pr(\hat{y} = 0 \mid Y = 0, g = 0) = Pr(\hat{y} = 0 \mid Y = 0, g = 1)$	Equal incorrect positive predictions for both male and females.
	False negative error rate	$Pr(\hat{y} = 0 \mid Y = 1, g = 0) = Pr(\hat{y} = 0 \mid Y = 1, g = 1)$ *True negative rate:* $Pr(\hat{y} = 0 \mid Y = 0, g = 0) = Pr(\hat{y} = 0 \mid Y = 0, g = 1)$	Equal incorrect negative predictions for both male and females.
	Equalized odds	$Pr(\hat{y} = 1 \mid Y = 0, g = 0) = Pr(\hat{y} = 1 \mid Y = 0, g = 1) \cap$ $Pr(\hat{y} = 0 \mid Y = 1, g = 0) = Pr(\hat{y} = 0 \mid Y = 1, g = 1)$	Equality of false positive and false negative error rates for both males and females.
	Treatment equality	(*FNR/FPR*) for males = (*FNR/FPR*) for females	Equal ratio of false positive and false negative error rates for both males and females.

TABLE 7:Subcategory 2: Definitions based on predicted outcome ŷ and true label Y. Accuracy metrics (shown above) and misrepresentation metrics (shown below).

Separation allows correlation between the prediction and the protected attribute to the extent that it is justified by the target variable. In this context, the true labels are split into two categories of approve and reject, separation requires for the population defined by each category that gender does not influence the statistical prospects of a decision. This intuition can be made precise with a simple conditional independence statement: separation is satisfied when the prediction (and therefore decision) ŷ is statistically independent of group-membership g conditional on the true label Y: $\hat{y} \perp g \mid Y$. The disparate mistreatment notions, such as equalized odds and its weaker forms, false positive rates and false negative rates, as well their mathematical counterparts can be described as fairness notions satisfying separation (Heidari et al., 2019).

Separation: Individuals with the same true label have the same statistical prospects of either decision, regardless of their irrelevant feature.

Sufficiency, on the other hand, allows correlation between the true label and the protected attribute to the extent that it is justified by the prediction outcome. Mathematically, it can be expressed that sufficiency requires the true label Y to be statistically independent of group-membership conditional on prediction outcome ŷ (or decision): $Y \perp g \mid \hat{y}$. The accuracy notions, such as predictive parity, along with its false discovery rate counterpart, can be described as metrics satisfying the sufficiency fairness criteria (Heidari et al., 2019).

Sufficiency: Individuals about whom the same decision is made have the same statistical prospects of being either true label, regardless of their irrelevant feature.

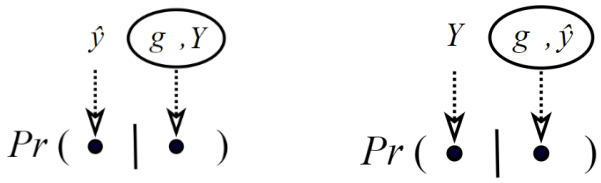


FIGURE 12:Statistical fairness criteria: Separation and Sufficiency.

In the illustration above, the mathematical formulation of separation is shown on the left and sufficiency on the right. Understanding the difference between the two statistical criteria is essential for determining which of the two are suitable for a particular application.

Subcategory 3

The fairness notions in this subcategory are based on some distribution between predicted probability S, sensitive attribute g, and the actual outcome Y. The four metrics belonging here are described in the table below. Noting that the predicted probability S and prediction ŷ are closely related- ŷ is derived from prediction probability S and a certain threshold S*- the four metrics shown in the table can be thought of as some similar metrics from subcategory 2. For instance, if a classifier satisfies a positive rate balance, that is, the protected and unprotected groups of applicants with true positive label to have equal average predicted probability score S, a consistent result would require false negative error rates to be satisfied.

Metric	**Mathematical Notation**	**Requires**	**Similar to**
Calibration	$Pr(Y = 1 \mid S = s, g = 0) =$ $Pr(Y = 1 \mid S = s, g = 1)$	Equal probability to belong to true positive label.	Predictive parity (except it considers fraction of positive predictions)
Well-calibration	$Pr(Y = 1 \mid S = s, g = 0) =$ $Pr(Y = 1 \mid S = s, g = 1) = s$	Equal probability to belong to true positive label and the probability to equal *S*.	Predictive parity (except it considers fraction of positive predictions)
Balance for positive class	$E(S \mid Y = 1, g = 0) = E(S \mid Y = 1, g = 1)$	Equal expected value of probability S for individuals with positive true label.	• False negative error rate • True positive rate.
Balance for negative class	$E(S \mid Y = 0, g = 0) = E(S \mid Y = 0, g = 1)$	Equal expected value of probability S for individuals with negative true label.	• False positive error rate balance • true negative rate

TABLE 8:Subcategory 3: Definitions based on predicted probability S and true label Y.

The fairness criteria imposed on these metrics are in correspondence to their similar metric in Subcategory 2, that is: calibration and well-calibration satisfy sufficiency while balance for positive class and balance for negative class satisfy separation.

2.3.2.1 Group-level measures and guiding principles of justice

So far, the metrics which measure fairness across groups have been introduced. These various measures are then categorized based on their mathematical notation and, more importantly, their statistical fairness criteria. This categorization makes it easier to make comparison across the measures. The table below summarizes the group-level metrics.

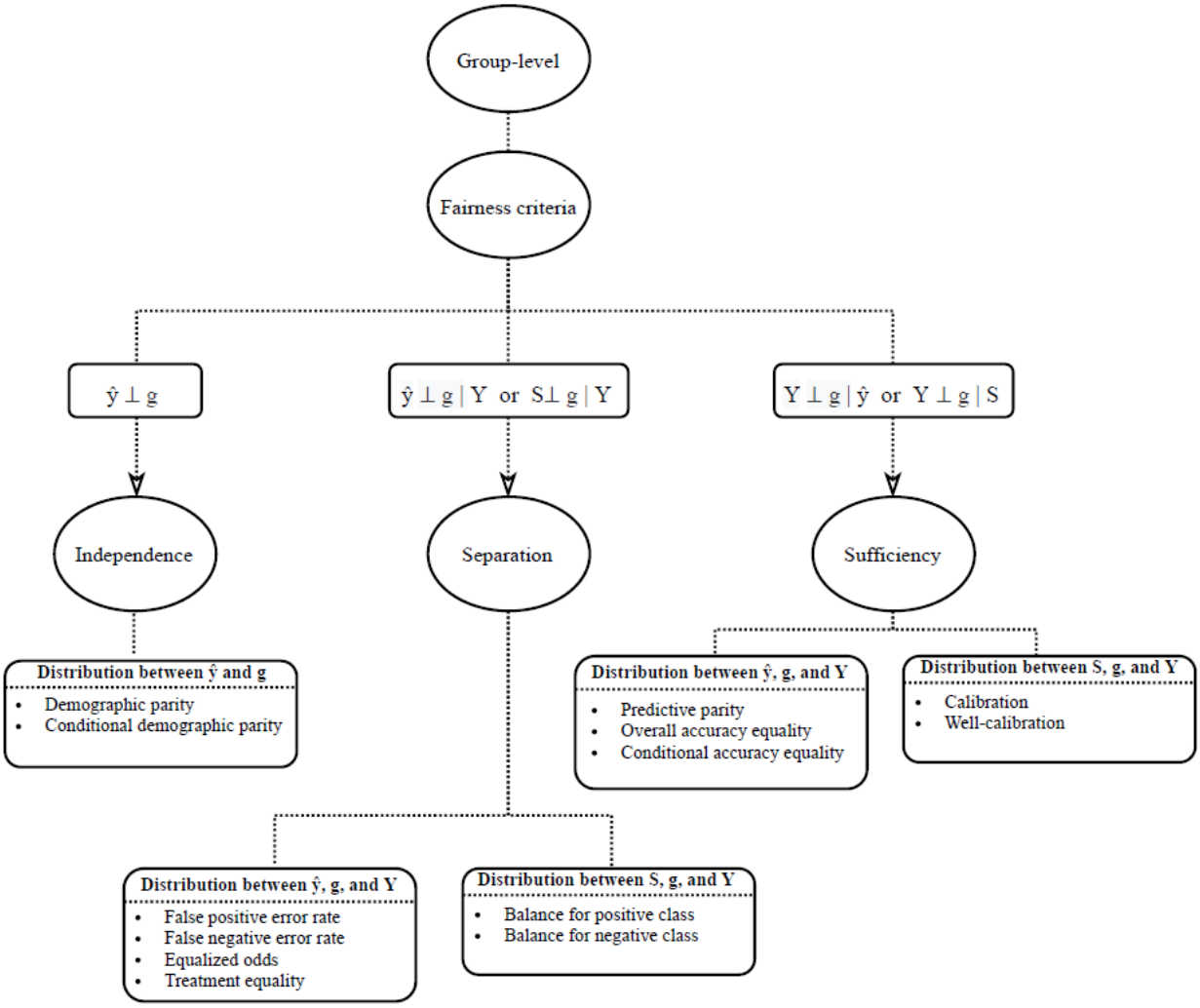


FIGURE 13: Categorization of group-level metrics based on their statistical fairness criteria

Ideally, one would desire to achieve an algorithmic decision-making model that satisfies all the group-level notions, but as it will be explained here, this is not achievable unless for some improbable circumstances. Consequently, careful choices have to be made between one subcategory and another and this is what the remaining part of this synthesis on algorithmic fairness will investigate.

The two fairness criteria of separation and sufficiency cannot be satisfied together, except for a narrow range of circumstances that are extremely rare: when the features values are identical across the two groups or the model is 100% accurate (Chouldechova, 2016; Kleinberg et al., 2016). For instance, calibration and equalized odds cannot be achieved simultaneously unless the groups have equal base rates, that is the same fraction of positive or negative instances. This implies that those developing the models have to choose between different types of fairness requirement and it seems like inevitably some discrimination is unavoidable. The machine learning community considers this to be a genuine dilemma where one needs to make a choice between different fairness criteria but both choices seem wrong. It

has been described as the "trolley problem of machine learning" and some have compared it to Kenneth Arrow's famous impossibility theorem in social choice theory (Heidari et al.,2018).

The impossibility in simultaneously achieving multiple metrics goes beyond the mutual exclusiveness of separation and sufficiency. It can be shown mathematically that independence and sufficiency, and independence and separation also cannot be satisfied at the same time unless for very specific circumstances (Shira et al., 2018) (see table below). While the mathematical proofs of the impossibility results are beyond the scope of this research project, the list of tensions between the three fairness criteria is expressed in the following table for reference (Kleiberg et al., 2016; Chouldechova, 2016; Shira et al., 2018).

Separation and sufficiency	Case 1	Case 2	Case 3
Assuming (S or $\hat{y}$)	$(S \perp g \mid Y)$ and $(Y \perp g \mid S)$	Balance for positive class, balance for negative class and, calibration within groups.	Equal false positive rates, false negative rates, and positive predictive parity across groups.
Conditions to hold (at least one)	• $(Y, S) \perp g$ • An event in the joint distribution has probability zero.	• Equal base rates • Perfect prediction	• Equal base rates • False positive rate = 0 and positive predictive parity = 1 • False positive rate = 0 and false negative rate = 1

Independence and separation	
Assuming (S or $\hat{y}$)	$(S \perp g \mid Y)$ and $(S \perp g)$
Conditions to hold (at least one)	• Equal base rates • $Y \perp S$

Independence and sufficiency	
Assuming (S or $\hat{y}$)	$(Y \perp g \mid S)$ and $(S \perp g)$
Conditions to hold	$Y \perp g$

TABLE 9:List of impossibility results: Unless for very improbable and specific circumstances, the three fairness criteria cannot be held simultaneously. These cases are investigated in more detail in the Appendix.

The trolley problem of machine learning leads to the following challenge: if, outside the improbable and special circumstances, all the three statistical fairness criteria cannot be achieved simultaneously, how can group fairness be achieved?

To start investigating the question, it is useful to reiterate what the group-level notions require for satisfying fairness. Group-level notions of fairness require a certain metric, quantifying prospects or likelihood of receiving benefit or harm to be equal across the socially salient groups. This definition of group fairness is coupled with the definition of equality of opportunity that was in section 2.1.

Going forward, any factor that is considered a legitimate source of inequality, accountability, or ambition (e.g. shopping habits) is called effort or e. Circumstances c are all factors which are irrelevant

(e.g. gender) and should not affect the outcome. Finally, luck *l* represents involuntary factors which, under Rawlsian interpretations, are a legitimate source of inequality and, under luck egalitarian, are illegitimate and their effects should be neutralized (Heidari et al., 2019).

For the sake of simplicity, let's assume that all luck is 'fair luck' and that the decision-making model is to achieve Rawlsian equality of opportunity. It can be shown that a policy or practice satisfies the Rawlsian equality of opportunity if and only if (Heidari et al., 2019; Lefranc et al., 2009):

$$Pr\ (outcome\ U \mid e, c) = Pr\ (outcome\ U \mid e, c')$$

The expression above specifies that the prospect of receiving an outcome (harm or benefit) U, should be equal for individuals that have the same effort e, and should be irrespective of their circumstance c (Loi et al., 2019).

Since the above formalization is frequently used and mentioned throughout this work, it is vital to highlight some of its strengths and weaknesses that one should consider in conceptualizing fairness.

Strengths:

- The probabilistic equation has high resemblance with the way in fairness metrics are defined. This makes the conceptual mapping of the metrics a straightforward task as illustrated in Figures 8, 9 and 10.
- All the three fairness criteria can be analyzed using the above formalization.
- The formalization of equality of opportunity is neutral in the way it defines effort-based utility; Effort-based utility (e) is what justifies inequality or unequal distribution, and it can be different factors such as need, merit, or responsibility.

The third strength highlighted here is particularly important as it provides a descriptive way to formalize equality of opportunity but, when applied in practice, it can be used a normative guideline. To understand this notion, consider the following two banks that operate under the same economic conditions:

During periods of economic recession, bank reserves increase since consumers and businesses reduce consumption and lending. As such, in a post-recession era, often banks find they have too much cash in their reserves.

Bank A. Suppose that a bank A is heavily supported by insurance and governmental support (as is often the case in post-recession or post-financial crisis era). In order to boost lending and to circulate its cash, this bank offers loans to those consumers, but approves those who have more *need* for the loan. In return, it charges higher interests to earn higher gains for the risk.

In conforming to equality of opportunity, individuals who have the same level of *need* should be given equal prospects of receiving the loan, irrespective of their special traits such as gender or race.

Bank B. Suppose that bank B also has excess cash reserves, but it does not have the security of Bank A. Therefore, it decides to allocate its cash to those applicants who are perceived to be less risky.

In conforming to equality of opportunity, individuals who have the same level of *merit* should be given equal prospects of receiving the loan, irrespective of their special traits such as gender or race.

Weaknesses:

- The formalization does not provide any guideline on how to define/measure utility. For instance, for the case of loan allocations benefit (harm) can be defined as receiving (not receiving) the loan. Benefit (harm) can also be defined as not default (default). In the first example, the utility stems from decisions while in the second example, it stems from self-action of individuals.

- Utility and effort should be thought of as scalar factors. This can be problematic and a source of bias.

- Using this formalization requires the assumption that effort-based utility is not affected by irrelevant circumstance nor the predictive model (decisions).

When reading the remainder of this section, the reader is asked to bear in mind the strengths and weaknesses mentioned above.

Independence

It can be shown that a decision-making model satisfies independence if and only if the harm or benefit to the individuals is the decision (or prediction) and all the individuals have the same (constant) effort-level. Visually, this can be expressed below by mapping (U to ŷ), c to a binary group classification g and with e being a constant. Based on this mapping, the underlying moral assumption of independence or statistical parity metrics can be revealed: metrics explained by independence are suitable for scenarios in which all the individuals have the same effort-based utility.

$$Pr(U \mid e, c)$$
$$\downarrow \quad \downarrow \quad \downarrow$$
$$Pr(\hat{y} \mid constant, g)$$

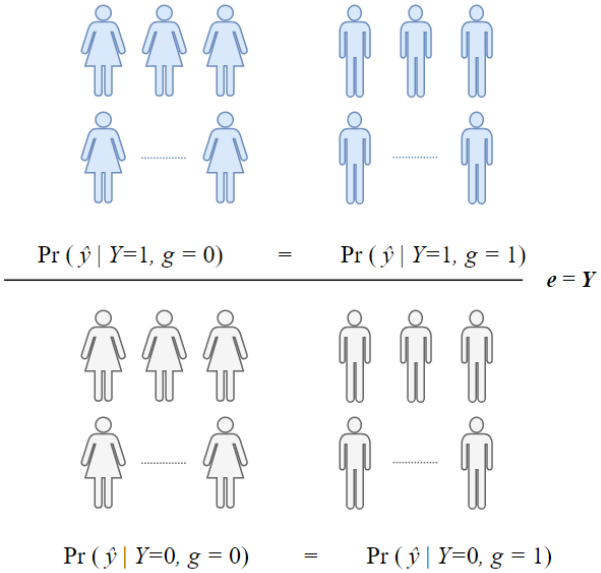
FIGURE 14: Conceptual mapping of statistical criteria independence as a form of Rawlsian equality of opportunity.

In the case of automated loan allocations, imposing independence as a fairness constraint results in the same rate of loans getting distributed amongst the two demographic groups defined by gender, regardless of the fact that individual's ability to repay might be inherently different in each group. In some cases, this would mean the allocation of loans to individuals with dubious repayment ability.

Separation

In a similar manner, separation can be thought of as a condition in which harm or benefit of the individuals stems from the decision and what justifies inequality is the true label. Visually, this can be illustrated by replacing U with ŷ, and e with Y, assuming circumstance c is the binary group classification g. This allows us to spill out two moral assumption underlying separation: harm/benefit

is a result of an algorithm's decision and that the true label/action of individuals reflects their effort-based utility. The latter assumption indicates that metrics belonging to separation are suitable for situations in which it can be shown that individuals with similar true labels are equally accountable for their true labels Y.



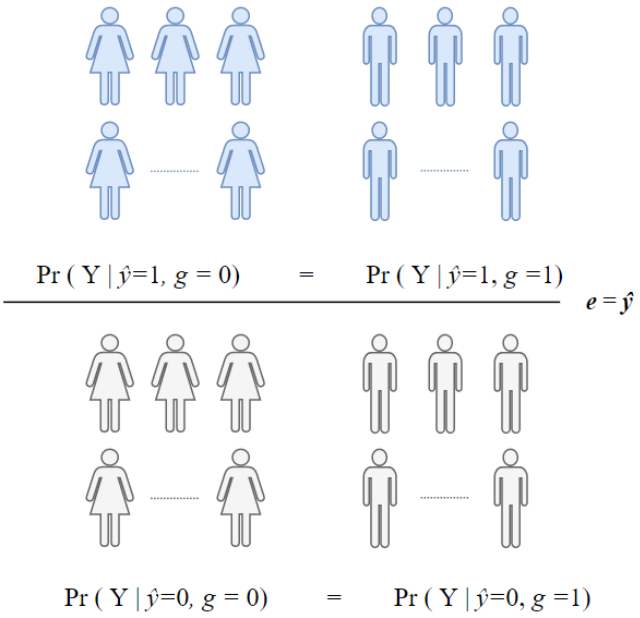
$$Pr(U \mid e, c) \rightarrow Pr(\hat{y} \mid Y, g)$$

FIGURE 15:Conceptual mapping of statistical criteria separation as a form of Rawlsian equality of opportunity.

By looking at separation through the lens of equality of opportunity, it can be concluded that harm/benefit is the decision outcome, effort-based utility or what justifies inequality is the candidates' true credit, or default (no default), and finally, the irrelevant circumstance is gender. Only after validating such moral equivalency assumption separation can be taken as an appropriate fairness criterion.

Sufficiency

Finally, sufficiency is satisfied when harm/benefit received is the true label and predictions or decisions are factors that justify inequality. By replacing *U* with *Y*, *e* with $\hat{y}$, and assuming circumstance c is the binary group-membership g in the formal notation of equality of opportunity, separation can be shown as below. There are two key moral assumptions that follow this conceptual mapping: the first one is that harm/benefits to the individuals are not from the algorithmic decision (prediction $\hat{y}$) received and are result of action of the individuals themselves, and secondly, decisions reflect the effort-based utility of individuals.



$$Pr(U \mid e, c) \downarrow \quad \downarrow \quad \downarrow$$
$$Pr(Y \mid \hat{y}, g)$$

FIGURE 16:Conceptual mapping of statistical criteria separation as a form of Rawlsian equality of opportunity.

Sufficiency seeks to equalize the prospect of truly having a bad/good credit score for individuals who are assigned the same prediction (decision).

By looking at sufficiency through the lens of equality of opportunity, it can be concluded that harm/benefit is grounded in applicants' true credit worthiness or default (no default), effort-based utility is the decision outcome that stems from predictions, and irrelevant circumstance is gender. After validating such moral equivalency assumption, sufficiency can then be taken as an appropriate fairness criterion.

2.3.3 Similarity-based measures (individual-level)

As the name suggests, this definition category takes a more comparative approach in explaining fairness. Similarity-based measures attempt to overcome the shortcomings of group-level measures by not marginalizing over non-sensitive features *X*. There are three fairness metrics that fall under this category: causal discrimination, fairness through unawareness, and fairness through awareness. Causal discrimination and fairness through unawareness strictly state that two individuals who only differ by a sensitive attribute (here gender *g*) and are otherwise *identical* should have the exact same classification outcome.

The difference between the two metrics is that causal discrimination requires the model to be trained on the sensitive feature *g* and fairness through unawareness excludes sensitive attribute *g* from the feature set. In a situation where fairness through awareness is satisfied, but causal discrimination is not, it can be concluded that some features in the data may be used as proxy for gender. As the comparison is made amongst two individuals, the two metrics can only be used if there exists an identical applicant who otherwise differs by gender in the training set. Fairness through unawareness eases this tension by

requiring that similar individuals should have similar classification outcome (Dwork et al., 2012). In this case, similar individuals are those who have similar repayment ability and creditworthiness.

Metric	Mathematical notation	Requires
Causal discrimination	$(X_i = X_j) \cap (g_i != g_j) \rightarrow \hat{y}_i = \hat{y}_j$	Two identical individuals that otherwise have different gender, receive the same classification.
Fairness through awareness	$(X_i = X_j) \rightarrow \hat{y}_i = \hat{y}_j$	Two identical individuals that otherwise have different gender, receive the same classification, and no sensitive feature is explicitly used in the model.
Fairness through unawareness	$(X_i \sim X_j) \cap (g_i != g_j) \rightarrow \hat{y}_i = \hat{y}_j$	Similar individuals receive similar classification

TABLE 10:Similarity based measures, also known as individual-level metrics.

Ensuring that identical or similar individuals that only differ by gender receive *similar classification outcomes* is adhering to *distributive justice* and failing to do so is an indication of direct discrimination or indirect discrimination in the absence of bad intentions and animosity.

In a more representative loan application scenario where the decision outcome is not just to approve/reject an applicant, but also represents the *amount* of loan granted between two individuals, distributive justice may be achieved in three operationalized distinct ways (Huang et al., 2019). These three definitions of fairness are explained in more details below in their respective order:

- Similar individuals are treated similarly (Dwork et al., 2012)
- A less creditworthy individual is never favored over a better one (Joseph et al., 2016)
- Individuals are selected in proportion to their merit (Liu et al., 2016).

According to a study done by Dwork et al, fairness is achieved by providing similar treatment to individuals deemed similar based on a stipulated similarity distance metric specific to a given context. This is the same as *fairness through awareness* metric that is mentioned above. In ML, fairness of an algorithm would thus be determined based on satisfaction the continuity and similarity metric, the Lipschitz condition, defined with regards to this metric (Huang et al., 2019). In the context of lending, this would translate into a decision-making algorithm allocating a similar amount of money to individuals with similar repayment rates (Dwork et al., 2012).

In another study by Joseph et al., fairness is described as always selecting, under a setting in which only one individual is to be favorably selected, an individual with higher expected value (of some measure of intrinsic quality). To put differently, selected is always the better individual with a probability greater than or equal to the probability of choosing the worse individual. According to this definition, rewards are attributed to an individual proportionally based on the level of her inherent quality, in other words, through meritocracy. In the context of lending, this translates into a decision-making algorithm allocating an individual with a higher repayment rate at least or more the same amount of money as the other candidate (Joseph et al., 2016).

Last but not least, in a study by Liu et al., fairness is defined, in the setting of consequential decision making, as selecting individuals *proportionally* based on their merit. This so-called 'calibrated fairness'

implies the fairness *principles of meritocracy* of Joseph et al., and a suitably chosen similarity metric of Dwork et al. In the context of lending, this translates into a decision-making algorithm allocating two individuals with repayment rates r_1 and r_2 with $\frac{r_1}{r_1+r_2}$ and $\frac{r_2}{r_1+r_2}$ amount of money, respectively (Liu et al., 2017).

In a recent study, Huang et al. (2019) attempted to find support and evidence for the suitability of these three definitions through investigating people's perceptions of algorithmic fairness in the context of loan allocations. Across two online experiments, they test which of the three definitions people perceive to be the fairest in the context of loan decisions, and whether fairness perceptions change with the addition of The results of the research indicated that the most supported definition, was the one in which the money was allocated to candidates in *proportion* to their repayment rates, was the notion of *calibrated fairness* (Huang et al., 2019).

2.3.4 Causal reasoning

Causal reasoning metrics are non-observational criteria that take a different approach at defining fairness. Definitions in this category assume a given causal graph (Kilbertus et al., 2017). In contrast to the statistical and similarity-based measures, where fairness is determined based on some joint distribution of prediction and true outcomes between groups or individuals, causal reasoning can be used to capture *relationships* between different attributes and their *influence* on outcome. In other words, causal reasoning metrics do not depend on values of the predictions $\hat{y}$.

There are four metrics that belong to this category as outlined in the table below. Causal metrics are especially useful criteria in situations where simply *ignoring* the sensitive attribute is ineffective if it is possible for the model to learn the sensitive attribute by means of non-sensitive attributes (Kilbertus et al., 2017).

Metric	Requires
Counterfactual fairness	The predicted outcome $\hat{y}$ to not depend on a direct descendant of the protected attribute in a causal graph.
No proxy discrimination	No path from the protected attribute to the prediction $\hat{y}$ through a proxy in a causal graph.
No unresolved discrimination	No unresolved path from protected attribute to the prediction $\hat{y}$ in a causal graph.
Fair inference	No illegitimate path from prediction $\hat{y}$ to the protected attribute.

TABLE 11:Causal metrics: these definitions depend on a causal graph that capture relationship between input features.

To visually express the information presented in the table above, assume the causal graph below is provided. Suppose that a prediction model is mandated to ignore the sensitive attribute *g*, and that the prediction outcome $\hat{y}$ is only to be drawn from the following four non-sensitive attributes: credit amount, income, age, and the industry the applicant is employed in. For simplicity, let's also assume that attributes *X* are uncorrelated (orthogonal) with each other.

Starting with counterfactual fairness, it can be observed that the causal graph below fails to meet this criterion as there is since income is a direct descendant of gender. Income in this situation also acts as a *proxy* for gender. In other words, based on the identifiable correlation between income and gender, applicant's gender can be derived from his/her income. Thus, it can be concluded that *discrimination by proxy* arises here due to the use of a feature (income) which is correlated to the sensitive attribute (gender). This definition is consistent with the notion of indirect discrimination that was introduced earlier. The feature industry in this hypothetical scenario is a deemed a resolving edge, meaning that it is influenced, by the protected attribute- and hence the prediction $\hat{y}$ is influenced indirectly- but in a non-discriminatory manner. Income on the other hand, is an unresolving edge which implies that the causal graph fails to meet this definition of fairness.

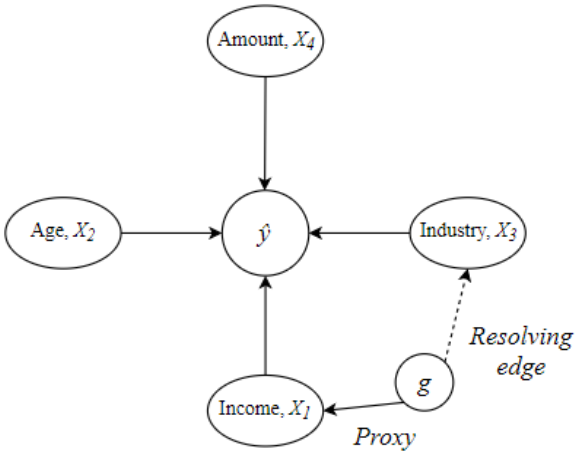


FIGURE 17:Causal graph describing relationships between features in a hypothetical loan approval process.

Finally, fair inference depends on the legitimacy of the path from the protected attribute to the predicted outcome. In an automated decision-making, application legitimacy means that the use of a particular feature is justified even though it might act as a proxy for the sensitive attribute. In practice, for instance, many banks calculate *debt-to-income ratio* as a parameter for determining capacity of the candidate to repay his debt. Thus, acquiring and verifying information regarding income of applicants is an essential part of the loan approval process. In the hypothetical scenario above, even though income is a proxy for the protected attribute, it is a legitimate feature to use and therefore, the causal graph satisfies fair inference.

As described earlier, this category of fairness emphasizes the importance of selecting features that are non-discriminatory and legitimate to use for making accurate and unbiased decisions. In contrast to the previous two categories, this category is not concerned with the state of affairs per se, but rather with the way in which that state of affairs was produced.

However, it is important to mention that it is a very difficult task to determine which features are influenced by a sensitive attribute in a model. In practice, it is hard to reach a consensus in terms of what the causal graph should look like and it is more difficult is to determine the degree in a which each feature influences the predictions. This is especially the case when the model uses vast number of features.

Even if it is possible to determine which factors are influenced by a sensitive feature with a high degree certainty, there is no consensus on whether the features are indeed legitimate, and if they can be deemed as resolving or non-resolving edges. A multidisciplinary approach involving sociology, economics,

philosophy, history, and other nascent disciplines is needed to reflect on local social values and to determine the specific ways in which some groups become to be unfairly disadvantaged.

While existing fair learning mechanisms can leverage input features and their correlations with the sensitive attributes in order to resolve indirect discrimination and achieve distributive fair outcomes, they overlook several important considerations which are addressed by procedural fairness. For example, these considerations include whether or not the perceived fairness of using an individual's feature in the decision-making process is affected by the following:

- Certain features represent the result of volitional decisions made by the individual (Beahrs, 1991).
- Unbiased, relevant, and legitimate features which represent accurate information about individuals are causally related to the decision outcomes (Kilbertus et al. 2017; Kusner et al. 2017).
- Collective opinions – rooted in prevailing cultural or social norms, political beliefs, and legal regulations – reflect societal consensus on the desirability of using particular features.

In conclusion, understanding how people exhibit a moral sense for whether or not it is fair to use a feature in a decision-making scenario captures the above procedural fairness considerations (Yaari & Bar-Hillel, 1984; Zafar et al., 2018). Evaluating moral judgements and opinions on input features can help ensuring that a process is free of any animosity or any indirect discrimination which serves for equality of opportunity. These opinions (implicitly) provide the missing background knowledge needed to evaluate the fair use of input features in a given decision process.

2.4 Chapter Summary

In this chapter, the first research question is investigated:

> *What fairness metrics are suitable for assessing gender equality in consumer lending?*

Before diving into the technical analysis of the 21 definitions of fairness proposed by Narayana, this chapter examines and conceptualizes gender equality in the context of consumer lending. If the good that is being distributed amongst the borrowers is considered to be opportunity to access loans, distributive and procedural justice, two widely supported guiding principles and aspects of fairness that are rooted in applicants' perception, can be used to specify how opportunity can be distributed amongst the individuals. In short, fairness can be defined as:

Individuals who have the same (merit, need, or whatever justifies inequality) should be given equal prospects of receiving the same treatment, irrespective of their gender.

Based on this definition, the three categories of notions of fairness (group fairness, individual fairness, and causal measures) are then viewed through their analogous guiding principles of justice. It is shown that group fairness definitions can be thought of as a special case equality of opportunity. Based on the application scenario, one of the three fairness criteria of independence, separation, and sufficiency should be chosen as an appropriate criterion for assessing fairness of the decision-making model.

Similarity-based measures, also known as individual-level metrics require the outcome to be similar amongst two individuals. As such, they can be thought of as measures of distributive justice. To use individual fairness metrics expert knowledge of the field of lending as well as statistics is required. One of the benefits of these metrics is that they can be used to determine whether there is a proxy for the protected attribute in the model and whether gender causally relates to the output. All three metrics in this category are suitable for measuring gender bias for the application scenario at hand. However, as mentioned earlier, one of their shortcomings is that *similarity* needs to be defined by a person and this can lead to implicit bias.

Finally, causal measures overlook the distribution of outcome and emphasize the importance of appropriate input feature selection through a multidisciplinary approach. Through selecting features that are perceived unbiased, fair, legitimate, and reflective of voluntary choices of individuals, causal measures adhere to procedural justice. Determining causal graphs is a very difficult task that requires careful experimental analysis of the input features to determine their relationship to the protected attribute and to the prediction outcome.

This page is intentionally left blank.

3

Bias detection

with the AIF 360 toolkit

3.0 Introduction

So far, a synthesis which puts conceptualizing gender equality in automated consumer lending at its heart has been provided. In doing so, some of the previous research and contributions of gender studies, philosophy, history, economics, and computer science are brought together to form a cognitive process which is to aid with the assessment of fairness metrics.

This chapter of the thesis turns the attention towards applying some of those metrics using IBM's AIF 360 toolkit, an open source Python toolkit for algorithmic fairness, to empirically investigate how those metrics can be implemented in practice to observe gender bias in the classification models. In this regard, this part of the work develops a classifier in a simple, yet, practically relevant process to classify a sample loan applicants in a dataset. The research question dealt with here is:

> *How can the metrics be applied to observe gender bias in lending history data?*

The dataset used in this work is an open source data, commonly known as "German Credit Data", provided by Prof. Hofmann. It contains information about 1000 candidates (instances), on the basis of

which they have been labeled as having a good or bad credit. More details about the dataset is provided in the following section.

To recap, there are 21 widely used fairness metrics belonging to three categories of a) group-level or statistical measures, b) individual-level, and c) causal measures. Broadly speaking, group fairness divides a population into two sub-populations of protected and unprotected groups while seeking for some statistical measure to be equal across the two. Individual fairness, in its broadest sense, seeks for identical or similar individuals to be treated similarly. Similar treatment implies receiving similar prediction outcomes. Causal measures emphasize the use of nondiscriminatory and legitimate features with aim to model how features influence the prediction outcomes.

It is important to mention that, going forward, this work narrows its focus to only *group fairness* for two main reasons. Firstly, group fairness metrics are controversial in a sense that they cannot all be achieved simultaneously. Such incompatibility (also known as the trolley problem of machine learning, see Appendix for more explanation) makes it encouraging to investigate the tensions between the metrics in this category.

Secondly, definitions belonging to this category depend on the availability of both classes of true labels. In practice, however, negative classifications (true negatives) are often not available for loan allocation settings (it is not possible to know if a rejected person, would otherwise default or not). German Credit Data offers both classes of true labels, hence this completeness makes the investigation of group fairness possible.

This chapter is structured as follows: An overview of the AIF 360 toolkit is provided in section 1. Section 2 provides a detailed description of the dataset used and utilizes the toolkit to test for bias in the mentioned dataset. In section 3, a classification model is built, and the AIF 360 toolkit is used to test the classifier for group fairness. Section 4 provides a summary of the chapter.

3.1 AIF 360 toolkit overview

The AIF 360 is an extensive toolkit that can be used for detection, mitigation, and explanation of bias in datasets and classification models. As outlined in the introduction, the primary focus of this chapter is to investigate how the toolkit can be used to *detect* bias. In providing an overview of the toolkit, it is also vital to show in what stage of the machine learning process the toolkit can be used to detect for bias.

A simple machine learning process can be divided into two phases: model development and model deployment. During model development, problems and business objectives are adequately defined, data is collected, prepared, and finally, the classification model is trained. In model deployment, the classifier is used internally or by external users, such as clients.

To visually investigate where in the development process the toolkit can be used to test for bias, a simple machine pipeline is provided below. In this pipeline, the process commences with preparing the raw data and randomly splitting it into the training and test partitions. Here, instances in each partition have two components: attributes X, g and the true label Y. Subsequently, a machine learning algorithm is trained on this training dataset in order to produce a classification model. Predictions (decisions) $\hat{y}$ can be then obtained for each instance using the classifier. At this stage, accuracy of the model can also

be evaluated using the test dataset. Next in the process comes model deployment. If the model is being deployed by a third-party user, external interventions may be implemented. Finally, the model may be used either once or repeatedly throughout an ongoing data mining process.

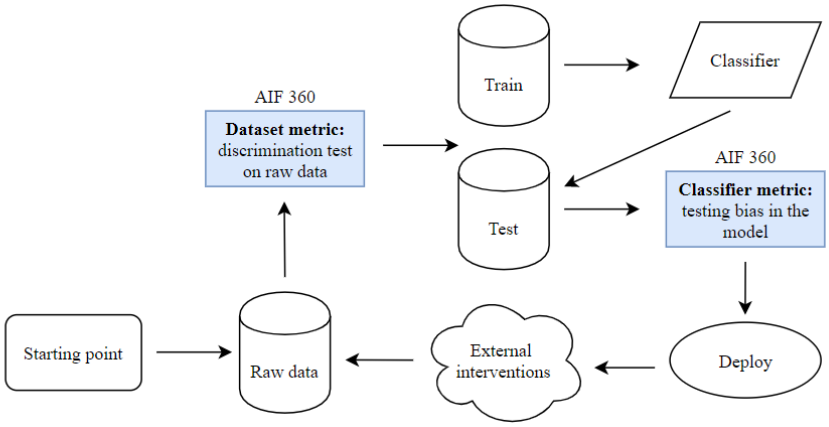


FIGURE 18:Two points of intervention by the toolkit are indicated on the diagram. Adapted from d'Alessandro et al. (2017).

The starting point of the pipeline, *raw data*, is the point where the toolkit can be used firstly to test for bias. Discrimination test on the training data can be executed by applying the metrics in the `DatasetMetric` class (and its subclass `BinaryLabelDatasetMetric`) of the toolkit (see Section 3.2). It is vital to mention that raw data represents all the stages that play a part in the selection, preparation, and process of input data. Therefore, numerous types of biases can be checked for at this starting point.

The second point in the pipeline where bias detection can be executed is shown on the right side of the figure. Here, the *classifier* is tested for fairness. When the application requires testing the classifier, the ones in the `ClassificationMetric` class should be used.

The toolkit offers group-level and individual-level fairness metrics and does *not* support causal measures. If the application is concerned with individual fairness, the metrics in the `SampleDistortionMetric` class should be used. If the application is concerned with group fairness, then the metrics in the `DatasetMetric` class (and in its children classes such as the `BinaryLabelDatasetMetric` class) as well as some of the metrics in the `ClassificationMetric` class should be used.

Now that the toolkit's intervention points have been demonstrated, attention can be steered towards identifying the types of bias, which can be measured with the toolkit in the German Credit. As a preliminary step, it is important to provide information about its features, true labels, and its limitations in predicting credit worthiness.

3.2 Dataset

3.2.1 Documentation and limitations

The dataset used in this work contains 1000 entries representing applicants who wish to request a loan from a bank. It includes a total of 21 variables, 20 of which are categorical/numerical input features

and 1 column representing categorical (binary) true labels based on the actual classification of the applicants.

	Features	Description	Type
1	Duration in months	Terms of financing. Longer term implies higher risk	Numerical
2	Credit amount	Principal amount of loan being requested	Numerical
3	Installment as percentage of income	Often referred to as debt-to-income ratio, It is an indication of repayment ability.	Numerical
4	Residency	Number of years living at the current house/apartment	Numerical
5	Age	Age in years	Numerical
6	Number of credits	Indicates number of revolving credit accounts at the financial institution.	Numerical
7	Number of people liable for	Number of persons entitled to maintenance	Numerical
8	Gender	• Male: Single / Divorced / Married • Female: Divorced / Married *(no single female applicant)*	Categorical
9	Status of checking account	• No running accounts • No balance or debit • 0 < ... < 200 DM • ... >= 200 DM or checking account for at least 1 year	Categorical
10	Credit history	• No credits taken/ all credits paid back duly • All credits at this bank paid back duly • Existing credits paid back duly until now • Delay in paying off in the past • Critical account/other credits existing (not at this bank)	Categorical
11	Purpose	• Car (new) • Car (used) • Furniture / Equipment • Radio / Television • Domestic appliances • Repairs • Education • Vacation • Retraining • Business • Others	Categorical
12	Savings	• ... < 100 DM • 100 <= ... < 500 DM • 500 <= ... < 1000 DM • ... >= 1000 DM • Unknown/ No savings account	Categorical
13	Employed since	• ... <= 1 years • 1 <= ... < 4 years • 4 <= ...< 7 years • >= 7 years	Categorical
14	Other debtors	• None • Co-applicant • Guarantor	Categorical
15	Collateral	A- Real estate B- (if not A) building society savings agreement/ life insurance C- (if not A/B) car or other D- Unknown / No property	Categorical
16	Installment plans	• Bank • Stores • None	Categorical
17	Housing status	• Rent • Own • For free	Categorical

18	Skill level	• Unemployed / Unskilled – non-permanent resident • Unskilled – resident • Skilled employee / Official • Executive / Self-employed / Highly qualified / Employee / Officer	Categorical
19	Telephone	• None • Yes, registered under the customer's name	Categorical
20	Foreign worker	• Yes • No	Categorical
	Label	**Description**	**Type**
21	Credibility	• Good • Bad	Categorical

TABLE 12: German Credit Data with list of numerical and categorical features and true label.

The input features listed above can be thought of as factors being used to assess repayment ability or credit risk. It easy to interpret the relevance of some of those features to the target variable. For instance, higher savings signal creditworthiness. On the contrary, it is harder to see how features such as gender, age, or foreign worker status are correlated to the target variable in the dataset.

One way to get a better understanding of the role of each feature is to calculate their influential power in the dataset through a concept commonly known as *weight of evidence* (WoE). Weight of evidence is a technique used to measure the strength of the independent variable to separate the dependent variable into two categories.

The concept has evolved from the credit scoring world as measure of separation of creditworthy and noncreditworthy applicants. Once WoE for each feature is calculated, their predictive power can be computed using a concept called *information value*. Information value reveals the *overall* strength of a variable.

Higher values indicate stronger relationship to the dependent outcome. Calculating information value can be particularly useful for the application scenario in this work as the following conventional ranges of *variable predictiveness* have been developed for the context of credit scoring (Sidiqqi, 2006):

INFORMATION VALUE	**VARIABLE PREDICTIVINESS**
Less than 0.02	Not Useful for Prediction
0.02 to 0.1	Weak Predictive Power
0.1 to 0.3	Medium Predictive Power
0.3 to 0.5	Strong Predictive Power
>0.5	Suspicious Predictive Power

TABLE 13:Benchmark for variable predictiveness used in financial industry. Source: Sidiqqi, 2006

Attribute	Information value	Interpretation based on benchmark (Sidiqqi, 2006)
Status of checking account	0.666	Suspicious predictive power
Payment	0.293	Medium predictive power
Savings	0.196	Medium predictive power
Purpose	0.166	Medium predictive power
Duration (month)	0.165	Medium predictive power
Credit amount	0.119	Medium predictive power
Collateral	0.113	Medium predictive power
Age	0.093	Weak predictive power
Foreign worker	0.087	Weak predictive power
Employment length	0.086	Weak predictive power
Housing status	0.085	Weak predictive power
Credit history	0.058	Weak predictive power
Gender	0.045	Weak predictive power
Other debtors	0.032	Weak predictive power
Installment plans	0.026	Weak predictive power
Number of credits	0.013	Not useful for prediction
Telephone	0.01	Not useful for prediction
Skill level	0.009	Not useful for prediction
Residency	0.004	Not useful for prediction
Number of people liable for	0.00004	Not useful for prediction

TABLE 14:German credit data: features ranked based on predictive power.

Finally, it is important to highlight two shortcomings of the German Credit Data:

Relatively high dimensions and small number of samples: As it can be seen from Table 12, the dataset contains 20 features and 1000 instances. However, after applying one-hot encoding in the pre-processing phase, the total number of features turned to 58. Having larger number of features for the same number of instances challenges the accuracy of the information extracted. Also, it is unclear what population this sample represents.

Lack of cogency on what true labels represent: According to the information provided by UCI Machine Learning Repository, where the dataset is available to the public, every instance's true label indicates whether he/she has actually good or bad credit. Two things remain unclear:

1- *Meaning of true labels.* What does 'bad credit' mean? Does it mean not having a credit history? Or does it mean missing a bill payment? Having a clear understanding and definition of what the true label represent is essential in supervised learning. In the earlier stages of model development, businesses should clearly identify what they intend to predict since the information they collect and the way in which they mine the data depends on the definition of the target variable. Furthermore, the *subjective* decision on how to define the target variable can be a source of bias itself. This idea is extensively explained in Chapter 4 of the thesis.
2- *Whether true labels are predictions or not.* In conventional lending practices, lenders collect relevant information on applicants and based on that they *predict* whether the applicants would default or not. If the true labels represent those human predictions, the ground truth is *not reliable*.

3.2.2 Discrimination test on the German Credit Data

At this point, let's examine the metrics that the toolkit offers for measuring bias in the German Credit Data. There are 5 metrics provided by the toolkit that can be applied to data before the cross-validation step. These metrics are listed in the table below:

Class	Metric	Computes	Results
DatasetMetric	`num_instances()`	Total number of instances. Number of instances in the privileged group. Number of instances in the unprivileged group.	All instances (N) = 1000 Males = 690 Females = 310
BinaryLabelDatasetMetric	`num_positives()`	Total number of instances with positive true label. Number of instances with positive true label in the privileged group. Number of instances with positive true label in the unprivileged group.	N =700 Males = 499 (~71%) Females = 201 (~29%)
	`num_negatives()`	Total number of instances with negative true label. Number of instances with negative true label in the privileged group. Number of instances with negative true label in the unprivileged group.	N =300 Males = 191 (~64%) Females = 109 (~36%)
	`disparate_impact()`	Calculates (Ideal value is 1): $\frac{\Pr(Y = 1 \mid g = 0)}{\Pr(Y = 1 \mid g = 1)}$ Alias: `statistical_parity_difference()`	0.89
	`Consistency ()`	Measures how similar true labels are for similar instances. Calculates (Ideal value is 1): $1 - \frac{1}{n \cdot \text{n\_neighbors}} \sum_{i=1}^{n} \lvert \hat{y}_i - \sum_{j \in \mathcal{N}_{\text{n\_neighbors}}(x_i)} \hat{y}_j \rvert$ *Note: In calculating this metric, ŷ is replaced with true label Y. See shortcoming of the dataset (above)for detailed explanation.*	Results depend on the number of neighbors (n) 0.73 (when n =3) 0.68 (when n =5) 0.66 (when n = 7) 0.64 (when n =9)

TABLE 15: AIF 360 toolkit metrics used to detect bias in the German Credit Data

The values obtained reveal important information about the dataset. Starting with `num_instances()`, this metric indicates that there is a higher density of males than females in the dataset. For statistical reasons, predictive models favor groups which are better represented in the sample because they can make their decisions with higher certainty.

`num_positives/negatives()` calculate the number of positive (negative) true labels for all instances, privileged groups, or unprivileged groups. 700/1000 instances have positive labels and 499/700 are males. This implies that the dataset is unbalanced. There are two points to note here: first, the model can predict positive classes better and, secondly, males dominate the positive class. Therefore, the model can classify males in the positive class with more certainty than women which, in itself, can lead to bias.

`disparate_impact()` calculates the proportion of individuals in each group that have positive true labels. The value 0.89 indicates that there is a higher distribution of males in the positive class than females.

`Consistency ()` is a measure of individual fairness. It is used to get an estimation on how similar individuals are classified by the *model*. This metric can also be used to test whether similar individuals have similar *true labels*. To apply this metrics, first similarity should be defined by choosing the number of neighbors (n). Note that there is no consensus on how to choose neighbors and this subjective choice can lead to bias itself. This can be observed by the change in the value of estimation when *n* is increased.

Thus far, an overview of the toolkit followed by various discrimination tests on the raw data has been provided. In the remaining part of this chapter, attention is steered towards detecting bias in a classification model that will be used to predict the instances in the German Credit Data.

3.3 Classifier

A machine learning binary classification model, e.g. logistic regression, predicts the categorical classifications for instances in the dataset. This is done by finding a decision boundary that splits the instances into two partitions (namely positive or negative classifications). In this report, it is often referred to as a *classifier*.

The classifier alone does not discriminate. Instead, the process in which it is developed can lead to unwanted bias. This idea is investigated in the following subsection of the report. Firstly, a classifier (see Section 3.3.1) that can be used to predict the instances in the German Credit Data is built. Here, possible sources of bias during the development process are identified. Secondly, the classifier is tested for group fairness using the AIF 360 toolkit (see Section 3.3.2).

3.3.1 Classifier development

There are various types of machine learning models that can be used for classification, among which are random forest and logistic regression. In practice, the choice of a model often depends on its ability to separate the instances into two classes.

To measure of how well a model is capable of separating the two classes, the Receiver Operating Characteristic curve (ROC), can be plotted. The ROC curve is a plot showing the performance or diagnostic ability of a binary classification model at *all classification thresholds*. It is obtained by plotting true positive rates (sensitivity) on the y-axis and false positive rates (1-specifity) on the x-axis in a two-dimensional coordinate system.

From the ROC curve, Area Under the Curve (AUC) curve can be computed. AUC is a single number that provides an aggregate measure of performance across all possible classification thresholds. When several algorithms are being compared, the model with the highest AUC is the one with the highest performance.

As illustrated below, three distinct algorithms are applied and the ROC curve for every one of them is plotted. The highest AUC obtained corresponds to linear regression at 0.77. As a result, linear regression will be the primary choice for classification algorithm throughout this work.

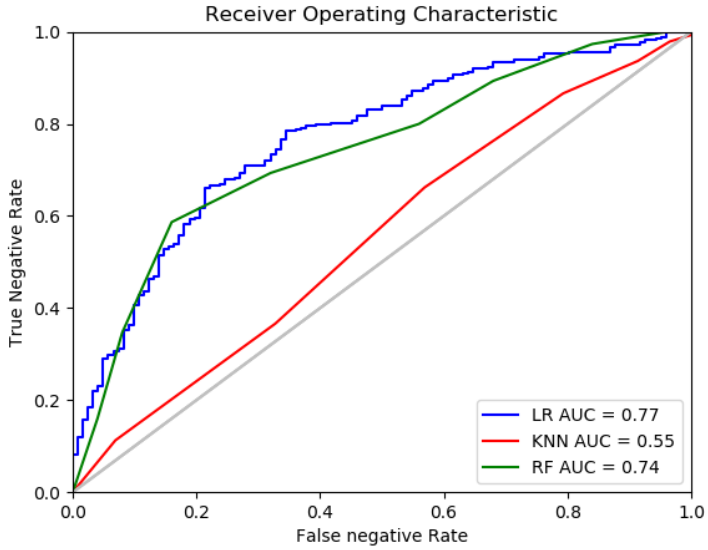


FIGURE 19:ROC curve and AUC for three binary classification algorithms: Logistic regression (blue), K-nearest neighbors (red), Random forest (green). The curve with the highest AUC is chosen as the preferred classification algorithm.

At this point, it is important to discuss the limitations of AUC as a measure of performance. There are two main points that should be highlighted:

1- AUC is more suitable for datasets that are balanced. The relatively high imbalance in the German Credit Data renders the use of AUC questionable. However, as the purpose of this work is not to maximize accuracy, but rather to investigate fairness with regard to gender, AUC is taken as a sufficient measure for model performance.

2- AUC assumes that positive and negative classes are equally important. In practice, however, this may not be the case. When there is a wide disparity between the cost of false negatives and false positives, AUC *is not a* desirable mechanism for determining performance. For instance, when a prediction model is used to detect email spam, it is necessary to prioritize minimizing false positives.

Where machine learning models are used to make key business decisions, such as in predicting the likelihood of default, achieving accuracy is an essential criterion for building classifiers. Inaccurate decisions made by classification algorithms are referred to as *misclassifications.*

When building classifiers, businesses need to have mechanisms that deal with misclassification of their observations in place. One such mechanism is to impose penalty (cost) to an incorrect prediction. In machine learning, this is referred to as a *cost* (*loss) function.*

An optimization problem seeks to minimize the cost function. In general, cost function can be thought of as a criterion in a classification problem. If the criterion is overall accuracy, an optimization problem then seeks to maximize the cost function.

Consider a scenario in which the cost of default to a bank is more than the opportunity cost of not granting a loan to a potential creditworthy candidate. In that case, banks are more likely to choose and optimize their automated decision-making models in a manner that minimizes the number *false positives errors*.

Expressed in technical terms, they choose their models based on a loss function that gives more penalties to false positives and optimize their models by minimizing false positive errors. While this choice of loss function seems logical from a business perspective, it is important to highlight that with minimizing false positives errors comes greater number of *false negatives errors*.

False negative errors pose threat to those applicants who are truly creditworthy but receive a negative (rejection) decision from the decision-making model. In their perspective, the cost of false negatives is higher than the cost of its counterpart, false positives. This implies that:

A subjective decision regarding the choice of cost function is a genuine value-judgement and this makes loan allocation algorithms essentially value-laden.

Since it is beyond the scope of this work to take a definitive stand, costs associated with errors are omitted by giving equal rank to positive and negative classifications. The cost function is selected to be *balanced accuracy* (a measure of overall accuracy more suitable for unbalanced datasets). The optimization problem then becomes choosing a classification threshold which corresponds to the highest balanced accuracy.

To start off, the dataset is split into 80/20 training and test partitions. The training set is then used to train a logistic regression model to find the predicted outcomes. Logistic regression requires choosing a classification threshold (S^*) that usually corresponds to the highest balanced accuracy or overall accuracy of the model. Given the choice of S^*, for every instance and a calculated probability score s above S^*, the model assigns a prediction value of $\hat{y} = 1$ and $\hat{y} = 0$ otherwise.

To find the optimal S^*, the classification threshold is swept from [0, 1] and values of balanced accuracy are calculated and plotted in the graph below. Balanced accuracy is the average of the correct predictions of each class separately. It is computed as follows:

$$\text{Balanced accuracy} = \frac{1}{2}\left[\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right]$$

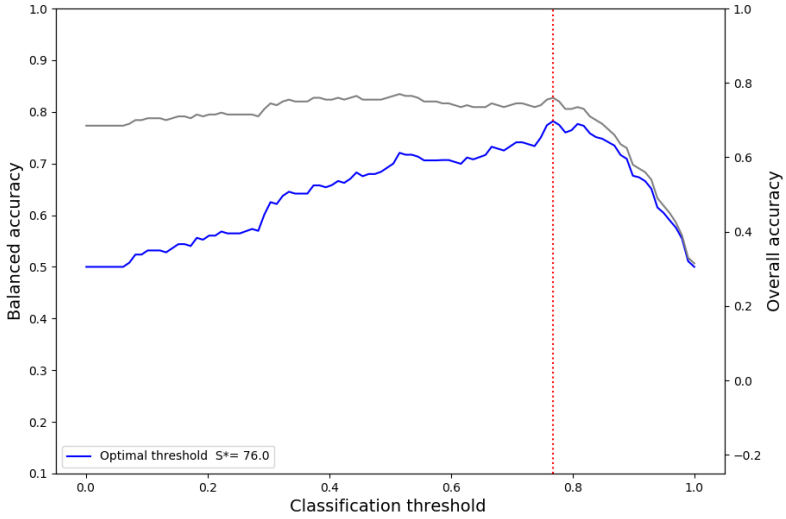


FIGURE 20:Balanced accuracy and overall accuracy versus classification threshold. Optimal threshold is S* = 0.76.

Overall accuracy as compared to balanced accuracy is more suitable for scenarios in which the training set is balanced; a scenario in which the dataset contains the same number of classifications for each (Brodersen et al., 2010).

$$\text{Overall accuracy} = \left[\frac{TP+TN}{TP+FN+TN+FP}\right]$$

For a perfectly balanced dataset, it can be mathematically proven that overall accuracy and balanced accuracy are equal. As the data becomes skewed towards either negative or positive class, balanced accuracy becomes smaller than overall accuracy: therefore, it serves as a more conservative choice for optimization.

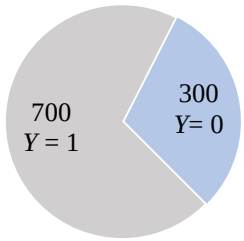


FIGURE 21:True label classification distributions.

Given the fact that the classification outcomes are unbalanced (700/300), the optimal threshold is chosen to be the one that maximizes balanced accuracy ($S^* = 0.76$). This *subjective* choice is logical in a way that for such unbalanced dataset, it is a better practice to use more conservative criteria.

Finally, the logistic regression model can be used to map the probability scores calculated for each instance to a binary classification. Any instance with score $s > S^*$ is designated as $\hat{y} = 1$ and $\hat{y} = 0$ otherwise. Now that predictions $\hat{y}$ obtained, group fairness can be examined with the AIF 360 toolkit.

There are three distinct ways in which group-level metrics check for bias:

a) Through definitions that are based distribution of predicted outcome $\hat{y}$ and demographic distribution of instances g. They can be computed from confusion matrix.

b) Through definitions that are based on distribution of predicted outcome $\hat{y}$, true label *Y*, and demographic distribution of instances *g*. They can be computed from confusion matrix.
c) Through definitions that are based on distribution of probability score *S*, true label *Y*, and demographic distribution of instances *g*.

Fairness measures that can be computed from a confusion matrix are supported (a, b). Confusion matrices for all instances (protected and unprotected groups) are provided below. The toolkit does not support metrics that are based on the distribution of probability score *S*, true label *Y*, and demographic distribution of instances *g* (item *c* on the list above). A list of commonly used group fairness metrics as proposed by Narayanan and supported by toolkit is provided below as a reference for the reader.

	Group fairness metric	Notation	Can be measured directly	Can be measured indirectly through its alias	Not included
Independence	Statistical parity	$Pr(\hat{y} = 1 \mid g = 0) = Pr(\hat{y} = 1 \mid g = 1)$	x		
	Conditional statistical parity	$Pr(\hat{y} = 1 \mid g = 0, L) = Pr(\hat{y} = 1 \mid g = 1, L)$		x	
Sufficiency	Predictive parity	$Pr(Y = 1 \mid \hat{y} = 1, g = 0) = Pr(Y = 1 \mid \hat{y} = 1, g = 1)$	x		
	Overall accuracy equality	$Pr(\hat{y} = Y, g = 0) = Pr(\hat{y} = Y, g = 1)$	x		
	conditional accuracy equality	$Pr(Y = 1 \mid \hat{y} = 1, g = 0) = Pr(Y = 1 \mid \hat{y} = 1, g = 1) \cap$ $Pr(Y = 0 \mid \hat{y} = 0, g = 0) = Pr(Y = 0 \mid \hat{y} = 0, g = 1)$		x	
	Calibration	$Pr(Y = 1 \mid S = s, g = 0) = Pr(Y = 1 \mid S = s, g = 1)$			x
	Well-calibration	$Pr(Y = 1 \mid S = s, g = 0) = Pr(Y = 1 \mid S = s, g = 1) = s$			x
Separation	False positive error rate balance	$Pr(\hat{y} = 1 \mid Y = 0, g = 0) = Pr(\hat{y} = 1 \mid Y = 0, g = 1)$	x		
	False negative error rate balance	$Pr(\hat{y} = 0 \mid Y = 1, g = 0) = Pr(\hat{y} = 0 \mid Y = 1, g = 1)$	x		
	Equalized odds	$Pr(\hat{y} = 1 \mid Y = 0, g = 0) = Pr(\hat{y} = 1 \mid Y = 0, g = 1) \cap$ $Pr(\hat{y} = 0 \mid Y = 1, g = 0) = Pr(\hat{y} = 0 \mid Y = 1, g = 1)$	x		
	Treatment equality	*(FNR/FPR)* for males = (*FNR/FPR*) for females		x	
	Balance for positive class	$E(S \mid Y = 1, g = 0) = E(S \mid Y = 1, g = 1)$			x
	Balance for negative class	$E(S \mid Y = 0, g = 0) = E(S \mid Y = 0, g = 1)$			x

As it can be seen from the table, fairness measures are grouped categorically based on their underlying

TABLE 16:Group fairness metrics proposed by Narayanan 2018. Those metrics that are based on distribution of probability score S and true labels Y are not supported.

fairness criteria. Before testing the model for group fairness, let us revisit some of the key points that were mentioned in Chapter 2:

- Independence requires the predictions to be uncorrelated with the protected attribute: $\hat{y} \perp g$.

- Separation requires the prediction outcome to be statistically independent from the protected attribute, conditional on the true label: $\hat{y} \perp g \mid Y$.
- Sufficiency requires the true label to be statistically independent from the protected attribute, conditional on the prediction outcome: $Y \perp g \mid \hat{y}$.
- Except for degenerate and improbable circumstances, they cannot be achieved simultaneously.

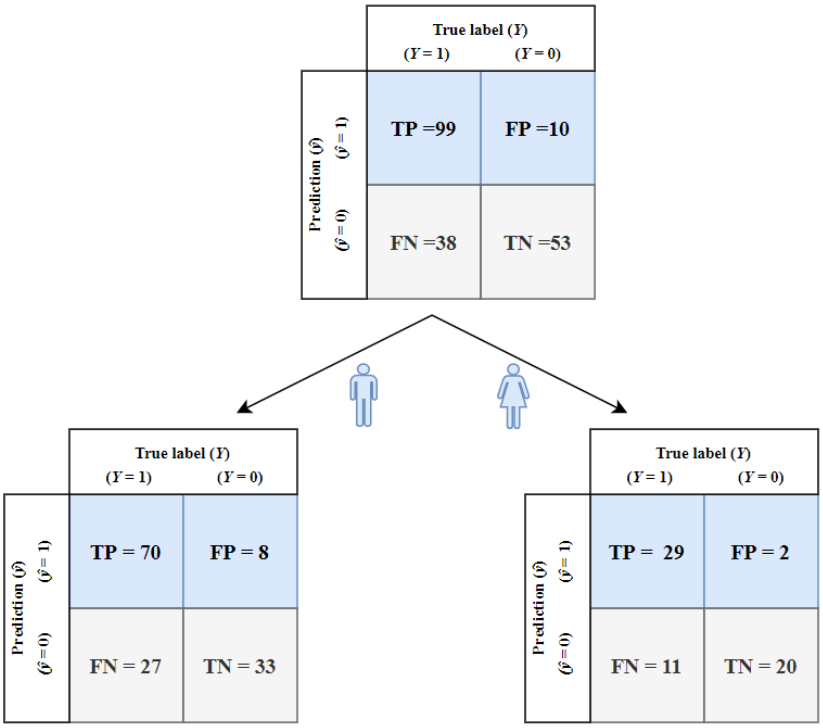


FIGURE 22:Confusion matrix for all the instances (top), for males (left), and for females (right).

3.3.2 Independence

The two metrics that evaluate independence are statistical parity and conditional statistical parity. Recall that a classifier satisfies statistical parity if individuals in both, privileged and unprivileged, groups have equal prospect of receiving a positive decision (here $\hat{y} = 1$).

Conditional statistical parity is similar to statistical parity, except that it requires the likelihood of receiving positive decisions to be conditioned on legitimate factors. To express this differently, conditional statistical parity seeks to legitimize predictions with the use features that have strong representation of applicants' creditworthiness.

Statistical parity

The AIF 360 toolkit offers two metrics that directly measure statistical parity, both of which belong to the `ClassificationMetric` class:

- `Statistical_parity_difference()`:

Computes $Pr(\hat{y} = 1 \mid g = 0) - Pr(\hat{y} = 1 \mid g = 1)$.
Smaller difference is more desirable; the ideal value is zero.
Results: - 0.0652 ~ -6.52%

Implication: It is more likely for a male applicant to have a positive predicted score.

- `Disparate_impact()`:

Computes $\frac{Pr(\hat{y} = 1 \mid g = 0)}{Pr(\hat{y} = 1 \mid g = 1)}$.

Values closer to 1 are more desirable; the ideal value is 1.

Results: 0.88

Implication: It is more likely for a male applicant to have a positive predicted score.

Conditional statistical parity

To calculate the conditional statistical parity, the first legitimate factors need to be selected from the feature set. There is no unique way of determining legitimate feature set *L*. Here, features are selected based on their overall predictive power. To recapitulate, this concept is commonly known as Information Value (IV) and is calculated from Weight of Evidence (WoE). WoE determines how well a particular feature (independent variable) separates two outcomes of a *binary* dependent variable.

There are three main reasons why selecting features based on their predictive power may be considered suitable for the application scenario at hand:

1. IV and WoE are concepts that were developed in the credit modeling and credit scoring world to separate creditworthy from not creditworthy applicants. Therefore, they are practically relevant.
2. There are conventional benchmark values for IV statistics in credit scoring that can help with feature ranking and selection. In this work, the IV table generated by Siddiqi (2006) is used as a benchmark
3. IV is an optimal feature selection technique when the classification model is binary logistic regression since IV builds strict linear relationship between the independent and dependent variables through log odds (logit function) (Smith Et al., 2002).

Some of the disadvantages of this feature selection technique are that:

1. IV is not an optimal input variable selection method when the model detects non-linear relationship between the features and the dependent variable such random forest or support vector machine (SVM).
2. Calculating WoE and IV can be computationally heavy.
3. Legitimacy is only conditioned on predictive power. This means that the protected group can still become disadvantaged if a *discriminatory feature* has a high predictive power.

In light of the information presented above, the feature selection is performed in the following *subjective* manner: Any factor that have medium and high predictive power (except for gender) will be considered as a legitimate factor.

That is, features for which their information value is greater than 0.1, are taken to be legitimate as they have medium to high ability to separate good from bad applicants in the dataset:

$$(\text{Information Value})_x > 0.1$$

$$\{\text{Status of checking acount, Credit history, Savings, Purpose, Duration , Credit amout, Collateral}\} \in L$$

Now that legitimate features are selected, a new dataset containing those features will be used to train a logistic regression model. The model will then be used to find the binary prediction values of $\hat{y}$ for each instance.

At this point, the AIF 360 toolkit can be used to calculate conditional statistical parity in two similar ways:

- `Disparate impact ():`

  Computes $\frac{Pr(\hat{y}=1 \mid g=0,\ L)}{Pr(\hat{y}=1 \mid g=1,\ L)}$

  Values closer to 1 are more desirable; the ideal value is 1.

  Results: 0.6745

  Implication: It is more likely for a male applicant to have a positive predicted score.

Bias detected: The classifier is more likely to assign the male applicants to a positive class. Does this mean the model is unfair?

The answer depends on whether *independence* is a suitable fairness criterion for the application scenario at hand. The answer also depends on whether the result obtained (~6.5% difference in likelihood) is significant enough to be labeled as unfair.

Independence is grounded in the belief that distribution of outcomes should be irrelevant and uncorrelated with certain human traits such as gender. Advocates of independence argue that achieving this fairness criterion should be a long-term societal goal for two primary reasons (Barocas et al., 2019):

1) The natural desire to live in a society in which outcomes, such as financial well-being, are statistically independent of protected attributes.
2) Independence serves as a proxy for a belief about human nature.

In the case of automated loan allocations, imposing independence as a fairness constraint results in the same rate of loans getting distributed amongst the two demographic groups defined by gender, regardless of the fact that individuals' ability to repay might be inherently different in each group. In some cases, this would result in distribution of loans to individuals with dubious repayment ability.

From a consequentialist view, such distribution pattern might be considered fair. Recall from Chapter 2 (see the Overview of Theories of Justice - Utilitarianism) that under the normative theory of utilitarianism, distributions are just as long as they maximize benefit for the greatest number of people in a society.

If a technical assessment of such loan allocations (equal rate among each group) confirms that its societal benefits outweigh the impoverishment that stems from the cost of default, then, under utilitarianism, independence is an appropriate fairness criterion.

3.3.3 Sufficiency

Recall that sufficiency allows correlation between the true label and the protected attribute to the extent that is justified by the prediction outcome. In some scenarios, allowing correlation between predictions and the sensitive attribute might be desirable if the prediction values already subsume the sensitive characteristic for the purpose of predicting the target variable. Sufficiency can also be thought of as positive and negative predictive value of the model.

Three metrics in this category can be measured either directly or indirectly using the AIF 360 toolkit.

Predictive parity

Predictive parity (precision) seeks to equalize the fraction of correct positive predictions for both genders. The AIF360 toolkit offers two metrics that can be used to measure predictive parity:

- `positive_predictive_value()`:

Computes $Pr(Y = 1 \mid \hat{y} = 1, g = 0,1)$, for each group separately.
Smaller difference is more desirable; the ideal value is zero.
Results: 0.9 for males, 0.93 for females. Difference: 0.03
Implication: Although the difference is minor, the classifier seems to have better precision for females than males.

- `false_discovery_rate()`:

Alias of positive predictive value.
Computes $Pr(Y = 0 \mid \hat{y} = 1, g = 0,1)$, for each group separately.
Smaller difference is more desirable; the ideal value is zero.
Results: 0.1 for males, 0.07 for females. Difference: -0.03
Implication: Although the difference is minor, the classifier seems to have better precision for females than males.

Conditional use accuracy

This definition of fairness, conjuncts *positive predictive parity* and *negative predictive parity*. As such, positive (negative) predictive parity can each be thought of as a relaxed version of conditional use accuracy. Note that this metric conditions fairness on the algorithm's predicted outcome and not the actual outcome. In other words, one is conditioning the probability of success (or failure) on the prediction of success (or failure).

To check whether a classifier satisfies conditional use accuracy, first, the AIF 360 toolkit should be used to measure positive predictive parity and negative parity separately. A second step is then to verify if the values are equal for males and females. Below this is illustrated.

- `positive_predictive_value()`:

Computes $Pr(Y = 1 \mid \hat{y} = 1, g = 0,1)$, for each group separately.
Smaller difference is more desirable; the ideal value is zero.
Results: 0.9 for males, 0.93 for females. Difference: 0.03
Implication: Although the difference is minor, the classifier seems to have better precision for females than males.

- `negative_predictive_value()`:

Computes $Pr(Y = 0 \mid \hat{y} = 0, g = 0,1)$, for each group separately.
Smaller difference is more desirable; the ideal value is zero.
Results: 0.55 for males, 0.65 for females. Difference: 0.1
Implication: It is more likely for females with negative predicted score to truly belong to a negative class.

- Conditional use accuracy:

Result/Implication: In combining the two conditions above, the classifier seems to perform better for the condition of positive predictive parity than negative predictive parity.

When comparing the results obtained for positive predictive parity with negative positive parity, it can be observed that the model makes more accurate decisions for the positive class. This could be the result of using an unbalanced dataset in which positive classes dominate (700/300: positive/negative).

Overall accuracy equality

This notion of fairness is achieved when overall accuracy is the same for each protected group category. The concept of overall accuracy was mentioned and used for finding classification threshold of the logistic regression model. It is important to reiterate the underlying assumption of this definition: *true positives and true negatives are equally desired.*

Adopting this notion of fairness means the application at hand makes no distinction between cost associated with true positives and true negatives. As such, this definition of fairness is not commonly used in practices where prediction for a specific class is preferred over another (Berk et al., 2017).
The AIF 360 toolkit offers a metric that directly measures overall accuracy.

- `accuracy ():`

Computes $Pr(\hat{y} = Y \mid g = 0, 1)$ for each group separately.
Smaller difference is more desirable; the ideal value is zero.
Results: 0.75 for males, 0.79 for females. Difference: 0.04
Implication: Although the difference is small, the classifier has a higher overall accuracy for females than males.

Bias detected: the classifier makes more accurate decisions (particularly in negative class) for females as compared to males in both classes. Does this mean the model is unfair?

No conclusion can be made about the model's fairness, unless, as the first condition, it can be verified that *sufficiency* is a suitable fairness criterion for the application at hand to begin with. To do this, sufficiency can be viewed through the lens of equality of opportunity. This approach helps with understanding the moral assumptions that underlie sufficiency as a fairness criterion.

Under the definition of equality of opportunity provided in Chapter 2, a decision-making system is fair if it provides the same statistical prospects of receiving harm/benefit to individuals who have the same effort-based utility, irrespective of their irrelevant circumstances, such as gender or race. In a similar manner, sufficiency seeks to equalize the prospect of actually having a good/bad credit (Y, U) for individuals who are assigned the same prediction ($\hat{y}$, e), irrespective of their special circumstance (g, c).

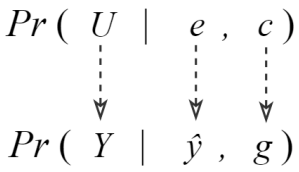


FIGURE 23:Conceptual mapping of sufficiency under the substantive equality of opportunity.

Finally, in imposing sufficiency as a fairness criterion on the classifier built for predicting instances in the German Credit Data, it can be concluded that harm/benefit (true label) is grounded in applicants' true credibility and effort-based utility (predictions) which together reflect *merit* of applicants. In other words, what justifies inequality is applicants' merit (merit comes from having higher savings, higher income, etc.). After validating such moral equivalency assumption, sufficiency can then be taken as an appropriate fairness criterion.

Let us assume for the sake of the argument that *sufficiency* is a valid fairness criterion, what does obtaining higher accuracy for females actually imply?

To answer the question, we can calculate $Pr\ (Y = 1 \mid \hat{y} = 0, g)$ from $Pr\ (Y = 0 \mid \hat{y} = 0, g)$:

Males: $Pr\ (Y = 0 \mid \hat{y} = 0, m) = 0.55$ → $Pr\ (Y = 1 \mid \hat{y} = 0, m) = \mathbf{0.45}$
Females: $Pr\ (Y = 0 \mid \hat{y} = 0, f) = 0.65$ → $Pr\ (Y = 1 \mid \hat{y} = 0, f) = \mathbf{0.35}$

It can be observed that it is more likely for males than females with a negative prediction to actually have good credit.

When inequality is justified by predictions, false distribution of utility (benefit) advantages women more than men (or disadvantaging women less than men). If the difference is large enough, then the model is deemed unfair!

3.3.4 Separation

Separation allows correlation between the prediction and the protected attribute to the extent that it is justified by the target variable. This fairness criterion acknowledges that, in some circumstances, the sensitive attribute may be correlated with the target variable. For instance, a bank might argue that it is a matter of business necessity to give out unequal lending rates among the two groups if one group truly has a higher default rate compared to another.

The AIF 360 toolkit can be used to directly/indirectly measure the following four out of six metrics belonging to this subcategory of group fairness.

False positive error rate balance

When the fairness is defined as equalizing the prospect of receiving false positives among the protected groups, false positive error rate balance should be used. Arguably, false positive errors are least desired by financial institutions if losses of revenue due to defaults are high. However, to make a definitive statement about lending institutions' preferences, the trade-off should be made explicit by a cost/benefit

analysis of false positives and false negatives and, thereby, choosing a threshold that optimizes some criteria based on costs and benefits (Khandani et al., 2010).

False positive error rates balance can be evaluated with the AIF 360 toolkit using the following metrics:

- `false_positive_rate_difference()`:

Computes $Pr(\hat{y} = 1 \mid Y = 0, g = 0) - Pr(\hat{y} = 1 \mid Y = 0, g = 1)$
Smaller difference is more desirable; the ideal value is zero.
Results: -0.1042 ~ -10.4%
Implication: It is more likely for the classifier to assign a positive score to males who actually have a bad credit score.

- `true_negative_rate ()`:

Alias of false positive rate
Computes $Pr(\hat{y} = 0 \mid Y = 0, g = 0,1)$ for each group separately.
Smaller difference is more desirable; the ideal value is zero.
Results: 0.80 for males, 0.91 for females.
Implication: It is more likely for the classifier to assign a negative score to females who actually have a bad credit score.

False negative error rate balance

When the fairness is defined as equalizing the prospect of receiving false negatives among the protected groups, false negative error rate balance should be used. One can argue that false negatives are the least desired outcome from the perspective of candidates who are truly creditworthy. Some reasons that strengthen such arguments are that, firstly, if a candidate is rejected, he/she should wait for a period before applying again at the same financial institution. Secondly, in most cases, a rejected attempt shows up on the credit bureau report, which can negatively affect the chances of getting approved for a loan at another financial institutions. Last but not least, candidates are always encouraged to seek explanations for the adverse decisions they receive, but this may not be such an easy task in the case of automated decision-making, because interpreting and explaining causes of those decisions is very complicated (Miller, 2017).

False negative error rates balance can be evaluated with the AIF 360 toolkit using the following metric:

- `false_negative_rate_difference()`:

Computes $Pr(\hat{y} = 0 \mid Y = 1, g = 0) - Pr(\hat{y} = 0 \mid Y = 1, g = 1)$
Smaller difference is more desirable; the ideal value is zero.
Results: -0.01
Implication: Although the difference is very small, the classifier is more likely to assign a negative prediction to females are truly classified as positive.

Equalized odds

Equalized odds seeks to equalize false positive errors and false negative errors for males and females. This notion of fairness requires two conditions: 1) Applicants with actual good credit should obtain an

equal chance of receiving a positive score, regardless of their gender and 2) applicants with actual bad credit should obtain equal chance of receiving a positive score, regardless of their gender. Relaxed forms of equalized odds are false positive rate balance and false negative rate balance. As mentioned earlier, often a choice has to be made between one of the relaxed forms.

The AIF 360 toolkit can be used to directly measure equalized odds.

- `average_odds_difference()`:

  Computes $\frac{1}{2}[Pr(\hat{y} = 1 \mid Y = 0, g = 0) - Pr(\hat{y} = 1 \mid Y = 0, g = 1)] + [Pr(\hat{y} = 1 \mid Y = 1, g = 0) - Pr(\hat{y} = 1 \mid Y = 1, g = 1)]$

  A value of zero indicates equality of odds.

  Results: -0.057

  Implication: The classifier is more likely to assign a good score to males who have an actual bad credit score.

Treatment equality

This notion of fairness is achieved when the ratio of false negatives and false positives is equal for both males and females. The term 'treatment' conveys that such ratios can be a policy lever with which other kinds of fairness could be achieved (Berk et al., 2017). For instance, if positives are less costly for males than females such that equalized odds is achieved, then males and females are being treated differently by the classification algorithm.

With the toolkit, treatment equality can be measured indirectly by taking the ratio of false positives and false negatives for both protected groups.

- `Treatment equality`:

  Computes (*FNR/FPR*) for males - (*FNR/FPR*) for females

  Smaller difference is more desirable; the ideal value is zero.

  Results: -1.60

  Implication: The classifier fails to satisfy treatment equality.

Bias detected: the classifier is more likely to assign a positive score to male applicants who actually have a bad credit. Does this mean the model is unfair?

No conclusion can be made about the model's fairness unless, as the first condition, it can be verified that separation is a suitable fairness criterion for the application at hand to begin with. This is done is a similar manner in which separation is viewed through the lens of equality of opportunity. This approach helps with understanding the moral assumptions that underlie this fairness criterion.

Under the provided definition of equality of opportunity in Chapter 2, a decision-making system is fair if it provides the same statistical prospects of receiving harm/benefit to individuals who have the same effort-based utility, irrespective of their irrelevant circumstances such as gender or race. In a similar manner, separation seeks to equalize the prospects of receiving the same predictions ($\hat{y}$, U) credit for individuals who have the same true labels (Y, e), irrespective of their special circumstance (g, c).

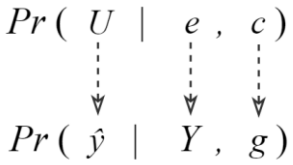


FIGURE 24:Conceptual mapping of separation under substantive equality of opportunity.

In imposing separation as a fairness criterion, it can be concluded that harm/benefit (predictions) is grounded in the algorithmic decisions made about the applicants, effort-based utility (true labels) reflects *merit* are what justify inequality. After validating such moral equivalency assumption, separation can then be taken as an appropriate fairness criterion.

Let us assume for the sake of the argument that *separation* is a valid fairness criterion for this application scenario, it can be observed that:

Males: $Pr\,(\hat{y} = 1 \mid Y = 0, m)$ = **0.2**
Females: $Pr\,(\hat{y} = 1 \mid Y = 0, f)$ = **0.09**

In contrast to sufficiency, when inequality is justified by the true labels, false distribution of utility (benefit) advantages men more than women. If the difference is large enough, then the model is deemed unfair!

3.4 Chapter summary

In this chapter, the second research question is investigated:

> *How can the metrics be applied to observe gender bias in lending history data?*

In investigating this research question, an overview of the toolkit is provided as a preliminary step. As illustrated in Figure 18, it is found that the toolkit can be used to detect bias at two different points in a machine learning pipeline: on the dataset and on the output of the classifier.

The AIF 360 toolkit is then used to investigate whether a logistic regression model built to classify loan applicants in the German Credit Data meets the definitions of group fairness proposed by Narayanan (2017). The toolkit does not offer definitions that are based on joint distribution of probability scores, true labels, and gender (*S, Y, g*). On the other hand, metrics that are based on confusion matrices can be computed. A table that summarizes the results is provided below.

	Fairness definition *Group fairness metrics proposed by Naraynan supported by the AIF360.*	**Results** *Expressed as difference in likelihood (female minus male), except for overall accuracy and treatment equality.*	**Implications** *Absolute value of difference:*
			|Δ| = 5%	|Δ| = 10%
Independence	Statistical parity	-6.5%	x	✓
	Conditional statistical parity	-23.3%	x	x
Sufficiency	Positive predictive parity	+3%	✓	✓
	Negative predictive parity	+10%	x	✓
	Conditional accuracy equality	Joint condition of the above two	x	✓
	Overall accuracy equality	+4%	✓	✓
Separation	False positive error rate balance	-10.4%	x	x
	False negative error rate balance	-1%	✓	✓
	Equalized odds	-4%	✓	✓
	Treatment equality	-1.6	-	-

TABLE 17: Summary of results obtained after testing the classifier for group fairness based on three statistical fairness criteria (independence, sufficiency, and separation). Negative sign indicates that the classifier favors males.

Without taking a stand on which of the fairness criteria and therefore, which metrics are suitable for assessing fairness with respect to gender for the context of consumer lending, two key points should be highlighted:

Firstly, the results obtained indicate that the classifier performs better at predicting instances in the positive class as compared to the ones in the negative class. One possible reason is that the dataset holds an unbalanced distribution of its binary classes. The dominance of positive true labels makes the model better at predicting positive instances, especially, if the loss function equally penalizes the false positive and false negative errors made.

Secondly, another key finding of this chapter is that, for the applicants in the German Credit Data, the choice of separation and sufficiency can have different repercussions for each demographic group. When false distribution of utility (benefit) is under inspection, sufficiency advantages females more than males while separation advantages males more than females. Such discordance highlights the importance of realizing how the relevant distribution of harm/benefit depends on the choices made by decision makers.

In closing this chapter, it is important to stress that the *underlying cause* of biases, shown in the table above, can be numerous factors which are discussed in the following discussion chapter of this work.

This page is intentionally left blank.

4

Discussion & Managerial implications

4.0 Introduction

Due to the complex nature of supervised machine learning processes, addressing unwanted bias amid its numerous root causes is not a straightforward task. As it will be explained in this chapter, there are multiple iterative stages involved in developing and deploying machine learning models, the majority of which entails making *subjective decisions* by the managers and data stewards such as how to define business objectives and how to collect data.

In continuation of the case study in Chapter 3, this part of the research investigates, using a taxonomy of causes of bias, how those subjective decisions and common practices can become a source of gender-related bias and, in tandem, describes some of the mechanisms and practices in place to help them address discrimination in their datasets and their machine learning models.

The structure of the chapter is as follows: Section 4.1 explains the stages of ML process. In Section 4.2, a taxonomy of bias is used to examine where in the process bias could be encountered. Section 4.3 provides some recommendations on how to mitigate bias. Section 4.4 provides some recommendations on how to use the AIF360 toolkit in practice.

4.1 Model development and deployment process

Bias can creep into the automated decision-making system at various stages. To illustrate the stages involved in a data mining process, the Cross Industry Standard Process for Data Mining (CRISP-DM) can be used. This standard divides the process of developing and deploying machine learning classification models into 6 different stages/tasks that are connected in an iterative process.

Before explaining each stage in the process, it is important to mention a shortcoming of CRSIP-DM: *the model is too generic*. Due to the fact that this model was developed to explain any data mining process without considering its application domain and the data mining problem type, it is hard to determine how, in practice, this standard can be converted into an application specific model.

Nonetheless, using the CRISP-DM model can be very useful for this work since all steps taken in developing the classifier (data-preparation, modeling, and evaluation) in Chapter 3 are embedded in this model.

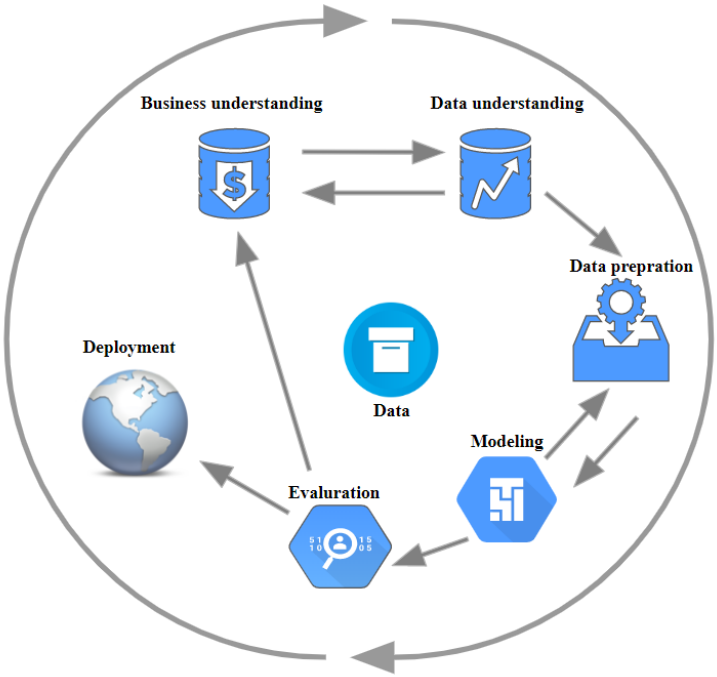


FIGURE 25:Process diagram for the CRISP-DM model. Adopted from Chapman et al. (2000).

Starting with *business understanding*, every stage in the CRISP-Model is explained below (Chapman et al., 2000)*:*

***Stage 1**: Business understanding*

This initial stage aims at identifying the objectives and criteria of the project from a business perspective, using the information to clearly define data mining issues and then form a concrete plan to achieve the stipulated objectives.

Stage 2: *Data understanding*

This phase begins with an initial data collection and then continues with data familiarization to gather insights, pinpoint data quality concerns, or identify interesting subsets to make hypotheses or hidden details.

Stage 3: *Data preparation*

The third stage includes all activities that will allow for curating the initial raw data into the final data set, which will be later used as an input to the modelling tool. These activities may include the *selection of tables, records, and attributes*, as well as transforming and organizing data for further modelling.

Stage 4: *Modelling*

This stage consists of choosing and applying diverse modelling methods and optimization of their parameters to optimum values. As several methods may be suitable for the same data mining problem, it is often necessary to return to the data preparation phase and understand specific requirements for the data form.

Stage 5: *Evaluation*

By now, an apparently high-quality model should have been created. However, in advance of its deployment, it is inevitable to reassess all the steps taken into building the model, including testing for its reliability at achieving the objectives of the business. The main goal is reviewing whether there are any significant problems that have not been adequately addressed. If not, a conclusion regarding the deployment the model should be made at the end of this phase.

Stage 6*: Deployment*

Model creation, typically, does not signify the end of the project. Even if the model's aim is to increase awareness of the data, the knowledge gathered will need to be structured in a way that can be easily used by the client. The deployment, if carried out by an analyst, may be as easy as producing a report or as complex as implementing an ongoing data mining process. If the deployment is carried out by the client himself, it is still necessary, however, that the client understands in advance exactly how to operate the model efficiently.

What makes bias so difficult to detect is the fact that it can creep into the system through each of the 6 stages of the CRISP-DM process (d'Alessandro et al., 2017). These stages can be thought of as 'where' bias may be encountered.

4.2 Taxonomy of bias

Attention can now be steered towards understanding the root causes of bias. To systematically investigate this, let's take a step back and revisit the two components of a classification process: *development* and *deployment*. Root causes of discrimination in each component differ as explained below:

Model development

The root cause of discrimination in the model (classifier) *development* can be divided into two categories:

1- Misspecification
2- Training data

If the training data has no existing bias (as it will be explained later, no prejudice and perfect representation of protected groups), discrimination can happen through misspecification of instances in the data. This can happen, for instance, by using features that discriminate against a protected class membership or by the subjective way in which a target value is defined.

Nevertheless, in most settings bias originates from the *input data* and it is the first essential place where stakeholders should probe when designing automated decision-making systems. In essence, bias in the input data is a human problem, which can be amplified by a model that imperfectly locates statistical relationships in that dataset. It can be concluded that the process of *building* a classification model free of bias if:

1) We can assume a large enough training data that is a prefect representation of its protected groups and does not reflect historical inequalities stemming from existence of systematic animosity against those groups.
2) We can assume an optimal classifier that reaches peak theoretical performance guarantees and perfectly discovers patterns and regularities in the training data.

Model deployment

Even if the model built is free of bias, the way in which it is deployed could lead to discrimination. Therefore, stakeholders need to remain attentive throughout the entire evaluation and deployment process. The primary root cause of bias in the deployment phase stems from failure to audit and monitor the machine learning model.

Now that the root causes of discrimination in the machine learning process have been explained, an extensive set of mechanisms in which discrimination can happen is presented in the taxonomy below. Using this taxonomy can expand the horizon of the reader on where and how bias can affect the system.

Component	Root cause	Layer 1	Layer 2
Classifier Development	Misspecification	Target variable	Heterogenous target variable (1,2,3,4)
			Proxy target variable learning (1,2,3,4)
			Target variable subjectivity (1)
		Feature	Inclusion of protected attributes (3,4)
			Inclusion of protected attribute proxies (3,4)
		Cost function	Omitted discrimination penalties (4)
			Failure to specify asymmetric error costs (4,5)
	Data	Prejudice in data (1,2,3,4)	
		Sample bias	Under-representation (2,3,4)
			Over-representation (2,3,4)

TABLE 18: Taxonomy of causes of bias in the life cycle of machine learning process. Numbers indicate the stages (where) in the process the bias listed in the second layer can be encountered. *Source: adapted from d'Alessandro et al. (2017).*

1: Business Understanding 2: Data Understanding 3: Data Preparation. 4: Modeling

4.2.1 Misspecification

A statistical model is only an approximation to the truth. For instance, if a linear function is used to predict logarithmic relationships between dependent and independent variables, misspecification occurs. In machine learning, misspecification due to functional approximations is not an issue since sophisticated models, such as Deep Learning and Random Forest, can be applied to estimate complex patterns in the dataset (d'Alessandro et al 2017). Discrimination issues that are of concern here arise from misspecification of the three *underlying components* used for supervised machine learning:

- Target variables
- Features
- Loss function

Target variables

Target variable subjectivity: Target variables are what classification models aim to predict. This requires the stakeholders to have a clear and formulated definition of the target variables. Bias may happen due to the subjective decision in which the target variable is defined. In explaining this idea, Barocas and Selbst compare a machine learning model designed to detect spam emails to another model that predicts credit worthiness of loan applicants (Barocas & Selbst, 2016). Defining and measuring credit worthiness is a much more subjective tasks than defining spam or fraud. Let's expand on this example provided by Barocas and Selbst to get a better understanding of how bias due to target variable selection can happen in consumer lending.

Creditworthiness has no clear definition. As such, it is artifact of the problem itself (Barocas & Selbst, 2016). It can be defined, for instance, as having good saving habits in which lenders would mine their data to measure their applicants' net savings, or it can be defined as having stable jobs, in which they would extract information on how long the applicants remained in their current positions. There are far more logical ways to define creditworthiness and its definition itself will, ultimately, determine the information that should be extracted from the data.

How can the choice of target variable be a cause bias? The answer is that is the choice of true label leads to some distribution between its classes and the members of a protected membership which, ultimately, causes a group to be systematically disadvantaged. For instance, consider a scenario in which females in the training dataset have on average higher net savings, better credit history, but lower income than male applicants. In a case where creditworthiness is defined as having high income, it is more likely that a higher distribution of rejected female applicants be present in the training data. This, in turn, causes the females to be ranked as less favorable applicants even if they have higher average savings or better credit history. In contrast, if creditworthiness was defined as having a better credit history, male applicants would be disadvantaged in the selection process.

Where can it be encountered? Bias due to target variable subjectivity appears during the business understanding phase. As mentioned earlier, there is no consensus on how to define credit worthiness and this flexibility can play to the advantage of the financial institution as the definition can be taken in a way to maximize utility.

Lessons learned from the case study: In the case study, it was found that one of the limitations of the German Credit Data is having an ambiguous definition of the true labels (good/bad credit). Even though it cannot be verified, this is a possible source of bias since the ground truth behind the true label classes stems from an arbitrary person's definition of good/bad applicant.

Having knowledge of the ground truth in how creditworthiness (true labels) is defined in a prelabeled training dataset is essential for evaluating whether the choice of labels systematically disadvantages female applicants.

Finally, bias due to target variable subjectivity cannot be directly tested for with the AIF 360 toolkit in the business understanding stage as it depends on the choices made by outside users. In the data preparation stage, however, users are encouraged to use the toolkit to investigate the distribution of true label binary classes for each protected and unprotected group because a large disparity in the distribution signals correlations between the selected target variable and the class membership.

Target variable proxy: For statistical reasons, higher occurrence of an event in a dataset leads to higher ability in predicting that event. In scenarios where information on the event getting predicted is sparse, stakeholders may decide to predict a label that acts as a proxy for the initially intended true label (d'Alessandro et al., 2017).

For instance, a bank that uses predictive tools to classify risky applicants is better off by using datasets that have higher density of risky applicants. If this is not possible, the bank can predict for other aspects, such as having a low income, as a proxy for riskiness. Again, it can be seen that the choice of target variable (here, substituted target variable) may adversely impact female applicants.

Bias arising from the use of target variable proxies can leap into the deployment process at four possible stages: business understanding, data understanding, data preparation, and modeling. At every of these stages, data users and managers may change the intended target variable for a proxy if they believe that their data is negatively unbalanced in representing the outcome they wish to predict.

Heterogenous target variable: In a similar case where the primary target to be predicted is believed to be scarce in the dataset, other target variables can be grouped together with the intended label in order to obtain higher accuracy.

Suppose that a lender is given the task of predicting applicants who are defined to be of high risk. Suppose now that a risky applicant is someone who has missed 5 or more payment obligations in the past. If the occurrence of applicants who have missed 5 or more payments is very low, accuracy in predicting risky applicants can be increased by relabeling, for instance, applicants who missed 3 or more payments as high risk.

The practice of augmenting the density of high-risk applicants in the training data may lead to bias. Suppose that in the example above, initially 150/1000 applicants are labeled as high-risk, out of which 100 are male. After the augmentation practice, the total number of high-risk applicants becomes 400/1000, but this time, a total of 150 are male. This means that 200 females who had missed 3 or more payments are now added to the pool of high-risk applicants.

Similar to biases stemming from proxy target variables, biases arising due to heterogenous target variables can leap into the system stages: business understanding, data understanding, data preparation, and modeling. At every of these stages, data users may redefine the target variable to boost the number intended events they wish to predict.

Finally, the AIF 360 toolkit does not offer a direct way to detect biases stemming from heterogenous target variable selection, however, the external user is recommended to apply the toolkit to test for the correlation of the *newly defined* target variable with the protected attributes.

Feature

European laws (Article 22 Paragraph 4, European Union, Parliament and Council) prohibit direct use of sensitive features in the data mining process. Recall from Chapter 2 that there are two ways in which sensitive features can be interpreted. The first interpretation relates to direct use of protected attributes.

Inclusion of protected attributes: It is generally believed that bias can leap into the machine learning process if the feature set used to train the model includes the protected attribute, such as gender. This belief is debatable for two reasons:

Firstly, from a practical standpoint, it may be advisable to collect and keep information about the data subjects' protected group membership within the scope of governing data protection laws. Such information can be used later to assess fairness of the classification model deployed with respect to the protected attribute. Furthermore, by keeping features in the dataset, correlation studies between sensitive and non-sensitive features becomes an efficient task.

Secondly, given a large enough dataset containing relevant and abundant information about its data subjects, direct use of protected attribute should be, *in theory*, irrelevant to the outcome in terms of discriminatory effect (Barocas & Selbst, 2016). If there is a high dependency between the protected class and the true label, it is safe to assume that the information about the trait is redundantly encoded in the data through other features, either through causal relationships or high correlations. In this case, removing the protected attribute should not have any influence on the outcome.

Recalling from Chapter 2, if financial institutions make their algorithms blind to the feature representing gender, they adhere to *formal equality of opportunity*, which restricts *direct discrimination.* However, managers should be wary of the fact that exclusion of the feature gender alone might not remove hidden (indirect) discrimination as this practice is merely formal and does not consider differences between men and women which might be present in their datasets.

When users wish to exclude the sensitive feature from their dataset, they can do so in the data preparation and modeling stages of the classifier deployment process (Stages 3 and 4).

Inclusion of protected attribute proxies: The second interpretation follows that sensitive features are not only the ones whose use formally discriminates, but also those that are correlated with the protected attribute. For instance, a bank that makes its algorithm blind to gender, but uses employment length to evaluate its applicants' credit worthiness, may indirectly discriminate against women if employment length is a proxy for gender in their dataset. Why was employment length used in this example?

In conceptualizing fairness from perspective of women, some of the key differences between men and women is identified (Table 4). One of those differences relates to disproportionate time women spend

on domestic affairs as compared to men. This implies that men tend to have more stable career prospect as compared to women. Going back to the example above, using employment length to determine credit worthiness might have an adverse impact (indirect discrimination) on female applicants if those in the sample are reflective of the difference mentioned.

EU laws governing consumer lending prohibit both direct and indirect discrimination. Substantive equality of opportunity takes the differences between applicants into consideration and thus, it is better suited for addressing indirect discrimination. In practice, addressing indirect discrimination is a much more difficult task when large datasets are used since correlation between features becomes increasingly complex.

According to the taxonomy of bias in Table 18, bias due to inclusion of proxy variables can be encountered during the data preparation and model training phase. Using appropriate experimental designs, managers and data stewards should check for causal relationships between variables during these stages to detect proxy relationships between the features. If this cannot be achieved, they can get a sense of how features relate to each other and relate to the predictions by calculating correlations between features.

The AIF 360 toolkit does not offer correlation or causal measures. To investigate relationships between variables, IBM's AIX 360 is a more suitable toolkit as it supports an extensive set of local and global explanation algorithms.

There is a unique way in which presence of a proxy variable(s) for gender can be checked for using the AIF360 toolkit: Train the model without the feature representing gender. If the predicted outcomes are not identical for two identical individuals that only differ by gender, then a proxy exists in the feature set. However, nothing can be said with certainty about which feature(s) serves as proxy for gender.

Cost (loss) function

Failure to specify asymmetry costs: A cost function is a measure of the performance of a classification model. Bias can creep into the system if asymmetry of misclassification costs is ignored. For instance, if false negative errors critically impact the individuals in a protected class, the classifier should consider the asymmetry cost of misspecification associated with false negatives.

In practice, managers and data scientists may face controversies when choosing cost functions since they need to make tradeoffs between false positives and false negative errors. This idea was briefly mentioned in Section 3.3.1, but continues to be elaborated on here.

In Chapter 3, cost of errors was ignored which can be a cause of bias. To see how, let's revisit the distribution of the positive and negative classes in the German Credit Data (700/300). If a classifier makes 1000 predictions with 100 errors (50 false positives and 50 false negatives), the negative class is getting penalized more than the positive class. For such unbalanced data, often a more severe penalty is assigned to the model for making false negative errors to reduce chances of bias.

In addressing credit risk, the general belief is that it is more important to avoid false positives than false negatives (Khandani et al., 2010). Without taking a stand on whether this belief should be considered as the prevailing norm in the financial industry, a supporting argument that tallies with consequentialist intuitions is provided here: Financial institutions are necessary for smooth functioning and the overall

health of an economy. Due to interdependencies between financial institutions, there exists a system-wide risk of failure if one of the member institutions fails (Schweser, 2018). Therefore, financial institutions should not only minimize risk for the benefit of their businesses, but also for the overall well-being of society.

Error by omission: This idea is very similar to the one above except that it requires the cost function to incorporate an *additional* penalty for discrimination. Said differently, fairness here can be thought of as an additional criterion that attaches itself to the cost function. The fairness criterion ensures that during the learning phase, the algorithm assigns similar classifications to similar individuals. As it will be explained in the next section, prejudice is a term that describes a scenario in which two individuals that are similar in every aspect except for a sensitive trait, such as gender, receive different treatments.

Finally, the AIF360 toolkit offers a bias mitigation technique (see Table 19) called *Prejudice remover* which considers the differences in how the learning algorithm classifies the protected and non-protected classes. Penalization is determined by the extent of the difference.

4.2.2 Training data

Prejudice in data

Data reflects the decisions made about humans in the past and often, prejudice and animosity are deeply rooted in past decisions against certain groups. Algorithms can learn prejudice in the dataset which they are trained on and further reinforce humans' prejudicial thinking. Prejudice and wrongful discrimination against women in consumer lending have a long history. The reader is encouraged to refer to the work of Bowdish as this work provides an extensive history of how discrimination affected women in consumer lending and, more importantly, how it has transformed over the past four decades (Bowdish, 2010).

Prejudice in the dataset can be encountered at any stages that involves planning, choosing, preparing, and using the dataset (Stages 1-4). Therefore, it is highly recommended that data scientists and managers make effort to access more even-handed and diverse datasets. Prejudice should be tested for early on during the classifier development stage.

The process in which class labels are manually encoded is a critical point where prejudice can present itself in the dataset. Here, the judgement of the person who classifies the instances is being taken as the ground truth and, ultimately, into a formalized rule that would systematically change the prospects of future applicants.

Although it is almost impossible to demonstrate with certainty that prejudice is present in the raw data, it can be investigated for by calculating the correlations between the members in a protected class and the classes of true labels (Kamishima et al., 2012).

Another method to check for prejudice is to apply the toolkit's consistency metric to the raw data. This metric gives out a single number that can be used to evaluate the similarity of the true labels for the similar individuals. For the German Credit Data, the consistency metric indicated that there is some disparity between individuals' treatment, although, this has to be verified since the choice in which

similarity is defined can lead to bias itself. Finally, the toolkit offers an in-processing algorithm, prejudice remover, that can be used to reduce the dependencies between the two variables.

Sample bias

Underrepresentation and overrepresentation. Bias can arise when members in a protected class are disproportionately underrepresented in the dataset. By treating the underrepresented group as noise, predictive models revert to average trends to avoid overfitting. Bias due to underrepresentation is particularly an issue in consumer lending as women tend to suffer more than men from lack of financial inclusion (Table 4) (Holloway, 2017).

Overrepresentation of a particular group in a dataset can also lead too bias as the learning models may skew the decision toward a particular result. For instance, including disproportionately more men in the dataset augments their likelihood of receiving erroneous decisions by the predictive model.

An important distinction should be made here - lack of representation (or over representation) in the data is not always a source of bias. Take the German Credit Data as an example, where ratio of male/females is (610/390) and clearly males dominate the dataset, however, one should ask whether the population it represents has a *similar ratio* of male/female or not. If that is the case, it may be nothing more than a population that is naturally skewed towards men.

In conclusion, when dealing with sample bias, managers should look beyond the distribution of demographic groups in their datasets and investigate how the sample was collected and obtain details about the questions asked, time and location of collection and, more importantly, about the population which their sample data represents. These factors interact with social norms which can result in unintended bias.

4.3 Bias mitigation

In this section, some of the most recent techniques which can be used to mitigate bias are discussed. Broadly, bias mitigation methods fall under two distinct categories: *technical approaches* and *discursive strategies* (Rovatsos et al., 2019). One of the latest developments in technical approaches to bias mitigation is IBM's AIF 360 toolkit. As the main objective of this section is to provide recommendation on how to mitigate using the AIF360 toolkit, emphasis is put on the overview of technical approaches.

Technical approaches relate to use of statistical and software-based techniques and tools such as the AIF 360 toolkit. Generally, statistical methods can mitigate bias at three different points in the machine learning pipeline: raw data, learning procedure, and output of the mode.

Discursive strategies constitute approaches that involve life-long governance and interaction with stakeholders who are affected by algorithmic decision-making systems. Discussion forums and workshops are some of the practices that fall under this category. Broadly speaking, there are three tasks that are essential in implementing discursive strategies in practice.

The first involves translating the existing research into the organizational context. The second task entails developing processes that can be integrated into the existing product lifecycles. Finally, the third task requires engaging with external communities and experts to share experiences and knowledge to stay up to date with the most recent solutions and practices (Carmer et al., 2018).

Some of the key strengths and weaknesses of each approach are listed below.

Technical approaches strengths:

- In principle, technical approaches are considered to be more consistent and efficient.
- Technical approaches can be used to intervene at a single point in the machine learning life cycle.
- They are less costly to implement than discursive approaches.

Technical approaches weaknesses:

- They require precise instructions since they cannot navigate moral grey areas in the same way that humans can.
- They require a clear and formulated definition of fairness.
- They are too complex and hard to implement by a layperson.

Discursive strategies strengths:

- They provide an interdisciplinary and collective approach to mitigating bias.
- They are better suited for navigating moral grey areas where mitigation algorithms struggle.
- These approaches are generally better understood by different stakeholders.

Discursive strategies weaknesses:

- They are more costly and inconsistent.
- They are effective as long as they are given serious consideration throughout the entire lifecycle of an algorithm.
- Results cannot be obtained immediately.

For more information on discursive strategies, the reader is encouraged to refer to the following recent works: Rovatsos et al. provide overview of several algorithmic assessment tools and frameworks such as a commonly used assessment framework offered by US-based AI Now Institute (Rovatsos et al., 2019). Carmer et al. investigate how algorithmic bias can be addressed in practice and discuss organizational challenges to implementing mitigating solutions. They also provide a list of questions that teams can refer to when they collect data, choose decision-making models, and assess the outcomes. Finally, Resiman et al. provide an extensive analysis of Algorithmic Impact Assessments (AIA), a framework that combines assessment tools that are already being implanted in other domains such as environment protection and human right policy (Resiman et al., 2018).

Next, technical approaches in removing bias from machine learning systems are described.

Technical approaches to bias mitigation

Recent years have witnessed the development of a multitude of statistical approaches to mitigating bias in decision-making systems. As it was mentioned earlier, one of the advantages of using statistical

approaches is that unlike discursive strategies, they can target specific points (raw data, algorithm, and prediction outcomes) in the machine learning process. To visually investigates this, a more comprehensive version of the pipeline that is illustrated in Figure 18 is provided to show the three interventions points.

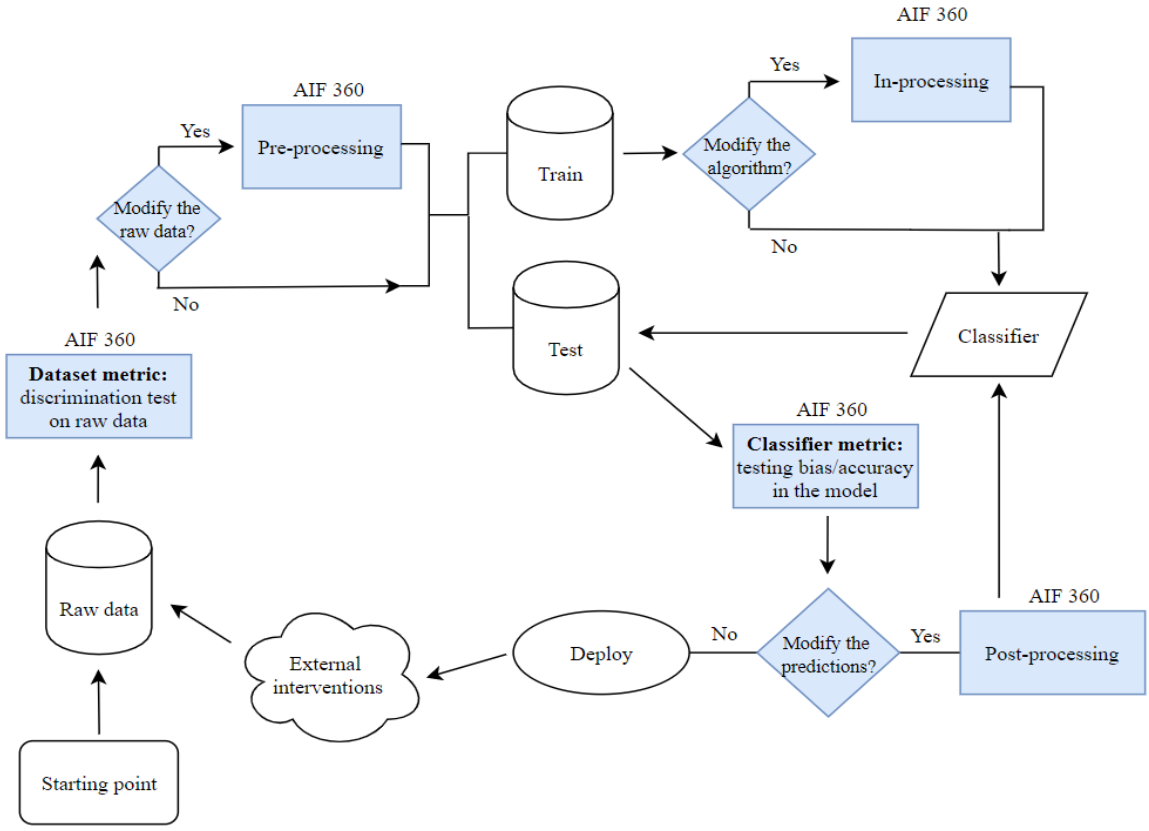


FIGURE 26:Machine learning pipeline with bias mitigation intervention points. Adapted from d'Alessandro et al (2017).

According to the figure above, the machine learning process starts with the raw data. This is the first point in the pipeline where the first mitigation algorithms can be implemented. Any mitigation technique that is applied to the raw data *before* the training step is categorized as *pre-processing*.

If the practice does not permit any change to the raw data or it is believed to be free of harmful bias, the dataset will be split into test and train partitions. Test dataset is used for model training and any intervention during this phase happens by applying *in-processing* algorithms.

Once the classification model is built and tested for accuracy, prediction outcomes can be obtained for instances the test dataset. This is where the third point of intervention can take place. If the practice requires modifying the predictions, *post-processing* techniques should be applied.

As illustrated in the table below, the AIF 360 toolkit offers all the three mitigation strategies. For each, the toolkit offers various algorithms that are listed in the table below for reference (Bellamy et a., 2017). Please note that technical and in-depth analyses of these techniques are not provided here as they fall outside of the scope of this work. Instead, emphasis is put on investigating strengths and weaknesses of each mitigation strategy.

Strategy	Technique	Description
Pre-processing	Reweighting	Generates weights for the training examples in each group combination differently to ensure fairness before classification.
	Disparate impact remover	edits feature values to increase group fairness while preserving rank-ordering within groups.
	Optimized pre-processing	Learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives.
	Learning fair representations	Finds a latent representation that encodes the data well but obfuscates information about protected attributes.
In-processing	Adversarial debiasing	learns a classifier to maximize prediction accuracy and simultaneously reduce an adversary ability to determine the protected attribute from the predictions
	Prejudice remover	Adds a discrimination-aware regulization term to the learning objective.
	Meta-fair classifier	Develops a meta-algorithm for a large family of classification problems with convex constraints. takes the fairness metric as part of the input and returns a classifier optimized with respect to that fairness metric.
Post-processing	Equalized odds postprocessing	Solves a linear program to find probabilities with which to change output labels to optimize equalized odds (Hardt et al., 2016)
	Calibrated equalized odds post processing	Optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odd objective
	Reject option classifier	Assigns favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty (Kamiran et al., 2012)

TABLE 19: List of bias mitigation techniques offered by the AIF 360 toolkit. Source: (Bellamy et al., 2018)

Pre-processing algorithms

Pre-processing methods modify the training data. In doing so, they prevent the classification model from getting exposed to prejudice in the training data and from learning discriminatory decision-making rules. This is perhaps the greatest benefits of using pre-processing algorithms; they produce clean data for any downstream task that is to follow. They can be used to achieve independence, individual fairness, and causal fairness.

Recall form the discussion on encountering bias during the model deployment process in Chapter 3 that biases in the training data can be amplified by the classification model. As such, cleaning the data as an initial step is a logical action to take. However, not all practices permit external or internal users to access the data and this can limit the use of pre-processing techniques.

Data access is one of the greatest challenges to this mitigation strategy. Most enterprises such as those in the financial industry deal with sensitive data that contains vast amount of information on data subjects. As such, they are often under heavy data protection regulations, which prevent access to sensitive data. If access is not an issue, information in the dataset may be anonymized which may cause other sorts of problems such as not knowing the underlying features that can act as proxies of the protected attribute.

Another shortcoming of the family of post-processing techniques is that they often trade-off accuracy in return for reducing discrimination. This is mainly due to the fact that pre-processing techniques transform the features into a new set by reducing or entirely removing any correlation with the protected attribute (Barocas et al., 2019).

Correlations can be thought of as hidden patterns in the data that machine learning algorithms extract to boost their accuracy. Therefore, by removing feature correlations the ability of models to maintain their overall accuracy is often reduced. Maintaining correlations can also be useful for investigating causal relationships between attributes. High correlation between features does not prove causality, but it *may* signal presence of causal relationship.

In-processing algorithms

In-processing techniques intervene during the model training phase. They can be used to achieve independence, separation, sufficiency, and individual fairness. Generally, these techniques offer more flexibility in picking trade-off between accuracy and fairness. Often applying pre-processing can result in the greatest utility since the user can optimize for a specific fairness criterion during the learning phase. This characteristic highlights one of the differences between statistical methods and discursive strategies, a notion that was mentioned earlier in this chapter.

This same characteristic of in-processing algorithms can also work to their disadvantage. Family of techniques here sacrifice generality for a more targeted approach by solving optimizing problems. This implies that in-processing algorithms are best suited for task specific scenarios (Hardt et al, 2018).

Another limitation is that many in-processing methods require access to personal data regarding protected attributes. As described above, this is a barrier to implementing such techniques. Apart from the raw data, the training pipeline needs to be accessed which augments limitations of this approach.

Post-processing algorithms

Post-processing algorithms artificially alter the predicted outcomes to ensure equal treatment across the protected and unprotected groups. They are the primary method of achieving separation and independence. When the cost of false negative and false positives is known, post-processing techniques can be used to minimize the expected cost of either misclassification according to the fairness constraint (Barocas et al., 2019).

The greatest advantage of post-processing techniques is that they do not require access to the raw data and the training phase of the pipeline. In cases where access to upstream stages of the machine learning process is limited, these techniques become practical. Another advantage of post-processing methods is that they are agnostic to the choice of classifier. This characteristic makes their implementations relatively easier, but it may come at a cost of losing accuracy,

One of the shortcomings of post-processing techniques is that they trade accuracy in achieving fairness measures. Also, majority of the post-processing algorithms are suitable for addressing group unfairness and very few proposed techniques can address both individual and group fairness (Lohia et al., 2018).

4.4 Recommendations for using the toolkit from the case study

The AIF360 toolkit can be used to test for bias at two different points of the machine learning pipeline: raw data and classifier. In practice, it is beneficial to detect bias as early as possible in the data before it enters the cross-validation stage. When using the AIF360 toolkit, the data should be pre-processed into a specific format that is compatible with the toolkit.

The German Credit Data was tested for bias using metrics in Table 15. These metrics are based on the distribution of demographic groups and the binary classes of true labels. The dataset was also tested for individual fairness using the `consistency ()`, the results indicated that there is a high disparity between in classes between similar individuals. It is, however, important to mention that the number of neighborhoods was chosen at random. In practice, stakeholders need to have a consensus on how to choose a criterion for similarity between individuals as this subjective choice can lead to implicit bias.

In chapter 3, the classifier was then tested for bias using group fairness metrics. Some metrics such as false positive errors rate balance can be measured directly while some other metrics require indirect calculation from the confusion matrices. The toolkit does not support metrics that are based on the distribution of probability scores, true labels and gender (S,Y,g).

When using the toolkit, one should keep in mind that:

- It can only be used for binary classification problems.
- Majority of its metrics require the distribution of both class labels ($Y = 0$, $Y = 1$) to be available.
- An appropriate bias metric should already be selected (Stage 1,2,3,4). Not all the metrics are supported by the toolkit and simply choosing one metric without considering some of the specifics of the settings may not suffice as an effective way to assess the model's fairness.
- Bias due to proxy features and target variable subjectivity are already checked for.
- Possible bias from formulation of business problem (stage 1) has been already checked for.
- Antidiscrimination laws and ethical standards governing the application have been carefully examined (stage 1). Managers should consult legal advisors and experts from other disciplines in process of feature selection early on to get a better understanding on whether use of sensitive features in their practice causes discrimination.

Finally, managers should keep in mind that the toolkit is more suitable for more targeted and short-term solutions and as such, it does not suffice as a long-term managing solution, especially, after models have been released 'in the wild'. Therefore, the toolkit must be considered as one promising avenue for mitigating bias that will only be effective if they are embedded in a long-term, iterative governance process which entail the classifier's entire lifecycle.

This page is intentionally left blank

5 Conclusion

This study started with identifying some of the main issues that stakeholders in consumer lending face when ML models are used to make approve/reject decisions about loan applicants. As it was mentioned earlier, there is a lack of agreement on which notions of fairness and their corresponding metrics should be considered when evaluating gender equality in consumer lending. Beyond that, metrics function only as a measure of bias without clearly explaining its potential source and placement in the process. These issues served as a motivation to investigate, as the main objective, *how can automated loan approval processes be assessed for gender equality*.

In order to systematically examine the stated problem, two essential elements for assessing fairness were considered and studied in sequence: choosing appropriate fairness metrics and testing the dataset and the model for possible bias.

This chapter concludes the thesis by providing answers to the research questions and the main findings in Section 5.1. Then, in Section 5.2, it discusses the limitations of this work and recommendations for future research. Finally, in Section 5.3, it provides the author's general reflection on artificial intelligence in banking.

5.1 Answers to the research questions and main findings

RQ1. *What fairness metrics are suitable for assessing gender equality in consumer lending?*

To answer this question, the stakeholders should have in mind a clear conception and, ultimately, formulation of fairness encompassing the perspectives of different stakeholders, particularly, those who are more likely to receive unfavorable treatment, such as women. Since the metrics serve as quantifiable measures of fairness, formulation of an appropriate definition of fairness is necessary beforehand to allow for serving as a basis of evaluation of thereof.

To conceptualize fairness, a systematic approach, in which the *good* to be distributed is identified, can be used as a first step. Secondly, the guiding principles of justice specifying fair allocations of that good are identified. Lastly, the normative theories of justice can be used to gain an understanding of what the fairest pattern of distribution of the good could be.

To illustrate, the figure from Chapter 2 is provided:

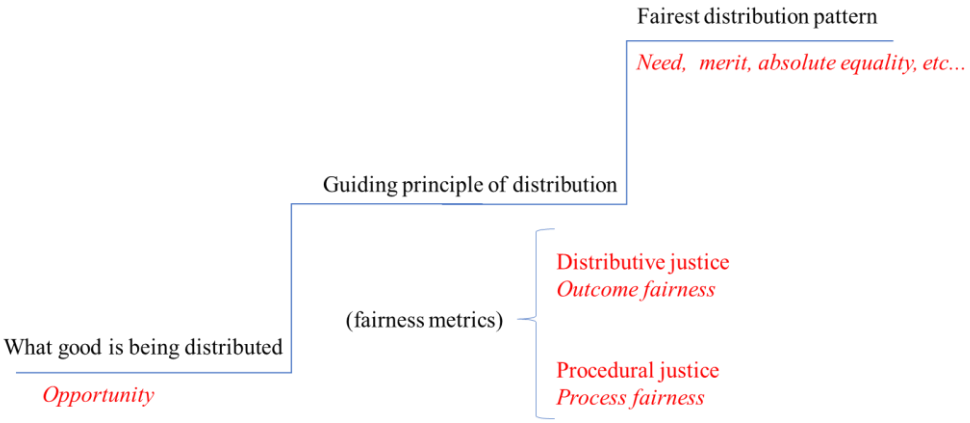


FIGURE 27: Systematic conceptualization of gender equality in consumer lending.

Based on the good (equality of opportunity) and the European laws that prohibit (indirect and direct) discrimination, one possible definition of fairness can be taken as:

Individuals who have the same effort-based utility (needs, merit, etc.,) should be given equal prospect of receiving the same utility (harm/benefit), irrespective of their gender.

Based on the above definition, the suitability of the metrics can be assessed by considering that:

- *Group fairness* metrics are a measure of distributive justice. To test whether the above definition is met, they verify, from the distribution of outcome,s whether the opportunity of receiving similar treatment is the same for each demographic group in their associated class.

- The choice of *fairness criteria* (independence, sufficiency, or separation) should be carefully examined based on their moral assumptions and the application scenario at hand, since all three cannot be achieved simultaneously.

- *Independence* assumes that all individuals have the same effort-based utility. As such, the choice of independence can lead to distribution of the same rate of loans between the two demographic groups.

- By looking at *separation* via the above definition of equality of opportunity, it is found that separation assumes the predictions to be the harm/benefit applicants receive while true labels are what justifies inequality.

- By looking at *sufficiency* via the above definition of equality of opportunity, it is found that sufficiency assumes the true labels or ground truth to be the harm/benefit applicants receive while predictions are what justifies inequality.

- *Individual fairness* metrics are a measure of distributive justice. In order to check whether similar individuals receive similar treatments, *similarity* needs to be clearly identified and this is a normative choice. For instance, similarity can be based on need: if two people have the same level of need, they should be receiving the same decision, regardless of their gender.

- *Causal measures* can be thought of as a measure of procedural justice. They check whether features selected to train the predictive models are discriminatory and/or illegitimate. Causal measures are beneficial for understanding the relationships between gender and other features in the dataset.

Above all, assessing the suitability of the definitions requires a multidisciplinary approach that draws knowledge from other domains such as philosophy, law, economics, and gender studies. This way, the perspectives of various stakeholders can be examined more efficiently, and the normative choices required to conceptualize fairness (Which guiding principle? Which distribution pattern?) can be evaluated thoroughly.

RQ2. *How can the metrics be applied to observe gender bias in lending history data?*

This research question was investigated through an exploratory case study. To apply the fairness metrics, this work used IBM's AIF360 toolkit as it provides a comprehensive set of fairness metrics in a single Python package. German Credit Data, an open source dataset containing financial information about 1000 loan applicants together with their true label classes was used throughout the research.

When applying the group fairness metrics to the model, access to both classes of the true labels in the dataset, positive and negative, is required since those metrics depend on some distribution of the predicted classes/probability scores and true labels. This is one of the primary reasons behind the usage of the German Credit Data throughout this research.

In answering this research question, Chapter 3 also investigated how to interpret the results obtained after the application of the fairness metrics. To review, let us revisit some of the biases that were detected and put them side by side to illustrate how they can be interpreted based on the moral assumptions that underlie the fairness criteria

As shown below, it is found that the choice of separation and sufficiency may have different repercussions for each demographic group in the German Credit Data. When *false distribution of utility* [P(+|-)] is under inspection, sufficiency advantages females more than males since it is more likely for male applicants who are assigned to negative class (what justifies inequality) to, in fact, have good credit (utility). On the other hand, separation advantages males more than females since it is more likely for males who truly belong to the negative class (what justifies inequality) to receive a positive predicted score (utility). Such inconsistency highlights the importance of realizing how relevant distribution of harm/benefit depends on the choice of fairness criteria made by decision makers.

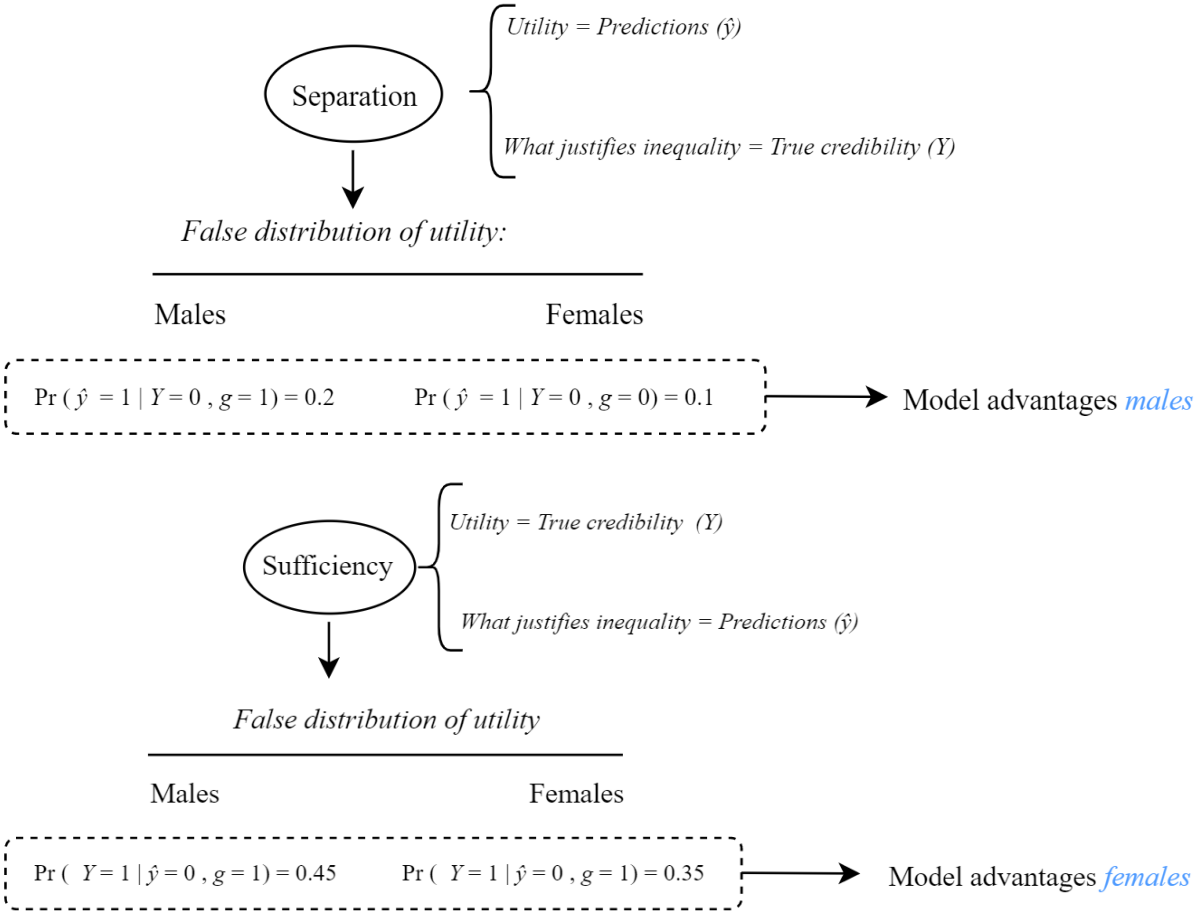


FIGURE 28:False distribution of utility under separation (above) and sufficiency (below). The results indicate that males are advantaged more than females when what justifies are the predictions. On the other hand, when inequality is justified by the true labels, females are advantaged more than males.

5.2 Limitations and recommendations

5.2.1 Limitations

Practical relevance

The second chapter is intended to form a cognitive process which may help with the interpretation of fairness metrics based on factors that lie outside of the domain of computer science. This requires understanding and synthesizing ideas that explain abstract concepts, such as fairness, discrimination, and gender equality. While this is beneficial in many ways, it is also one of the limitations of this research as its practical relevance cannot be confirmed.

For instance, one of the findings of the research was that separation and sufficiency make different moral assumptions about what constitutes utility and effort and this inherent difference can be used to evaluate which of the two is more suitable, given the application scenario at hand. However, it is very challenging to clearly define what constitutes utility and effort in the first place. For example, is utility (harm) rejecting an applicant or is it the case of a default brought about by the applicant himself?

Despite the fact that the research introduced this new line of thinking in attempt to reveal the moral equivalency assumptions of the fairness metrics, it does not provide any information on how to actually *validate* the assumptions in practice.

Gap between idealized machine learning process and complex problems in practice

It is customary to think about algorithmic fairness in terms of idealized sequence of steps in a model development process; understand business needs, collect and prepare data, select a model, train the model, predict, etc. In this framing, achieving fairness requires that all of these sequential steps are executed fairly. However, many of the problems that businesses face require more complex system interactions and, therefore, questions of ethics often fall above and beyond the scope of this sequence.

This research framed the problem of fairness around the same idealized machine learning process and conducted a uniform investigation of how bias can be encountered. For this reason, some of the key issues and problematic implications for metrics are not addressed here. For instance, even for a simple idealized process, it is more likely that risk is not uniformly distributed across its steps and to effectively tackle bias issues, one should first identify those steps which carry a bulk of the risks. Especially for more complex systems, it is often required, as the first step, to identify and tackle the steps deemed most risky or with the potential to cause greatest harm.

5.2.2 Recommendations

Build a taxonomy which shows potential barriers and disparities between members of protected groups

One of the main contributions of this research lies in demonstrating the importance of conceptualizing fairness before analyzing the suitability of fairness metrics. In order to understand what constitutes fairness, one has to take the perspectives of different individuals who may be affected by a decision into consideration.

While this may not be possible for every decision-making scenario, having access to a taxonomy that shows the potential barriers (particularly those that stem from involuntary choices) between the demographic groups, can significantly help the decision-makers and data scientists to incorporate different viewpoints. For instance, apart from being too generic, Table 4 in Chapter 2, listing the barriers between men and women in finance, can be used a simple guideline for feature selection in consumer lending.

While it is beyond the scope of this work to conduct a cost-benefit analysis of building such taxonomy and to carefully examine its effectiveness in practice, it recommends considering it as a potentially valuable area for research for two main reasons:

- In most scenarios, access to data is not permitted and therefore, evaluating process fairness is a difficult task as features are hidden to customers and other stakeholders. Thus, knowing that there are practices in place which prevent or limit the use of discriminatory features can help greatly help with building trust and integrity. This is particularly important in high-impact settings, such as lending.

- Causal studies can be computationally heavy and time consuming. By excluding discriminatory features from the dataset, the need for causality can be reduced in most scenarios.

5.3 Reflections

To finalize the research project, this section takes a broader perspective and reflects on the future prospects of artificial intelligence fairness in banking. There is no doubt that artificial intelligence has and will continue to revolutionize the banking industry amid its potential to save costs and streamline customer interactions. However, it is the general belief of the researcher that, alongside these benefits of AI, stakeholders should expect to encounter some of its ethical issues, such as prejudice and discrimination, against women for a great length of time in the future.

What shapes such sceptic belief? This reflection is being written only two days after the latest UN study found that 9 out of 10 people are biased against women in politics, economics, education, and violence despite the progress, efforts, policies, and practices which have aimed, over the years, to close the gender inequality gap. The UN study, extended over 75 countries across the globe, found that almost half of people consider men as superior political leaders and more than 40% believe that men make better business executives. Another finding of the study is that there are *no* countries in the world with gender equality.

If the very notion of gender equality that we try to achieve in ML has not been achieved yet, we cannot expect our algorithms to make bias-free decisions for us. While a great body of literature continues to contribute to this evolving field and stakeholders increasingly attempt to untangle these complex problems, stakeholders will continue to struggle to find effective ways to cope with ambiguous and rapidly evolving discriminatory data or to interpret and execute what human intentions would be if us humans could have coped with complex and multifaceted data.

Even having a person who continuously scrutinizes the automated decisions in the 'loop' may not be sufficient to produce a 'fair' decision: as cognitive AI does not make decisions in the same way as humans would, the human would not be equipped with the knowledge and information required to decide if the data-driven action fulfills our intentions. Moreover, the stochastic characteristic of cognitive AI, together with our consequent inability to know why a specific choice has been made by the system, implies that the decision is less likely to be trusted.

Throughout this research, it became increasingly apparent to the writer that every technical aspect of the research tackling fairness issues should only be considered as a short term and temporary solution. Recently, long-term discursive strategies and less-technical methods have emerged with the goal to detect and measure bias thus providing lifelong governance inclusive solutions to effectively address algorithmic bias and discrimination. In conclusion, as long as we generate biased and discriminatory datasets, we will require improved understanding among academics, experts, and policy makers about the nature of the problems and the array of bias detection and mitigation techniques available.

This page is intentionally left blank.

6
References

Adams, J. S. (1965) Inequity in social exchange. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 2, pp. 267-299). New York: Academic Press.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). "Machine Bias".

Arneson, (2015), "Equality of Opportunity", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2015/entries/equal-opportunity. Avoiding Discrimination through Causal Reasoning. In NIPS.

Barocas ,S. Narayanan, A. Hardt, M. (2019). "Fairness and Machine Learning". Fairml.org.

Barocas, S., & Selbst, A.D. (2016). Big Data's Disparate Impact.

Beahrs, J. O. (1991). Volition, Deception, and the Evolution of Justice. Bulletin of the American Academy of Psychiatry & the Law.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K. et al., (2018). "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias". arXiv preprint arXiv:1810.01943.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of Machine Learning Research 81 (2018), 1–11.

Binns, R. Van Kleek, M. Veale, M. Lyngs, U. Zhao, J. and Shadbolt, N. (2018) 'it's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In Proceedings of the 2018 CHI. Conference on Human Factors in Computing Systems, page 377. ACM. URL

Blader, S. Tyler, T. (2003). A four-component model of procedural justice: Defining the meaning of a "fair" process. Personality and Social Psychology Bulletin 29, 6 (2003), 747–758.

Blond, P. Milbank, J (2010). "No equality in opportunity: By synthesizing old Tory and traditional left ideas a genuinely egalitarian society can be achieved".

Borgesius. F. Z. (2018). *Council of Europe*: "Discrimination, artificial intelligence, and algorithmic decision-making".

Bowdish, L. (2010). *Invidious Distinctions: Credit Discrimination Against Women, 1960s–Present.* (Electronic Thesis or Dissertation). Retrieved from https://etd.ohiolink.edu/.

Brodersen, K. Ong, C. Stephan, K. Buhmann, J. (2010). "The Balanced Accuracy and Its Posterior Distribution," 2010 20th International Conference on Pattern Recognition, Istanbul, pp. 3121-3124.

Carmer, H., Garcia-Gathright, J., Springer, A., Reddy, S. (2018). Assessing and addressing algorithmic bias in practice.

Chan, D. (2011). Fairness: Processes as important as outcomes What research on fairness perception tells us about policy and politics. The Straits Times, p A30.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C.R., Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Chouldechova, A. (2016) "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." arXiv:1610.075254v1.

Colquitt, J. A, Rodell, J. B. (2015). Measuring justice and fairness. The Oxford handbook of justice in the workplace place187, 202.

Corbett-Davies, S. Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018).

Cropanzano, R. Molina, A . (2015). Organizational Justice. 10.1016/B978-0-08-097086-8.22033-3.

Custers B., Calders T., Schermer B., Zarsky T. (eds) Discrimination and Privacy in the Information Society. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 3. Springer, Berlin, Heidelberg.

d'Alessandro, B. O'Neil, C. LaGatta, T. (2017). Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. Big Data. 5. 120-134. 10.1089/big.2016.0048.

Datta, A., Sen, S., Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. 2016 IEEE Symposium on Security and Privacy (SP), 598-617.

Datta, A., Sen, S., Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. 2016 IEEE Symposium on Security and Privacy (SP), 598-617.

Dattner, B., Buchband, R., Schettler, L. Chamarro-Premuzic, T. (2019). *Harvard Business Review*: *"*The Legal and Ethical Implications of Using AI in Hiring".

Denmark, F., & Paludi, M. A. (2008). *Psychology of women: A handbook of issues and theories*. Westport, Conn: Praeger.

Denning, P. J. A New Social Contract for Research, Communications of the ACM (40:2), February 1997, pp. 132-134.

Dobbie, F., Arthur, F. S., & Jones N. (2010). Building Understanding of Fairness, Equality and Good Relations in Scotland. Equality and Human Rights Commission Research Report, 53, 1-122.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference, 214–226. ACM.

ECHR. (1953). *The European Convention on Human rights: "*Article 7 of the United Nations Declaration of Human Rights; Article 26 of the International Covenant on Civil and Political Rights; Article 21 of the Charter of Fundamental Rights of the European Union".

ECtHR. (2016). *European Court of Human Rights*: "Biao v. Denmark (Grand Chamber), No. 38590/10, 24 May 2016", para. 89. et al., 2016).

Fair Employment Practice Commission. (1974). "Fair Employment Practice Commission Report: July 1, 1971 - June 30, 1972" California Agencies. Paper 49.

FAT*. [n.d.]. ACM Conference on Fairness, Accountability, and Transparency.

Friedman, B., Nissenbaum, H. (1996). "Bias in Computer Systems" . *ACM Transactions on Information Systems*. 14 (3): 330–347. doi:10.1145/230538.230561

Gheaus, A. (2016). Gender and Distributive Justice: Forthcoming in Serena Olsaretti (ed.) Oxford Handbook of Distributive Justice, Oxford University Press.

Gilliland, S. (1993). The Perceived Fairness of Selection Systems: An Organizational Justice Perspective. The Academy of Management Review, 18(4), 694-734.

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". AI Magazine, 38, 50-57.

Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., & Weller, A. (2018). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. AAAI.

Hao, K. (2019). Massachusetts Institute of Technology Review. "*This is how bias really happens-and how to fix it".*

Hardt, M., Price, E., Srebro, N., et al. (2016). "Equality of opportunity in supervised learning. In Advances in neural information processing systems". (pp. 3315–3323).

Hausman, D. M., McPherson, M. S. (2006). Economic analysis, moral philosophy, and public policy (2nd ed.). New York, NY: Cambridge: Cambridge University Press.

Heidari, H. Loi, M. & Gummadi, K. Krause, A. (2019). A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. 181-190. 10.1145/3287560.3287584.

Hevner, A. R. (2007). A three-cycle view of design science research. *Scandinavian journal of information systems*, 19(2), 4.

Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004), "Design Science in Information Systems Research, MIS Quarterly, 28(1), pp. 75-105.

Hinnefeld, J. & Cooman, Peter & Mammo, Nat & Deese, Rupert. (2018). Evaluating Fairness Metrics in the Presence of Dataset Bias.

Holloway, K., Rouse, R., & Niazi, Z. (2017) Women's Economic Empowerment through Financial Inclusion: A Review of Existing Evidence and Remaining Knowledge Gaps. Innovations for Poverty. Action. http://hdr.undp.org/en/content/table-4-gender-inequality-index.

Holmes, T. E. (2019). *Research and statistics:* "Credit card and market share statistics: The largest card issuers held more than $800 billion in outstanding loans as of 2018".

Huang, K. Saxena, N, A. DeFilippis, E. Radanovic, G. Parkes, D. C, Liu, Y. (2019). "How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness."

Hurley, M., Adebayo, J. (2016). Credit scoring in the era of big data. Yale JL & Tech., 18, 148.

Hurthouse, R., & Pettigrove, G. (2016). Virtue ethics. In E. Zalta (Ed.), Stanford encyclopedia of philosophy. Retrieved from https://plato.stanford.edu/.

Joseph, M. Kearns, M. Morgenstern, J. H.; and Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In Advances in Neural Information Processing Systems, 325–333.

Kamiran, F., Calders, T. (2011). Data Pre-Processing Techniques for Classification without Discrimination. Knowledge and Information Systems. 33. 10.1007/s10115-011-0463-8.

Kamishima, T., Akaho, S., Asoh, H., Sakuma, J., (2012) "Fairness-Aware Classifier with Prejudice Remover Regularizer," Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Kaplan Schweser. (2018). *Schweser Notes 2019 Level II CFA Book 2:* "Financial Reporting and Analysis and Corporate Finance", Page 126- 139.

Kendig, D. (1973). *Discrimination Against Women in Home Mortgage Financing*, 3 Yale Rev. L. & Soc.Action. https://digitalcommons.law.yale.edu/yrlsa/vol3/iss2/3.

Khandani, A.E., Kim, A.J., & Lo, A.W. (2010). Consumer Credit Risk Models Via Machine-Learning Algorithms.

Kibe, T. (2011). The relational approach to egalitarian justice: A critique of luck egalitarianism. Critical Review of International Social and Political Philosophy. 14.1-2110.1080/09692290.2010.517984.

Kilbertus, N. Hardt, M. (2017). "Avoiding Discrimination through Causal Reasoning", Advances in Neural Information Processing Systems 30, 2017, p. 656--666, arXiv:1706.02744.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Sch¨olkopf, B. (2017).

Kleinberg, J.M., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *ArXiv, abs/1609.05807*.

Koene, A. Perez, E. Ceppi, S. Rovatsos, M. Webb, H. Patel, M. Jirotka, M. Lane, G. (2017). Algorithmic Fairness in Online Information Mediating Systems. In Proceedings of ACM Web Science Conference, Troy, NY, USA.

Kraemer, F., van Overveld, K., Peterson, M. (2011) Is there an ethics of algorithms? Ethics Inf Technol 13, 251–260. https://doi.org/10.1007/s10676-010-9233-7

Kulik, C.T., Lind, E.A., Ambrose, M.L. et al. Soc Just Res. (1996). Understanding gender differences in distributive and procedural justice 9: 351. https://doi.org/10.1007/BF02196990.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. (2017). Counterfactual Fairness. In NIPS.

Lamont, J. Favor, C. (2017). "Distributive Justice", The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.), URL *https://plato.stanford.edu/archives/win2017/entries/justice-distributive/*

Lohia, P., Ramamurthy, K., Bhide, M., Saha, D., Varshney, K., Puri, R., (2018). Bias Mitigation Post-processing for Individual and Group Fairness.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society.

Lee, M.K., Jain, A., Cha, H.J., Ojha, S., & Kusbit, D. (2019). Procedural Justice in Algorithmic Fairness:

Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. PACMHCI, 3, 182:1-182:26.

Lefranc, A. Pistolesi, N. Trannoy, A. (2009). Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. Journal of Public Economics 93, 11-12 (2009), 1189–1207.

Leventhal, G. S. 1980. What should be done with equity theory? In Social exchange. Springer, 27–55.

Lind, A. E, Tom. R Tyler, T. T. Huo, J. (1997). Procedural context and culture: Variation in the antecedents of procedural justice judgments. Journal of Personality and Social Psychology 73, 4 (1997), 767.

Lippert-Rasmussen, K. (2014). Indirect Discrimination is Not Necessarily Unjust. Journal of Practical Ethics Volume 2 Issue 2, 2014. Available at SSRN: https://ssrn.com/abstract=2536870

Liu, Y.; Radanovic, G.; Dimitrakakis, C.; Mandal, D.; and Parkes, D. C. (2017). Calibrated fairness in bandits. arXiv preprint arXiv:1707.01875.

Llewellyn, D. (2005). *Journal of Financial Regulation and Comp*liance: "Trust and confidence in financial services: A strategic challenge".

Loi, M., Herlitz, Anders., Heidari, H. (2019) A Philosophical Theory of Fairness for Prediction-Based Available at SSRN: https://ssrn.com/abstract=3450300 or http://dx.doi.org/10.2139/ssrn.3450300.

Lomazzi, V.; Israel, S.; Crespi, I. (2018). Gender Equality in Europe and the Effect of Work-Family Balance Policies on Gender-Role Attitudes. *Soc. Sci*. 2019, *8*, 5.

Malhotra, A. Schuler, S. (2005). Measuring Women's Empowerment as a Variable in International Development. Measuring Empowerment: Cross Disciplinary Perspectives.

March, S. T. & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, 15(4), 251–266.

Messick, D. M., & Sentis, K. (1983). Fairness, preference, and fairness biases. In D. M. Messick &

Mil, J. S. (2016). Utilitarianism. In Seven masterpieces of philosophy. Routledge, 337–383.

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artif. Intell., 267, 1-38.

Narayanan, A. (2018). *Translation tutorial*: "21 fairness definitions and their politics". In Proc. conf. fairness accountability transp., New York, NY, USA.

Nickel, J. W. (2000). "Discrimination", in Edward Craig and Edward Craig (eds.) Concise Routledge.

Oliver, A. Massialos, E. (2004). Equity of Access to Health Care: Outlining the Foundations for Action. Journal of Epidemiology and Community Health 2004 August; 58(8): 655-658

Olsaretti, S. (2018). The Oxford Handbook of Distributive Justice, Oxford University Press Option consommateurs. (2014). *Option consommateurs and presented to Industry Canada's Office of Consumer Affairs:* "How well do newcomers understand credit?"

Pereira, R. Schwanen, T. Banister, D. (2016). Distributive justice and equity in transportation. Transport Reviews. 1-22. 10.1080/01441647.2016.1257660.Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society 2nd.

Pruitt, D.G., Peirce, R.S., McGillicuddy, N.B. et al. Law Hum Behav (1993) 17: 313. https://doi.org/10.1007/BF01044511

Quinones. P., (2016). Female Entrepreneurs: Adding a New Perspective to EconomicGrowth. Fundación Rafael del Pino.

Rawls, J. (1999). A Theory of Justice. Harvard University Press.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: a practical framework for public agency accountability.

Rescher, N. (2002). Rescher's *Fairness: Theory & Practice of Distributive Justice.*

Ridgeway, C. L. (2011). *Framed by gender: How gender inequality persists in the modern world*. New York: Oxford University Press.

Roemer, J. (2002). Equality of opportunity: A progress report. Social Choice and Welfare 19, 2 (2002), 455–471.

Romei A., Ruggieri S. (2013). Discrimination Data Analysis: A Multi-disciplinary Bibliography. In:

Rovatsos, M., Mittelstadt, B., Koene, A. (2019). Landscape summary: Bias in algorithmic decision-making.

Salles, M. (2017). Felix Brandt, Vincent Conitzer, Ulle Endriss, Jerôme Lang, and Ariel Procaccia (eds), Handbook of Computational Social Choice. OEconomia. History, Methodology, Philosophy 7-4 (2017), 609–618.

Shira, M. Potash, E. Barocas, S. (2018). Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions.

Siddiqi, N. (2006). Credit Risk Scorecards: developing and implementing intelligent credit scoring. Wiley, New Jersey.
Simon, H. (1996). *The Sciences of Artificial*, 3rd Edition, MIT Press, Cambridge, MA.

Smith, E., Lipkovich, I., Ye, K., (2002). Weight of Evidence (WOE): Quantitative Estimation of Probability of Impact.
Srivats, K. R. (2016). "FICO, Lenddo partner to develop credit risk scores for consumers in India".

Stancil, P. J. (2016). Substantive Equality and Procedural Justice. Iowa Law Review, Forthcoming; BYU Law Research Paper No. 16-06. Available at SSRN: https://ssrn.com/abstract=2764240

Sweeney, L. (2013). Discrimination in online ad delivery. arXiv preprint arXiv:1301.6822.

Taylor, T. (2015). Equality of Opportunity and Equality of Result. http://conversableeconomist.blogspot.com/2015/06/equality-of-opportunity-and-equality-of.html

Thibaut, J. W. Walker, L. (1975). Procedural justice: A psychological analysis. L. Erlbaum Associates.

Trankell, A. (1972). Reliability of Evidence: Methods for Analyzing and Assessing Witness Statements.

Tsichritzis, D. The Dynamics of Innovation, in Beyond Calculation: The Next Fifty Years of Computing, P. J. Denning and R. M. Metcalfe (eds.), Copernicus Books, New York, 1998, pp. 259-265.

United Nations Development Programme. (2014). Gender Inequality Index. Available at:

United Nations. (2010). The World's Women.

Van den Bos, K., Lind, E. A., & Wilke, H. A. M. (2001). The psychology of procedural and distributive justice viewed from the perspective of fairness heuristic theory. In R. Cropanzano (Ed.), *Justice in the workplace: Vol. 2* (pp. 49-66). Mahwah, NJ: Erlbaum.

Verma, S. Rubin, J. (2018). "Fairness Definitions Explained". In Fair-Ware '18: IEEE/ACM International Workshop on Software Fairness, May 29, 2018, Gothenburg, Sweden.

Vossenberg, S. Rappoldt, A. D'Anjou, J. (2018). Beyond Access: Exploring gender-transformative approaches to financial inclusion.

Walster, Elaine & Berscheid, Ellen & Walster, G. (1973). New Directions in Equity Research. Journal of Personality and Social Psychology. 25. 151-176. 10.1037/h0033967.

World Economic Forum. (2016). *The World Economic Forum:* The global gender gap report.

Wright, S. Boese, G. (2015). Meritocracy and Tokenism. International Encyclopedia of the Social & Behavioral Sciences. 10.1016/B978-0-08-097086-8.24074-9.

Yaari, M. E, Bar-Hillel, M. (1984). On Dividing Justly. Social Choice and Welfare.

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded generalized reciprocity: Ingroup boasting and ingroup favoritism. In E. J. Lawler & M. W. Macy (Eds.), Advances in group processes (Vol. 16, p. 161-197). Jai Press Inc.
York: Praeger.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". In Proceedings of the 26th international conference on world wide web (pp. 1171–1180).

This page is intentionally left blank.

Appendix

Investigating the trolley problem of machine learning with AIF 360 toolkit.

A machine learning classification model would be ideal if it can satisfy all the criteria simultaneously. If this cannot be achieved, a choice for one must be made based on the application scenario at hand. The investigation of the trolley is done in a pairwise manner as illustrated in the table below. For every case listed except for case 1, the condition to hold will be checked, starting with *separation* and *sufficiency*.

Separation and sufficiency	Case 1	Case 2	Case 3
Assuming (S or $\hat{y}$)	$(\hat{y} \perp g \mid Y)$ and $(\hat{y} \perp g \mid S)$	Balance for positive class, balance for negative class and, calibration within groups.	Equal false positive rates, false negative rates, and positive predictive parity across groups.
Conditions to hold (at least one)	• $(Y, \hat{y}) \perp g$ • An event in the joint distribution has probability zero.	• Equal base rates: $(Y \perp g)$ • Perfect prediction	• Equal base rates: $(Y \perp g)$ • False positive rate = 0 and positive predictive parity = 1 • False positive rate = 0 and false negative rate = 1

Independence and separation	
Assuming (S or $\hat{y}$)	$(\hat{y} \perp g \mid Y)$ and $(\hat{y} \perp g)$
Conditions to hold (at least one)	• Equal base rates • $Y \perp \hat{y}$

Independence and sufficiency	
Assuming (S or $\hat{y}$)	$(Y \perp g \mid \hat{y})$ and $(\hat{y} \perp g)$
Conditions to hold	Equal base rates: $(Y \perp g)$

Separation and sufficiency

Case 2: balance for positive class, balance for negative class and, calibration within groups

Recall that calibration within groups is subject of sufficiency while balance for positive class and balance for negative class belong to separation. In order for all three to hold simultaneously, at least one of the following conditions should be met:

1. Equal base rates
2. Perfect prediction

In a population, base rate is determined by the marginal distribution of the response. It can be computed in two ways that complement each other: the proportion of actual positive labels to

the total number of observations, or the proportion of actual negative labels to the total number of observations. Mathematically, this can be expressed as:

$$\text{Base rate} = \frac{TP+FN}{TP+FN+FP+TN} \quad \text{or} \quad \frac{TN+FP}{TP+FN+FP+TN}$$

Base rates can be computed directly using the built-in `base_rate` metric of the `BinaryLabelDatasetMetric` class. For the German Credit Data, it can be shown that base rates for males and females are not strictly equal:

$$(\text{Base rate})_{\text{male}} = 0.72 \neq (\text{Base rate})_{\text{female}} = 0.64$$

Moving on to the second condition, the simplest way to check whether the classification model is perfect (100% accurate) is to check for accuracy. This has already been done in the previous section through estimating values of overall accuracy (or balanced accuracy) at different classification thresholds. Focusing on overall accuracy, the less conservative measure of accuracy of two, using the accuracy metric based on the `ClassificationMetric` class.

In conclusion, *none* of the two conditions required for balance for positive class, balance for negative class and, calibration within groups to hold simultaneously are met and therefore, case 2 of separation and sufficiency should be rejected.

Case 3: Equal false positive rates, false negative rates, and positive predictive parity across groups

To recap, equal false positive and false negative rates are metrics described by separation while positive predictive parity across groups adheres to sufficiency. In order for this case to hold, at least one of the three conditions listed below should be true:

1. Equal base rates
2. False positive rate = 0 and positive predictive parity = 1, for both groups
3. False positive rate = 0 and false negative rate = 1, for both groups

In the analysis of case 2, equal base rates were shown to be disparate across males and females. To check for conditions 2 and 3, false positive rates can be calculated for the test set either from the confusion matrices of privileged and unprivileged groups shown above, or directly using the `false_positive_rate` metric of the `ClassificationMetric` class. Either method gives identical results. If false positive rates are calculated to be different than zero, then both conditions 2 and 3 will automatically not hold.

$$(\text{False positive rate})_{\text{male}} = (0.1951 \sim 19.51\%) \neq 0$$
$$(\text{False positive rate})_{\text{female}} = (0.0909 \sim 9.1\%) \neq 0$$

In conclusion, none of the three conditions are met and therefore, case 3 is also dismissed.

Independence and separation

For the independence and separation to hold simultaneously, at least one of the following two conditions should be true:

1. Equal base rates

2. $Y \perp \hat{y}$

Equal base rates are not equal, and this only leaves the second condition to be investigated. Condition 2 states that true label and probability scores (or decisions) are statistically independent of each other. Condition 2 can be dismissed as predictions and true labels are correlated. Otherwise, the model would randomly classify each instance.

Independence and sufficiency

The only condition needed for sufficiency and independence to hold simultaneously is having equal base rates amongst the privileged and unprivileged groups. Thus, it can be concluded that for the German Credit Data, the two statistical criteria of independence and sufficiency cannot be achieved due to the face that base rates are different.
To summarize, the above investigation of the trolley problem of machine learning with the toolkit showed that for the dataset and the logistic regression model implemented, no two pair of the statistical fairness criteria can be achieved simultaneously. This wicked problem means two things:

- Some type of discrimination is *unavoidable.*
- A choice has to be made between different types of discrimination.

The moral tension between use of each criteria is particularly important in situations where decision outcomes may have long lasting impacts on data subjects such as in consumer lending. In chapter 2, the moral tensions between the fairness criteria were discussed comprehensively, but in the next section, they are restated after the model is independently investigated for each statistical fairness criteria through some of the metrics offered by the AIF 360 toolkit.