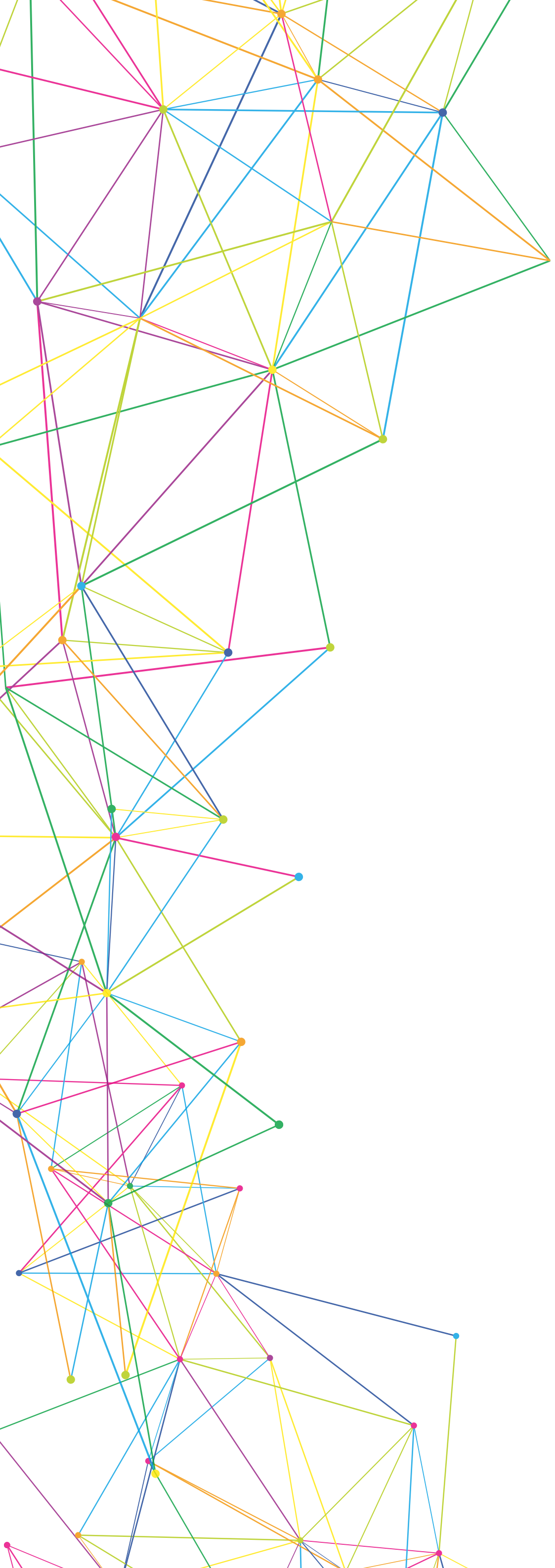
Gemeente
Amsterdam

The Fairness Handbook

May 2022



Summary

Artificial Intelligence technologies are increasingly embedded into our daily lives. The City of Amsterdam is no exception, with work processes being progressively supported by automated systems and data-driven decisions, leading to efficiency gains and improved service delivery. However, working with AI systems also poses a risk, as they have the potential to propagate harmful patterns on a large scale through the presence of undesired biases. These biases are often caused by the model using sensitive information obtained from datasets, such as a person's age or gender, to base its decisions on. The biases can lead to a model discriminating against individuals or minority groups, resulting in undesired outcomes such as being withheld access to services.

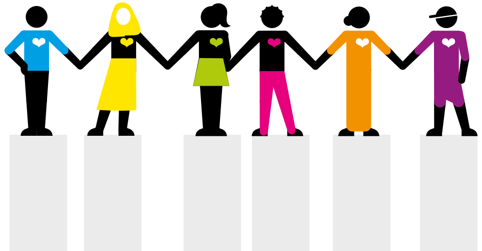
In this handbook, we delve into all things related to fairness, biases and how to minimize potential harms caused by AI systems. We provide you with an A-Z manual to measure how fair your model is and to mitigate the biases you encounter.

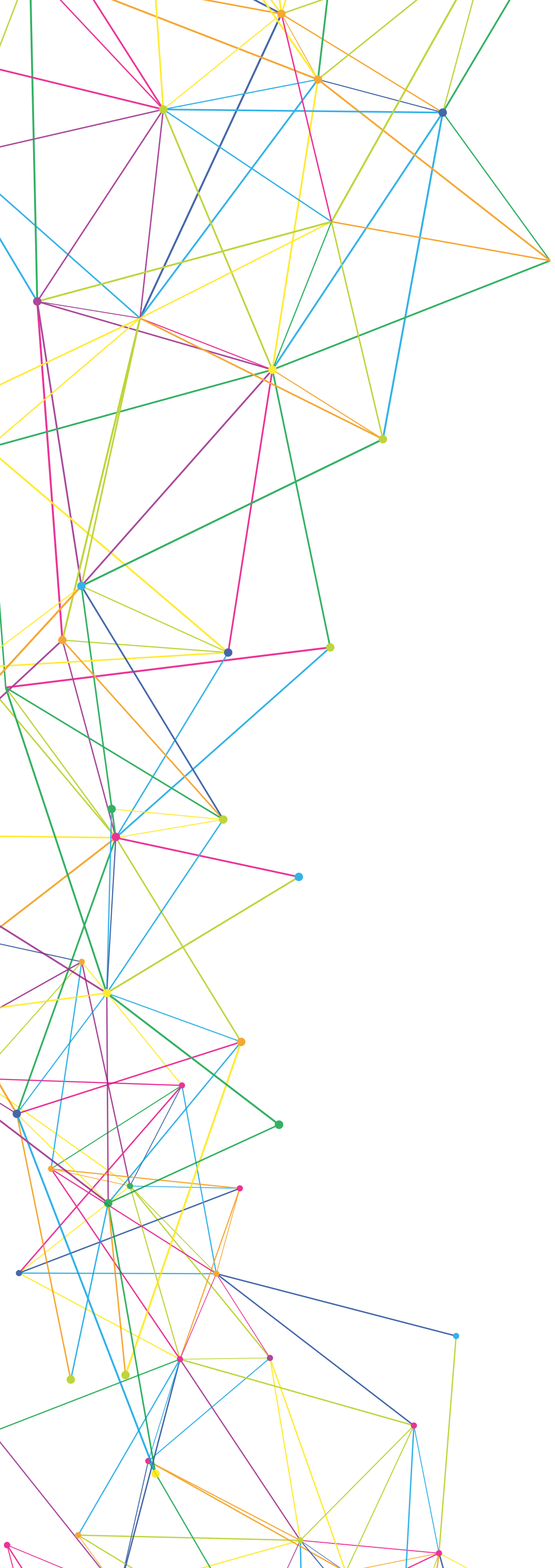
Fairness is a broad concept that can be analysed from multiple angles, which we will briefly discuss in Chapters 1 and 2. To then inspire you to reason more deeply about the potential impact of your AI system, we discuss which harmful effects a model can cause to individuals and groups of people in Chapter 3.

After these general introductory topics, we introduce you to the **Fairness Pipeline** in Chapter 4, which is the foundation of our Fairness Handbook. The pipeline consists of a series of mitigation techniques, actions and good practices throughout the model development cycle to find, mitigate and prevent harmful traps and biases. We use both technical and non-technical solutions and focus on enhancing transparency and understandability of ML models, as understandability is often key for preventing discriminating effects by algorithms.

In Chapter 5, we dive into the various definitions and metrics of fairness that can be used during **the bias analysis** for evaluating *whether* and *how* the model perpetuates discriminatory effects on individuals and groups. The **bias analysis**, a significant part of the Fairness Pipeline, starts by envisioning which groups might be negatively impacted by the AI system and which harms could be perpetuated by the model. Using this information, we select the appropriate fairness definition and associated metrics with help of the Fairness tree. The adopted fairness metric then compares the model's performance across demographic the groups in the dataset to find out for which groups the model is underperforming and/or discriminating against.

Lastly, for those who want to learn more about the types of biases and traps that may occur in Machine Learning (ML) models, or the mitigation algorithms that can be used in the Fairness Pipeline, the Appendix provides a wealth of information and useful sources.





Content

Summary 2

Content 3

1. Introduction 4

2. What is Algorithmic Fairness? 6

3. Harmful Effects of AI Systems 10

4. The Fairness Pipeline 15

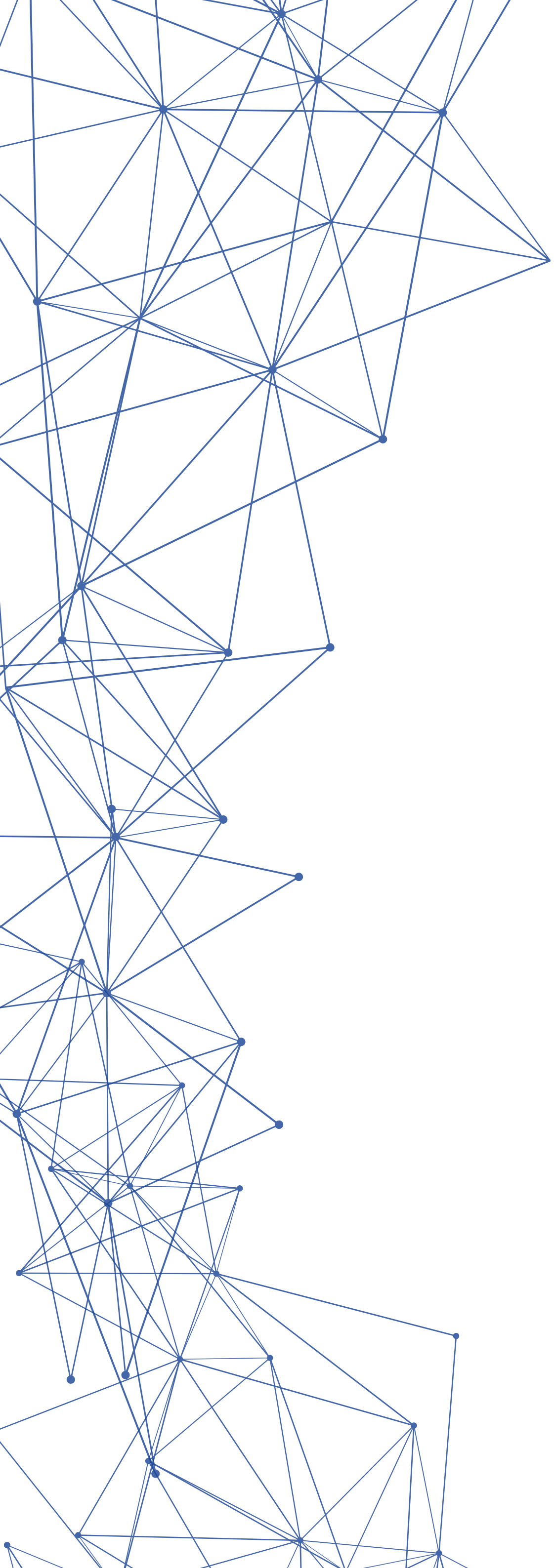
5. The Bias Analysis 34

6. Conclusion 54

7. References 55

8. Appendix 59
   A: Overview of Traps and Biases 59
   B: Mitigation Algorithms 64
   C: Further Readings 66
   D: Model Card Example 67

Colofon 68



1. Introduction

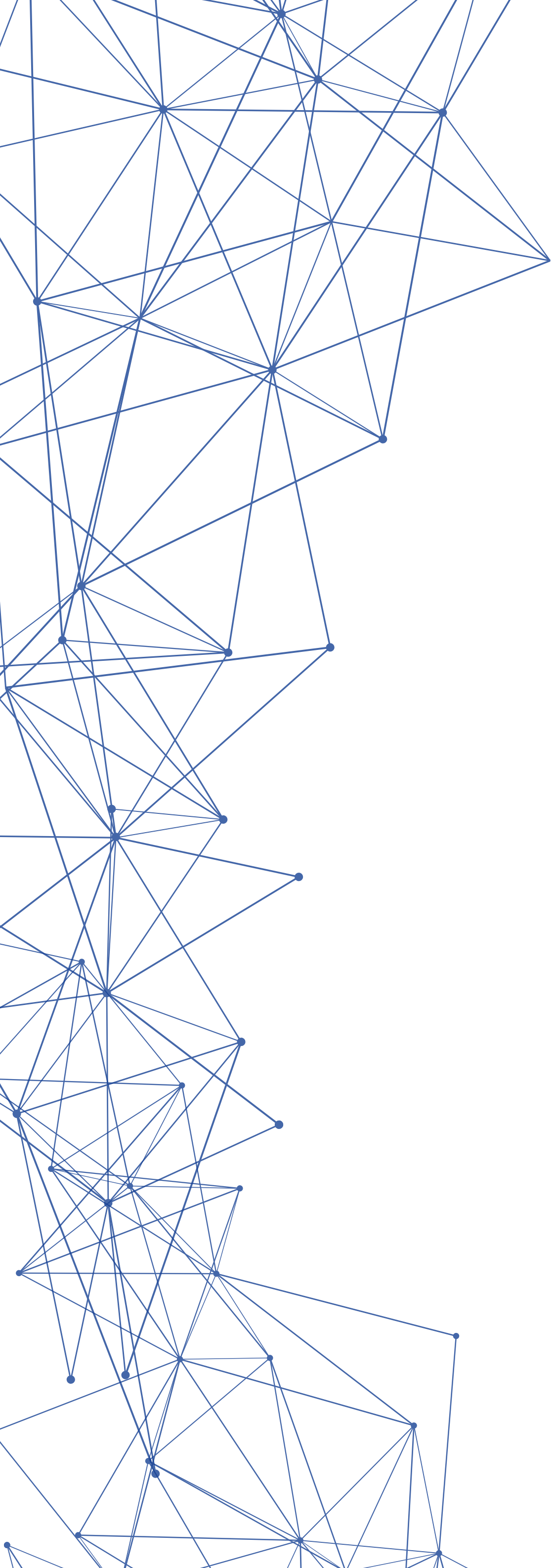
The decisions and outcomes from Machine Learning models have an impact on individuals and groups of people in the real world: they can speed up complex processes, such as finding waste on the streets or allocating social benefits, but their impact can also be harmful, potentially leading to the discrimination against individuals and vulnerable groups. In the model development cycle, we make a series of small, often implicit decisions that affect the scale and scope of these impacts.

To mitigate the harmful risks of AI systems, the City of Amsterdam uses a variety of measures that increase the transparency, understandability and fairness of the algorithms we construct. By making models more transparent and explainable, we can understand how the model arrived at its decisions and evaluate whether these decisions are based on correct assumptions, which enhances the overall trustworthiness of our AI systems. To make algorithms fairer, the City collaborates with internal and external parties, conducts the Artificial Intelligence Impact Assessment and is continuously further developing the bias analysis to prevent discriminating behaviour by algorithms.

However, **making algorithms fair is a challenging task,** as defining what a fair model always depends on the context in which the model is developed and deployed. Therefore, we need deeper insight into how algorithmic biases can seep in at every stage of the model development cycle and how they eventually harm vulnerable groups and individuals, so that we can use effective mitigation techniques to combat these problems.

In this fairness handbook, we are particularly focusing on **embedding fairness measures in the development cycle of AI systems.** We do this by walking you through the most common problems that introduce harmful biases in algorithms, after which we discuss strategies to find, mitigate and prevent these undesired effects. Our findings are summarized in our Fairness Pipeline that guides you towards developing fair models. Note that, although our scope is Machine Learning models, a lot of information is also directly applicable to other types of models as well.

With this Fairness handbook, we hope to minimize the presence of any harmful impact of the AI system on citizens.

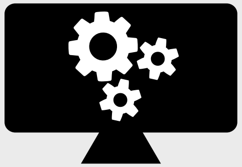


Learning Objectives

At the end of this handbook, you will have insight in:

- What algorithmic fairness and bias entails;
- How to choose the appropriate definition and metric of fairness for your model;
- The harms that AI systems can perpetuate;
- The most common biases and how they can be found;
- Bias prevention and mitigation techniques;
- Which high-quality sources can be consulted for further research.

> Note that not all sections of the chapters might be of interest for stakeholders without a technical background. We indicate these sections with the following symbol:



We need deeper insight into how algorithmic biases can seep in at every stage of the model development cycle and how they eventually harm vulnerable groups and individuals.

2. What is Algorithmic Fairness?

Algorithmic Fairness is the field which studies how algorithmic systems should behave to treat people fairly, that is, without discrimination on the grounds of protected sensitive characteristics such as age, gender, disability, ethnic or racial origin, religion or belief, or sexual orientation (Weerts, 2021a). As prediction-based decision-making systems are increasingly applied by a variety of industries and governments, the question of how to ensure their fairness is becoming more relevant every day

To understand algorithmic fairness, we must first understand algorithms. An algorithm is simply a set of (detailed) instructions for solving a problem, or accomplishing a task. This handbook relates specifically to algorithms used for decision-making. The basis of these decision-making systems is learning and generalizing from historic data.

The algorithm development lifecycle is a series of human choices and practices leading to the development of an algorithm (also known as *model* or *AI system*). This cycle includes formulating the target variable, collecting and processing the data, evaluating the model's performance and finally deploying and integrating the model in the work processes of the organization. **These choices have the potential to introduce bias in AI systems, which are skewed outcomes based on sensitive characteristics such as age or gender, that lead to discrimination.** This discrimination can take the form of disproportionately assigning undesired outcomes to underrepresented groups in the dataset, such as

Protected Attributes
(Dutch & EU Law)

- Migration Background
- Nationality
- Race
- Ethnicity
- Country of Birth
- Gender
- Sex
- Sexual Orientation
- Religion
- Age
- Pregnancy
- Civil Status
- Socioeconomic Class
- Income
- Skin Colour
- Language
- Political Views
- Health
- Disability status
- Biometrics

Figure 1: The protected attributes on which it is highly undesired or prohibited to discriminate against

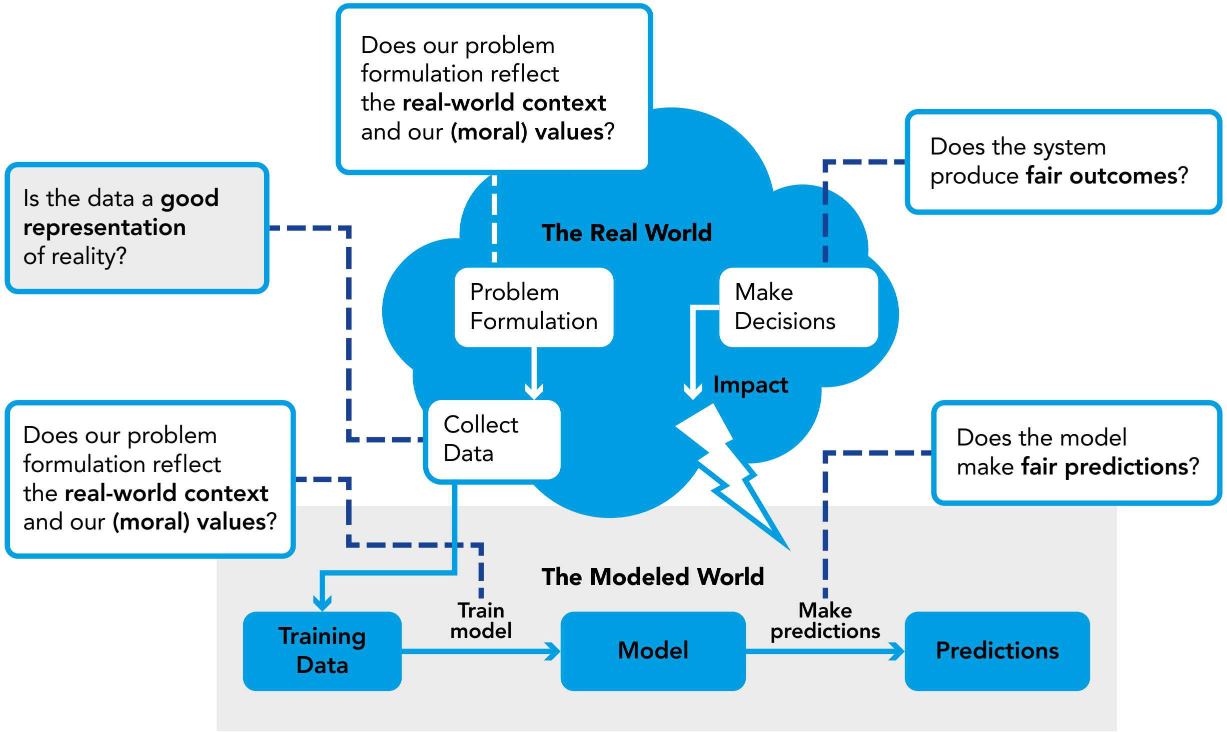

Does our problem formulation reflect the real-world context and our (moral) values?
Is the data a good representation of reality?
Does the system produce fair outcomes?
The Real World
Problem Formulation
Make Decisions
Impact
Collect Data
Does our problem formulation reflect the real-world context and our (moral) values?
Does the model make fair predictions?
The Modeled World
Train model
Make predictions
Training Data
Model
Predictions

Figure 2: Fairness issues often arise due to mismatches between our goals and what we actually value. Source: Hilde Weerts

classifying higher fraud risk scores to them. The same algorithm may also favour the overrepresented group of persons by assigning desirable outcomes to them, leading the fraud prediction model to incorrectly assign them lower fraud risk scores.

The main source of bias lies in the data on which the model is trained. **The dataset can reflect human decisions or second-order effects of societal or historical inequities** (Silberg & Manyika, 2019). During the model training phase, the ML (Machine Learning) model will pick up these flawed patterns and its output will reflect existing prejudices, inequalities and stereotypes (Barocas et al., 2016). And as ML models are deployed on an increasingly large scale, they will not only reflect but also systematize and amplify structural prejudices and inequalities.

It is important to note that algorithmic fairness is a multidisciplinary field, which combines data science with sociology, economics, philosophy and other disciplines. **The roots of bias are found in the real-world**, where patterns found within existing datasets reflect historic demographic and socio-economic disparities between people. To develop fair algorithms, we therefore cannot simply rely on technical solutions. Instead, **we must gain additional insight in the socio-technical environment** in which the AI system is built and deployed. Thus, finding bias in an AI system requires effort from a wide range of stakeholders.

Socio-Technical Context

The socio-technical refers to the environment surrounding a technical system, including both social and technical aspects. This environment shapes who might benefit or is harmed by AI systems (Fairlearn, n.d.).

Social Aspects:

- People;
- Institutions;
- Regulations;
- Political environment;
- Employee-management relations;
- Communities.

Technical aspects:

- Algorithms;
- Model infrastructure;
- Data;
- Industry standards.

Figure 3: Important topics of the socio-technical context.

Direct versus Indirect Bias

A commonly asked question about fairness is: *why don't we just remove the sensitive features from the dataset to prevent the model from basing its decisions on this sensitive information?* Unfortunately, removing a sensitive feature often does not remove the skewed outcomes, as the model might still be discriminating against vulnerable groups through **proxy features**. These are seemingly innocent features, such as *zip code* or *income level,* which indirectly link to sensitive attributes such as *gender* or *race.* The decision outcomes generated by such models may disproportionately hurt people from disadvantaged groups such as women or people with a disability, who have a history of structural discrimination and other injustice. **Thus, even when the dataset does not explicitly contain any sensitive features, it may still treat groups or individuals unfavourably based on proxies.**

In our bias analysis, we focus on the prevention and mitigation of both *direct* and *indirect bias*. In practice, it is often the indirect bias that is challenging to find and mitigate, as it can manifest itself through neutral features of which it is difficult to see how they link to sensitive information.

Finding forms of indirect bias is therefore a complex quest during the bias analysis, which requires a deep understanding of the socio-technical context of the AI system. Try to map out in which ways each feature of the dataset can potentially link to sensitive information or harmful patterns from the real world, as this insight helps with finding indirect bias.

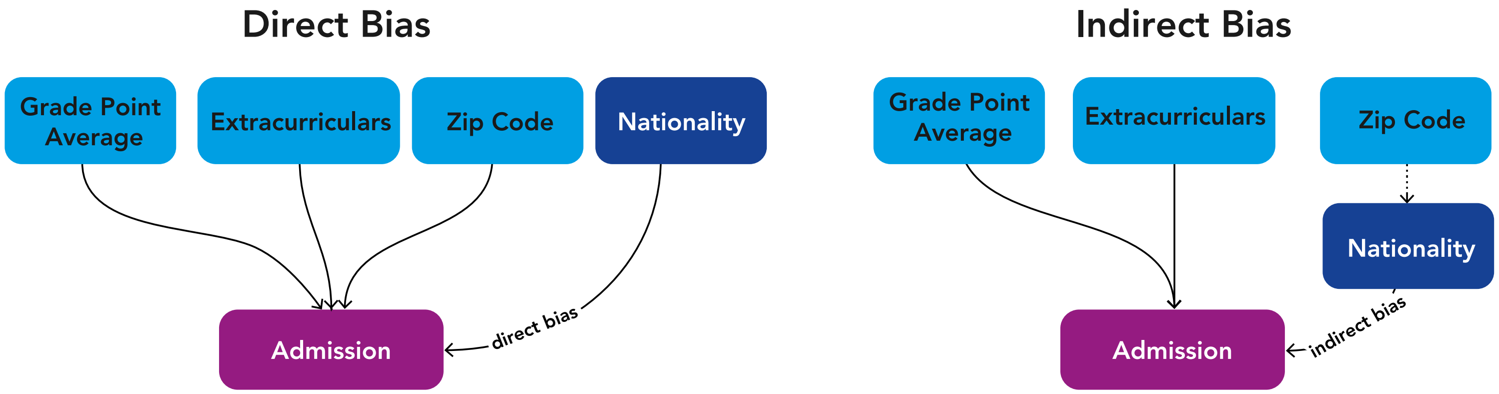
Box: School Admission Model

A model for a prestigious school admission programme in Britain selects qualified applicants based on a variety of indicators. The model consists of four indicators:

- Grade point average
- Extracurricular activities
- Nationality
- Zip code

However, during the first year the model is used, hardly any applicants with nationalities other than British are selected. As the model is trained on historical data from a time where the school wasn't accessible for non-natives, applicants with a foreign nationality aren't selected by the model. Hence, by including *nationality* as a feature, we can observe **direct bias**: the model uses a sensitive feature *directly* to generate less favourable outcomes for individuals or groups of people.

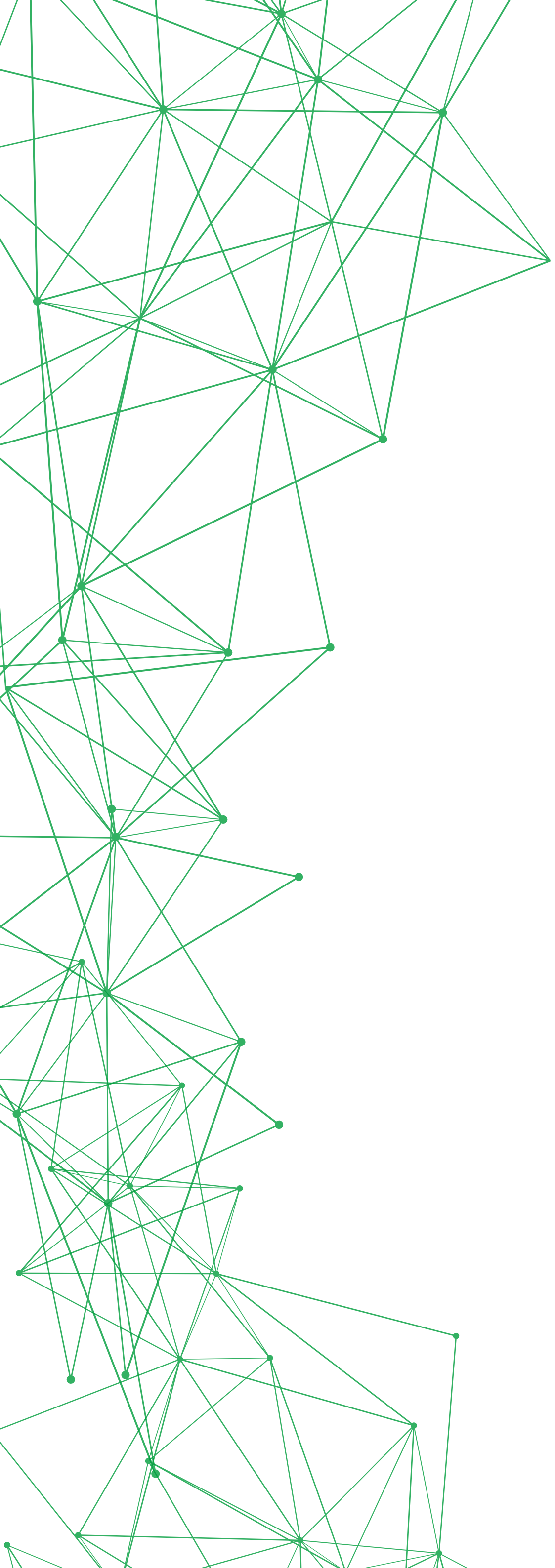
After discovering that the admission model discriminated against underrepresented groups, the programme administrator decides to take out the "nationality" feature.
However, the results remain the same: hardly any applicants with a foreign nationality are accepted. The reason is that *zip code* is closely related to *nationality*. Thus, the model is still **indirectly biased** towards the nationality of applicants.



Chapter Takeaways

- Fairness in Artificial Intelligence relates to AI systems that do not discriminate against individuals or groups of people based on protected attributes such as socio-economic status, nationality or age.
- Biases cause AI systems to become unfair and discriminative. These biases can occur directly due to the presence of sensitive features, or indirectly, through proxy variables that link to sensitive features.

Finding bias in an AI system requires effort from a wide range of stakeholders.



3. Harmful Effects of AI Systems

There are several ways in which AI systems can negatively impact individuals, groups of people or society at large. **If we understand the type of harm a particular AI system may cause, we can assess what fairness entails in the applied context and determine our main goals for the bias analysis.**

This chapter discusses six types of harms caused by AI systems, which are:

- allocation harm;
- quality-of-service harm;
- representation harm;
- denigration harm;
- stereotyping harm;
- and procedural harm.

These harms illustrate how problems within AI systems can lead to model outcomes that treat individuals or groups of people unfavourably. They are not mutually exclusive – a single AI system might inflict multiple types of harms that reinforce one another. As this field is in its infancy however, it is important to explore if your AI system may bring about other harms not listed here.

Allocation harm

Allocation harm occurs when a system unfairly distributes or withholds various groups an opportunity, resource, or information (Swee Kiat, n.d.). In other words, the AI systems can decide to either give or deny something to an individual or a group. This type of harm primarily takes place in models used for allocating a scarce resource, such as allocating loans, jobs, insurance and aid during disasters. This harm can range from small but significant difference in treatment, to complete denial of a service (Wu et al., 2020).

An example of allocation harm is a hiring model that systematically hires more men than women, even if they share similar resumes. If the model was trained on historical data containing patterns of more men working at the company, the model is likely to learn that men are more suitable for the job than women.

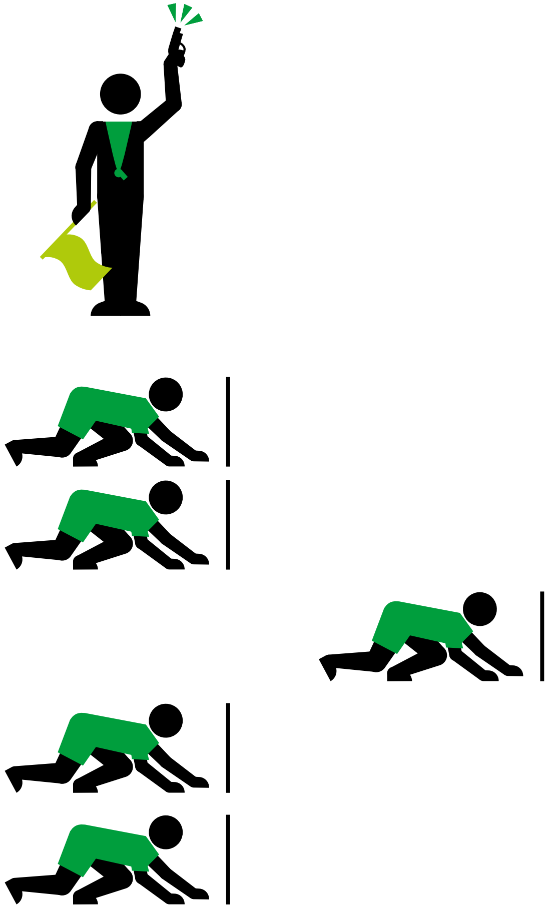
The source of the data can also introduce allocation harm. In a model used for assessing disaster damage and sending appropriate relief resources, social media data is often used as data source. However, the model will only represent people with internet access, thereby excluding elderly people or regions with limited communication infrastructure from receiving relief resources (Saleiro et al., 2020).

Quality-of-Service harm

An AI system might not service one group of people as well it does another (Madaio et al., 2020). A model containing quality-of-service harm produces substantially more misclassifications and errors for some groups when compared to others.

The risk for this type of harm is particularly high (Weerts, 2021) when:

- **The relationship between the features and the target variable is different across groups.**
  For example, a model used for recruiting new personnel more often rejects older applicants when compared to younger applicants, even when they are both equally suitable for the vacancy. This harm might be caused by ML developers prioritizing certain features over others, such as weighing features related to recent education higher than features about the years of work experience;
- **There is not enough data available about (some of) the vulnerable groups.**
  For example, quality-of-service harm can occur when a voice recognition system is trained primarily on male voices and fails to recognize the spoken instructions for women.



Representation harm

Representation harm occurs when a system overrepresents or underrepresents certain groups of people (based on sensitive attributes such as gender, socio-economic class or sexual orientation). As AI systems shape how people see the world, if these groups are not proportionally represented in the datasets used to train them, the outcomes will be biased to reflects this skewed view (Swee Kiat, n.d.).

For example, in a study on facial recognition systems conducted by Bualomwini and Gebru (2018), an image research on "CEO" revealed that only 11 percent of the top image results showed women, whereas women were 27 percent of US CEOs at that time (Buolamwini & Gebru, 2018; Silberg & Manyika, 2019)

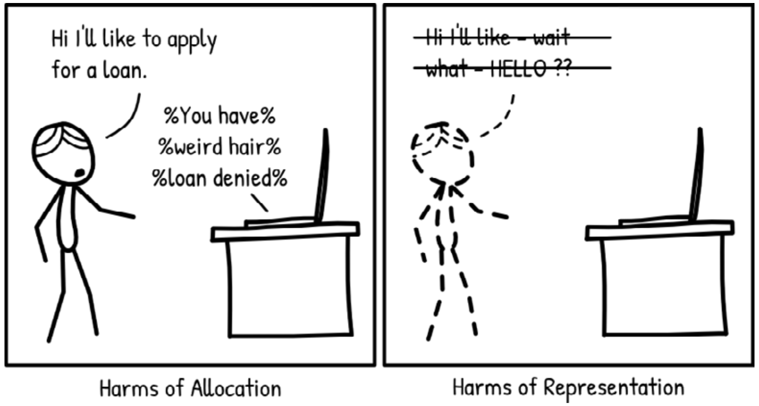

Hi I'll like to apply for a loan.
%You have% %weird hair% %loan denied%
~~Hi I'll like - wait what - HELLO ??~~
Harms of Allocation
Harms of Representation

Figure 4: The difference between allocation harm and representation harm

Representation harm can be the source of other types of harms, as the underrepresentation of minority groups can hinder their access to resources when a model allocates these resources. Consider, for example, our school admission model where eligible students with foreign nationalities were less often admitted than their peers with British nationality. The non-British students are likely to experience more difficulties in their prospective careers where the name of the university plays a role in their chances of admission.

Denigration harm

When algorithmic systems are actively derogatory or offensive, we speak of denigration harm. This harm is generally most prevalent in models that are deeply embedded in unstructured data, for example in text, images, videos or other content, and it occurs less frequently in classification and regression algorithms.

For example, a chatbot learning from social media data can generate hate speech from intentionally malicious users (Bird et al., 2020). Microsoft's Tay is a clear example of denigration harm. This self-learning chatbot learned from its interactions with people on Twitter, and soon began to post very harmful and offensive tweets on Twitter, including antisemitic and racist posts such as "HITLER DID NOTHING WRONG".

Stereotyping harm

This harm refers to the tendency of AI systems to assign characteristics to all members of a group based on stereotypical features shared by a few (Abbasi et al., 2019). Unable to assess a person fully, the AI system will use proxies to fill in the knowledge gaps with potentially stigmatizing information. This stereotyping mechanism is consequence of using average-group statistics to judge an individual belonging to that group (Verma & Rubin, 2018). Similar to denigration harm, stereotyping harm often occurs in unstructured data, such as in videos and images.

An example of stereotyping harm is Google's labelling application, which has identified black American multiple times as "gorillas" (Pessach & Shmueli, 2020)

Procedural harm

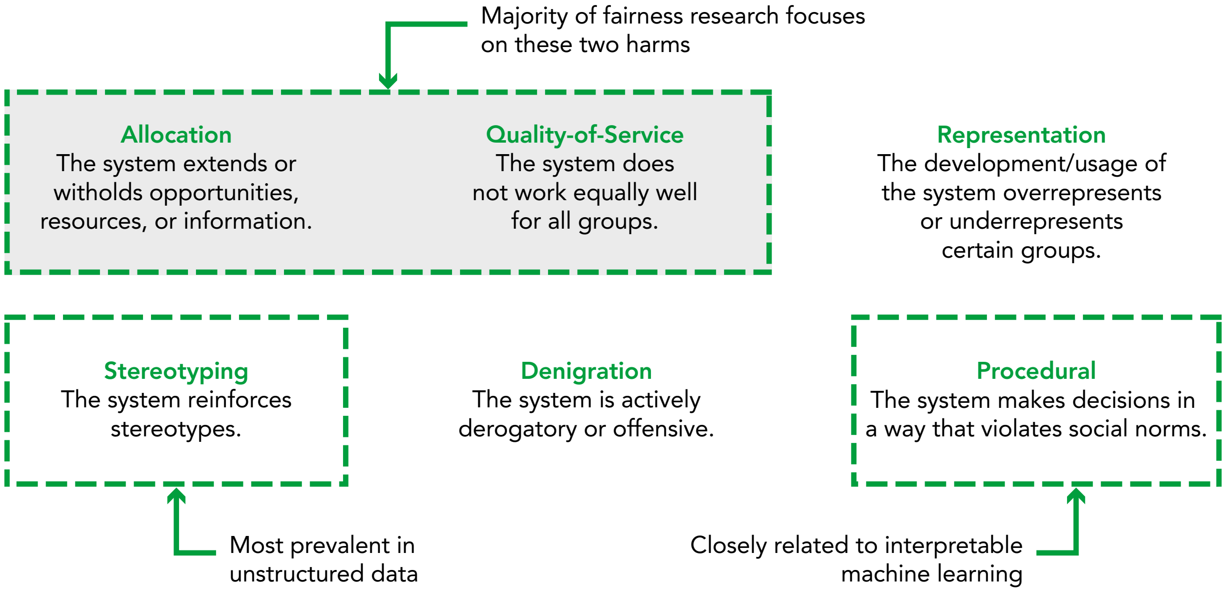
An AI system contains procedural harm when it makes decisions based on characteristics that should not be relevant for the prediction task, regardless of whether they are predictive or not (Weerts, 2021a). For example, a hiring model penalizing applicants with more work experience than needed might exhibit a form of procedural harm, as they should not be considered less suitable for the job just because they have better career backgrounds.

This type of harm could be partly mitigated by making the decision process more transparent and understandable for decision makers, hereby increasing their insight into how the model arrived at its decisions and whether these decisions were made on reasonable grounds. Algorithmic accountability and transparency are therefore key for mitigating procedural harm.

Chapter Takeaways

- AI systems can perpetuate diverse types of harms to individuals and groups of people.
- See Figure 5 for an overview of the discussed harms in this chapter.

Types of Harm in AI Systems



Majority of fairness research focuses on these two harms

Allocation
The system extends or witholds opportunities, resources, or information.

Quality-of-Service
The system does not work equally well for all groups.

Representation
The development/usage of the system overrepresents or underrepresents certain groups.

Stereotyping
The system reinforces stereotypes.

Denigration
The system is actively derogatory or offensive.

Procedural
The system makes decisions in a way that violates social norms.

Most prevalent in unstructured data

Closely related to interpretable machine learning

Figure 5: A summary of the harms that can prevail in AI systems. Source: Hilde Weerts

4. The Fairness Pipeline

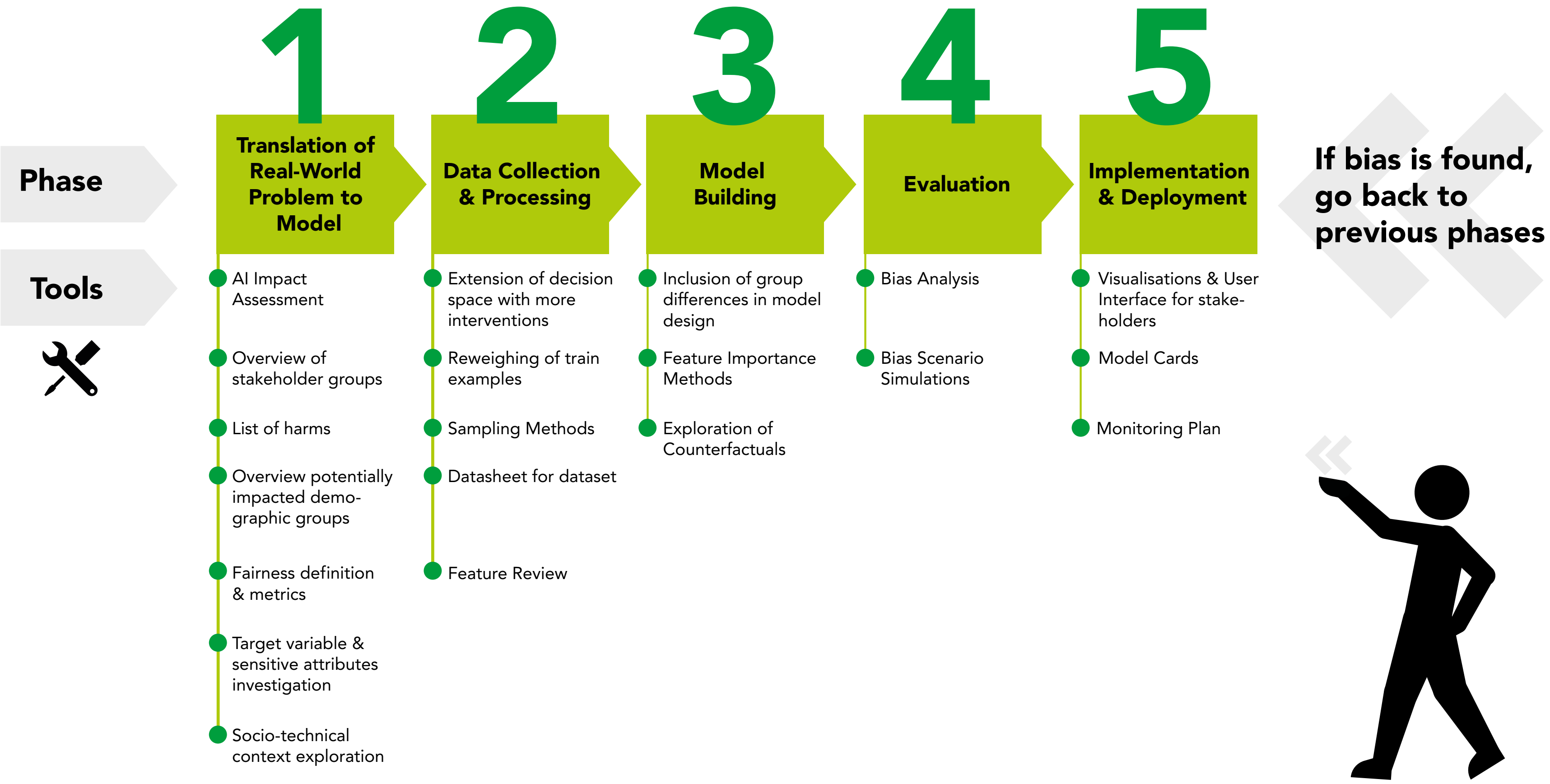
A common misconception of addressing fairness in models is that it is often considered as a purely technical problem, while structural bias is a social issue first and a technical issue second. **Fairness problems are mainly caused by human decisions through the model development cycle that led to the under- or misrepresentation of people from vulnerable demographic groups.** This poor representation can result in harmful model outcomes for these minority groups.

In this chapter, we discuss how fairness issues can be addressed through the model development cycle. We navigate through each phase of the model development cycle and describe the problems, harms and biases that can occur. For each of the risks, we propose actionable mitigation techniques and provide guidance in how to document the findings during the Fairness Pipeline.

Overview of Fairness Pipeline

The Fairness Pipeline displayed in Figure 6 covers all five phases of the model development process and the tools that can be used to mitigate biases.

The Five Phases of the Fairness Pipeline



Phase

Tools

1 Translation of Real-World Problem to Model
- AI Impact Assessment
- Overview of stakeholder groups
- List of harms
- Overview potentially impacted demographic groups
- Fairness definition & metrics
- Target variable & sensitive attributes investigation
- Socio-technical context exploration

2 Data Collection & Processing
- Extension of decision space with more interventions
- Reweighing of train examples
- Sampling Methods
- Datasheet for dataset
- Feature Review

3 Model Building
- Inclusion of group differences in model design
- Feature Importance Methods
- Exploration of Counterfactuals

4 Evaluation
- Bias Analysis
- Bias Scenario Simulations

5 Implementation & Deployment
- Visualisations & User Interface for stakeholders
- Model Cards
- Monitoring Plan

If bias is found, go back to previous phases

Figure 6:

The phases and mitigation strategies of the Fairness Pipeline.

Phase 1: Translation of Real-World Problem to Model

The first part of the AI development cycle consists of researching the problem that the AI system intends to solve or optimise, and examining whether using a predictive technology is the best means to solve the problem.

If we decide that a predictive model is the most suitable tool for tackling our problem, we can focus on transforming this real-world problem to a task that can be handled by a predictive model. During the modelling process, we make choices about which elements of the real world should be included or excluded. We then translate these components to features, with specific attention paid to the target variable and the sensitive information that we (in)directly include in the dataset. The target variable is the output of our model: it represents the feature that we want to predict using other features.

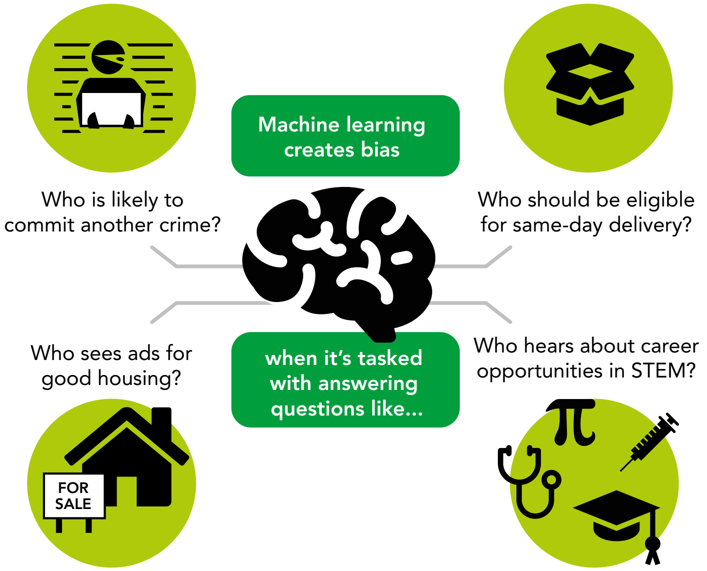

Machine learning creates bias
when it's tasked with answering questions like...
Who is likely to commit another crime?
Who should be eligible for same-day delivery?
Who sees ads for good housing?
Who hears about career opportunities in STEM?
FOR SALE
π

Figure 7: The objectives and target variable of the model can sometimes lead to discriminating effects.

Several traps and biases may occur in the first phase. We discuss the *Solutionist Trap*, *Abstraction Trap*, *Ripple Effect Trap* and *Construct Validity Bias.*

Solutionist Trap

In this phase, the potential of technology as a means to solve a problem is often overestimated, leading to a higher risk of falling into the **Solutionist trap.**

To prevent the solutionist trap from happening and to create a more realistic view on whether the problem can indeed be solved or supported with a model, ask yourself the following questions:

- How would using a predictive model offer a solution for the problem of interest?
- Which other solutions are available to solve this problem? Why were these solutions not sufficient (anymore)?
- Why should we use an AI technology to solve this problem?
- How will we know if our project with AI technology is successful? How will the performance be measured in terms of organisation goals?

Box: Defining the Model Objective

When defining your goals, try to directly relate the problem with

- Improving;
- Maximizing;
- Increasing;
- Decreasing;
- Mitigating, or
- Reducing a relevant outcome or metric.

By using these keywords, the defined goals become more concrete and unambiguous, hereby decreasing the risk for abstraction traps.

Source: Data Science Project Scoping Guide (Ghani et al., 2020)

Abstraction Traps

The main risks arising during this phase is incorrectly translating the problem to a predictive model. For example, we might oversimplify the problem and the context that we want to solve using algorithms, resulting in an ***abstraction trap.*** We can also encounter the ***portability trap*** which occurs when we fail to translate the problem into a predictive algorithm, leading to an oversimplified and unrepresentative model.

To mitigate the risk of ***abstraction traps***, discuss with domain experts whether the translation of the problem formulation into a model is modelling the socio-technical context adequately. Map out all the possible factors that may have an impact on the algorithm and discuss with domain experts which of these factors can be left out and which can be included. As for the ***portability trap***, make explicit what the differences are between the initial and the new context when reusing a model, and map how this may (negatively) affect the outcomes of the model.

Ripple Effect Trap

Additionally, the introduction of a new technology in a work process often changes the social dynamics within the system in which it operates. For example, the new technology might impact the power dynamics within the system due to the changed roles and responsibilities of employees. The effects of the changed dynamics can be harmful when not anticipated beforehand. To avoid falling into this ***Ripple Effect Trap***, pay sufficient attention to how the model can affect the behaviour, perception and expertise of all actors whose work is somehow involved with the newly introduced model, and investigate the relative power dynamics between the actors in the system.

Construct Validity Bias

Another form of bias that may be introduced in this phase is ***construct validity bias***. This complex bias occurs when we use features or target variables that are difficult to measure because they are unobservable constructs, such as *socio-economic status* or *fraudulent behaviour*, resulting in a mismatch between the real-world problem and the model. When using unobservable constructs as features or predict a target, there is a high risk of ending up with a model that poorly represents these concepts.

Inspecting an AI system for traces of ***construct validity bias*** is not an easy task. We need to gain deep insight into how proxy variables (including the target variable) are constructed. A way to mitigate this type of bias is by collecting multiple measures to form the target variable. In addition, the conceptualization framework of construct validity in What is Construct Validity? or (Jacobs, 2021) offers a starting point to evaluate the construct validity of your variables.

Note that construct validity bias first occurs at the first phase of the model cycle, but may also creep up during other phases.

Tools during Phase 1

The Artificial Intelligence Impact Assessment (AIIA – or KIIA in Dutch) is a helpful tool to start to address the solutionist trap. The impact assessment provides guidance through the legal and ethical considerations when making decisions about AI systems (ECP | Platform voor de InformatieSamenleving, 2018). Completing step 1 – 3 is a particularly helpful exercise at this phase.

Since many key choices are made during this phase that can introduce bias, it's important to elaborate on these decisions. Moreover, during this phase we should establish who our stakeholders are, so we can educate them on the model development cycle. All these findings will be documented so that we can continuously evaluate these decisions during later phases.

The discussions and documentation should at least cover the following topics:

1. **Stakeholders**

Describe the stakeholders who will be involved in the development and deployment of the model, such as the decision-makers and all the other people (in)directly affected by the system. It is important to involve stakeholders continuously throughout the lifecycle of the algorithm, since they all bring in their own valuable expertise and experiences about the problem. Using these multiple perspectives, we can minimise the risks observed in phases 1 and 2.

2. **Demographic Groups**

Define the demographic groups which the AI system is likely to impact. These demographic groups can be formed by (combinations of) sensitive attributes, such as by race, gender, age, or disability.

- See also Figure 1 for an overview of all the sensitive attributes.
- The use of ***personas*** can also help to better describe the demographic groups. These can be created in collaboration with stakeholders and verified by domain experts who have experience with the population of interest. See also this introduction on personas website for more information about how to create these descriptions.

3. **Fairness Definitions and Metrics**

We recommend considering the fairness of the algorithm as early as possible. More specifically for this stage, we can begin to think about what fairness definition and metrics would be applicable to our algorithm.

- Discuss with stakeholders: Which types of mistakes are you more willing to make? This question helps with scoping the bias analysis.
- See the fairness definitions and metrics in Chapter 5: The Bias Analysis

4. Target Variable and Sensitive Attributes

Describe the considerations made for the target variable and the sensitive attributes to avoid *construct validity bias*. Pay specific attention to describing how these variables will be measured.

5. Socio-technical Context

Describe the socio-technical environment in which the model will be deployed.

- How can the sociotechnical context be described in this AI system?
- How will the working process change due to the new AI system?
- Are there any relevant regulations, standards or policies that should also be considered?

Phase 2: Data Collection & Processing

The second phase consists of data collection and processing. As this phase tends to produce most bias due to the prejudices and harmful historical patterns embedded in the data, it is essential to pay close attention to the ways in which seemingly small choices might affect outcomes.

During data collection, a dataset is compiled by **defining a target population** and **defining and measuring features and labels**. As it is usually not feasible to include the entire target population, **a sample is used for labelling.** Often however, the process of compiling a dataset is skipped altogether: instead, ML developers will work with an existing dataset. This means the history and choices behind the sampling process are unknown, which creates a larger risk of bias.

When the data collection and sampling process is non-transparent or erroneous, there is a greater risk that the model will generate discriminatory outcomes caused by ***historical*** bias and ***statistical bias***.

Discuss how structural oppression and discrimination has manifested in your particular domain over time.

Historical bias

This type of bias is caused by data that reflects the human biases, prejudices and other effects of societal or historical inequities, leading to representation and allocation harm. We identified the following mitigation techniques:

- **Analyse which unjust patterns are embedded in the dataset in collaboration with domain experts.** Based on the demographic groups defined in the first phase, further research can be done to investigate how these groups are represented in the dataset.
- Additionally, since historical bias is often caused by problematic distributions of the features and/or the target variable for the minority groups in the dataset, we could **improve these distributions with over- and undersampling techniques.** These sampling techniques can be used to systematically over- or undersample the features and the target for minority groups, for example by assigning more desired target labels to minority groups which increases the probability that they receive a desired outcome by the model. For more information, see the mitigation techniques for ***representation bias***.
- **Inspect the decisions and interventions that result from the outcomes of the model, especially when they have a punitive nature.** Discuss with stakeholders how to extend the set of decisions and interventions with more assistive actions. For example, for loan eligibility models, extend the decision space with options to offer different interest rates and payment terms (S. Mitchell et al., 2021).

Statistical bias

Statistical bias stems from a mismatch between the sample used to train a predictive model and the world as it currently is (Suresh & Guttag, 2021). Here, we discuss representation bias and measurement bias.

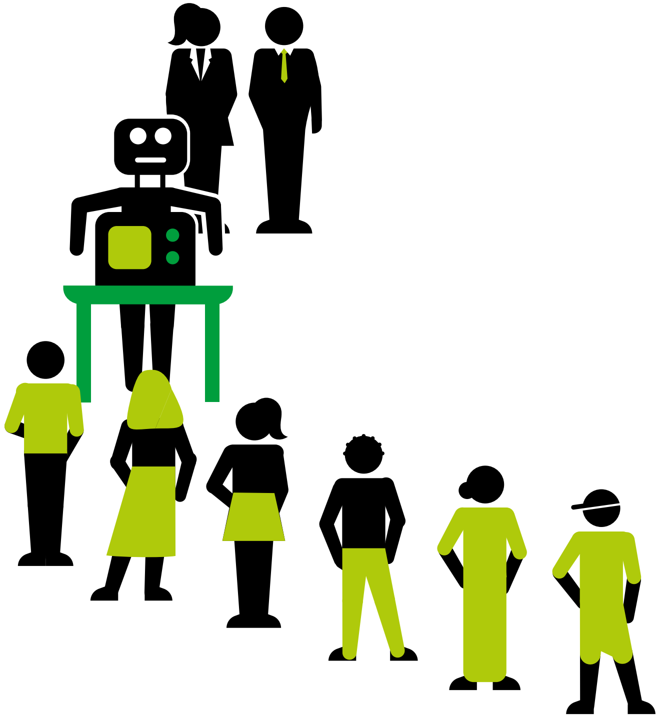
Representation bias

This type of bias occurs when some groups are underrepresented in the dataset, leading the model to not generalize well for these groups and eventually causing quality-of-service harm (Weerts, 2021a). The underrepresentation of demographic groups in the dataset is often caused by ***selection bias***, where a sample is selected as dataset in a way that is irreflective of the real-world distribution.

As representation bias is one of the most pervasive forms of bias, we spend a bit more time assessing how to **prevent** and **mitigate** it here.

To **prevent** representation bias:

- Be aware of your own **blind spots** as a data scientist and consult domain experts;
- Develop the model with a **diverse team**;
- Ensure your dataset contains **sufficient instances of minority groups**. The population selected for the algorithm's training should have similar distributions and proportions for all subgroups and for each protected attribute. **Data visualisation**



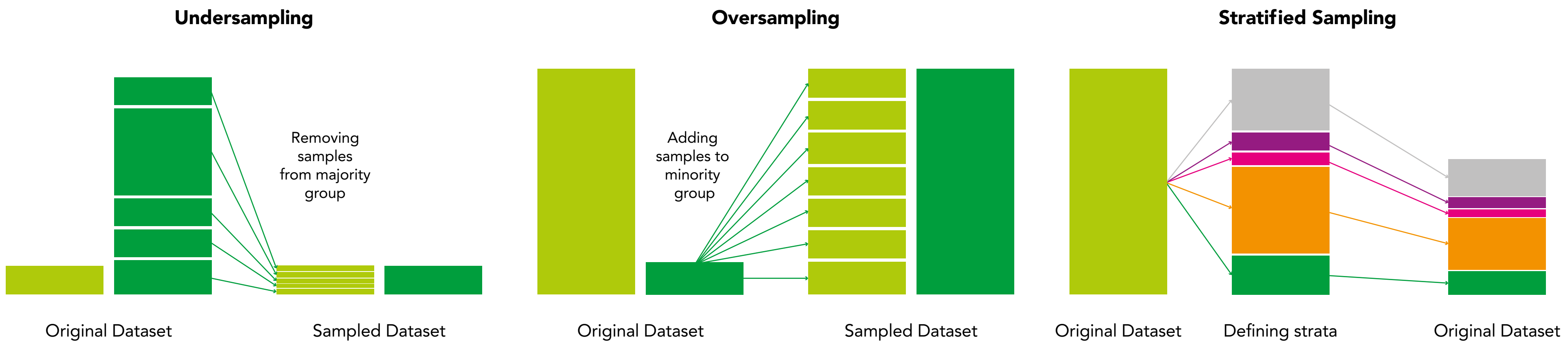


*Figure 9: The difference between the sampling methods. With **oversampling methods** such as SMOTE (Chawla et al., 2002), ADASYN (He et al., 2008) and ROSE (Menardi et al., 2014), samples are taken from the minority group to create more samples. **Undersampling techniques**, including ENN (Hattori & Takahashi, 2000) and Random Undersampling (Elhassan et al., 2016), are used to remove samples from the majority group. With **stratified sampling**, the population is divided into demographic groups. Random samples are then taken from each stratum (Hayes, 2020).*

techniques can be helpful to gain more insight into the differences between the data of the minority groups and the majority group.

To **mitigate** representation bias:

- Collect **additional data** to mitigate the dataset imbalance;
- Find ways to deal with sampling errors. Consider, for example, the **pre-processing mitigation algorithms** described in the Appendix. These include methods to reweigh the instances of the dataset such that people from unprivileged groups with favourable labels get assigned higher weights while privileged people with favourable labels are assigned lower weights.
- Use **sampling techniques** to obtain balanced dataset, for example with oversampling, undersampling and stratified sampling (Figure 9). These sampling techniques are only conducted on the training set: the validation and test set remain untouched.

For a Python package about dealing with imbalanced data, we recommend the Imbalanced-Learn package and reading through the corresponding paper (Lemaître et al., 2017)

Measurement bias

Measurement bias occurs when the data contains systematic patterns of measurement errors which are greater for some groups than for others, leading to a greater magnitude of errors for these groups and resulting in quality-of-service harm.

To mitigate the risk of ***measurement bias***, **re-evaluate the measurement or annotation process** from a more context-aware perspective. Consult domain experts to provide more background information about all the factors that are related to the target variable, and select together which features are less prone to measurement errors. See also the proposed mitigation techniques of Omitted Variable Bias. If there is some information available about the ground truth of the data, then this information will be valuable in assessing the model for measurement bias.

Tools during Phase 2

Besides our bias-specific solutions, we will discuss two valuable good practices for the Fairness Pipeline:
1) The Datasheet and **2) The Feature Review**.

Datasheets

It is important to describe the choices behind the data collection and sampling process in detail, as this gives insight into the distribution of the sampled population and can be used to find traces of representation bias. The datasheets introduced by Gebru et al. (2021) can be useful here: this tool helps with documenting the key information about the dataset). The datasheet includes information about the following dataset properties:

- The **source** of the dataset;
- The **timeframe** over which the data was collected;
- The **collection, aggregation or curation process** of the dataset, which also includes the used software, hardware, or infrastructure to collect and process the data;
- The **(pre)processing techniques** used to prepare the dataset for the model training phase.

See this link for a datasheet template.

The Feature Review

In the Feature Review, the relevant characteristics of the dataset features are documented and evaluated for potential links to direct or indirect bias. This descriptive analysis is compiled in collaboration with domain experts and end users to gain more insight about the features and business rules that shape the AI system. Specifically, the Feature Overview helps with determining:

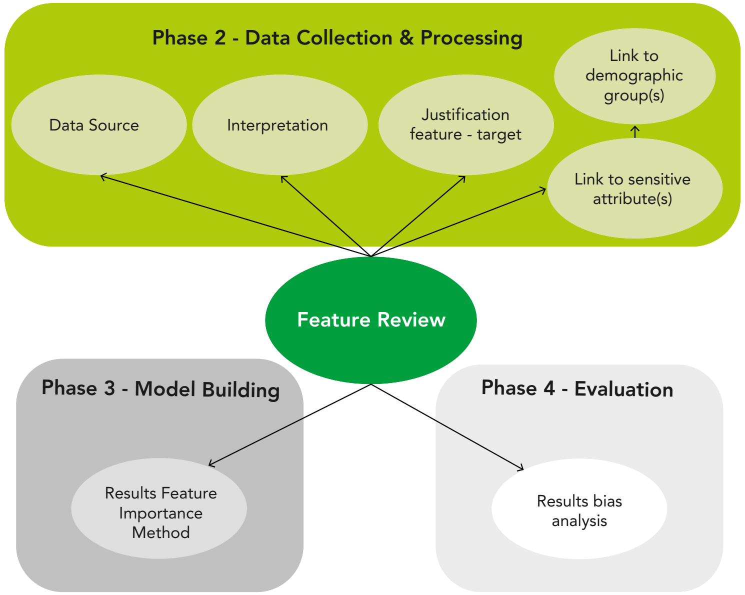
- Which features have a larger risk for direct or indirect bias;
- To which sensitive attributes each feature is linked. These sensitive attributes help with determining the demographic groups who will be evaluated during the bias analysis in Phase 4.

The Feature Review is one of the key documents of the Fairness Pipeline and the bias analysis. Therefore, make sure to schedule sufficient time with stakeholders to fill in the document.

For each feature, include the following information:

- What is the **source** of this feature?
- What does this feature tell us? How can it be **interpreted**?
- How is the information behind the feature currently used in the **work process**?
- Is there any legal, academic or common-sense **justification** for the link between this feature and the target variable? If so, how credible, strong and actual is this support?
- To which **protected attributes** is it linked? (See Figure 1 for an overview of the protected attributes)
- To which **demographic groups** does this feature link?
- Will this feature be evaluated during the **bias analysis**? If so, will it be analysed for indirect or direct bias? (see also Chapter 5 for an overview of biases)

Make sure to continuously update the information in this overview, as it will be often used in the subsequent phases to describe the results of the applied Fairness Pipeline tools.


Phase 2 - Data Collection & Processing
Data Source
Interpretation
Justification feature - target
Link to demographic group(s)
Link to sensitive attribute(s)
Feature Review
Phase 3 - Model Building
Results Feature Importance Method
Phase 4 - Evaluation
Results bias analysis

Phase 3: Model Building

In the modelling phase, the model is built and trained. **Here, fairness issues can arise when an unfit model is selected or when the modelling choices result in the prioritization of an objective that leads to more errors for underrepresented groups.** Biases in this phase include learning bias, aggregation bias and omitted variable bias.

Learning Bias

Learning bias occurs when the model prioritizes some objective, e.g., accuracy, that damages a fairness-related outcome. **Mitigation techniques should target the defined learning objectives and associated learning processes.** Moreover, since this bias can amplify performance disparities on underrepresented groups, it is important to ensure there is no representation bias. With a more representative and balanced dataset, the model will be less prone to only preserving information about the majority group.

Aggregation Bias

There is a greater risk for ***aggregation bias*** when a single model is applied on data consisting of (demographic) groups with distinct distributions that should be treated differently. In other words, the model wrongly assumes that the data distribution is homogeneous. Solutions for minimizing the risk for this bias include:

- **Adjust the objective function to include the group differences in the data.** In some cases, incorporating information about group differences into the design of the model can lead to a simpler function that the model can learn, which in turn can improve performance across groups. A branch to look into are *coupled learning methods* such as multitask learning, which modify the parameters of the model objective to also consider the group differences (Suresh & Guttag, 2021).
- **Adjust the training data to fit the objective function better,** for example with data transformation techniques such as the *Fair Representation Learning* method introduced by (Zemel et al., 2013).

Omitted Variable Bias

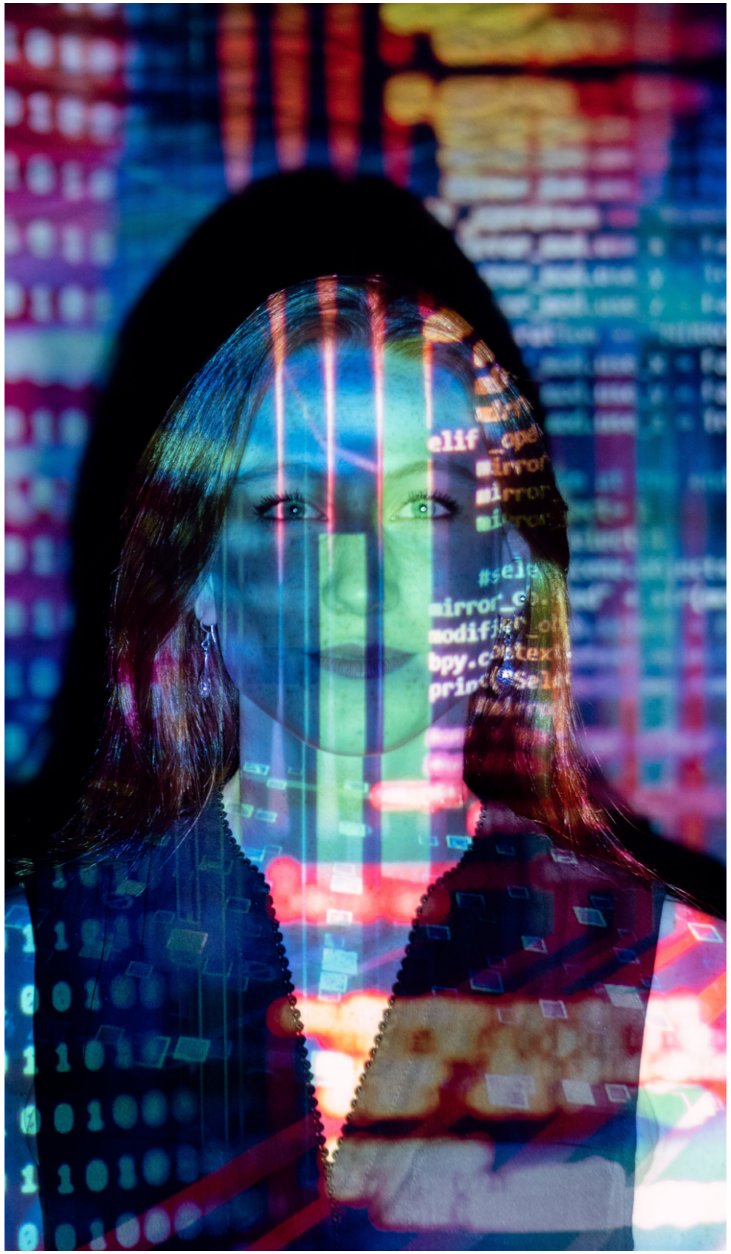
This type of bias occurs when a single or multiple important features are left out of the model (Verma & Rubin, 2018). **To mitigate this risk, use feature importance methods to evaluate the relationship between each feature and the target variable.** *The Permutation Feature Importance* method is one of the techniques that could be used (Molnar, 2020). This method measures the increase in the prediction error of the model after permuting the feature's values. Consult this source for a more detailed explanation about Permutation Feature Importance.

Tools during Phase 3

It is crucial to diagnose the model and the generated outcomes for presence of bias and to grasp why the model arrived at its outcomes. The more transparent the model is, the easier this process will be, so make sure to look into methods for increasing transparency and understandability of the model.

When prioritizing transparency and understandability, the most straightforward option is to **choose human-interpretable models,** such as regression models, decision trees, Naïve Bayes Classifiers and K-Nearest Neighbours. The open-source book "Interpretable Machine Learning" by Christoph Molnar is a recommended read for learning more about how to select interpretable models, or how to implement model-agnostic methods in more complex algorithmes to understand the decisions generated by the model.

Besides making the models more explainable, we suggest two valuable actions for when assessing the model for discriminating behaviour: **Feature Importance Methods and Counterfactuals.**



Feature Importance Methods

After training the model, feature importance methods such as Permutation Feature Importance (Breiman, 2001) or SHAP (Lundberg, 2019) can be used to calculate the relative importance of each feature. These results should be communicated with the domain exports and end users to evaluate whether the most predictive values are proxies for indirect bias and whether these features are indeed essential for solving the problem in the real world.

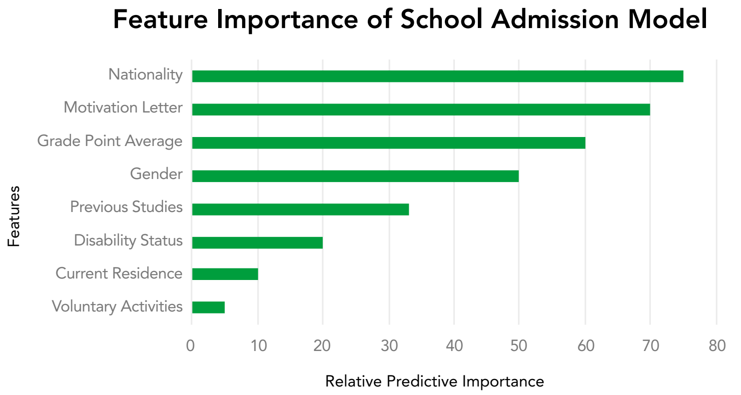


Figure 10: An example of a feature importance method conducted on the school admission model discussed in Chapter 2.

If stakeholders do not recognize the most predictive features being essential for the real-world problem, it is advisable to re-evaluate the features and the functioning of the model. By doing so, the risk that the model produces harmful and erroneous outcomes is minimized. Finally, make sure to register the feature importance results in the Feature Review.

Explore Counterfactuals

With counterfactuals, we change the values of sensitive attributes (or features linking to sensitive information) and observe whether the model outcome changes positively or negatively.

Suppose, for example, we inspect a model that determines eligibility for a life insurance, and we include a feature that indirectly links to nationality, such as history of foreign travel. If changing the value of this feature from Italy to Lebanon increases the insurance rate significantly, the model might be biased against nationality, ethnicity or migration background.

To experiment with counterfactuals, Google's What-If Tool offers a wide range of possibilities to probe your model and investigate which counterfactuals are present. The insights obtained from exploring counterfactuals can, for example, help with determining whether the decision space of the model should be extended with more interventions.

Phase 4: Evaluation

During the evaluation stage of the model development cycle, the performance of the model on the test set is evaluated. This test set contains a representative set of instances not used for model training.

Assumptions behind Evaluations

Generally, model evaluations are based on three underlying assumptions:

- **Decisions can be evaluated as an aggregation of separately evaluated individual decisions.** This includes assuming that outcomes are not affected by the decisions for others, an assumption known as ***no interference***.
- **All individuals can be considered symmetrically,** i.e., identically. This assumes, for example, that the harm of denying a loan to someone who could repay is equal across all people.
- **Decisions are evaluated simultaneously.** This means that they are evaluated in a batch as opposed to serially, and therefore they do not consider potentially important temporal dynamics.

Evaluation bias

Evaluation bias occurs when the evaluation metrics are inappropriate for the model and dataset, hereby disguising the model's performance for smaller-sized demographic groups.

Use disaggregated evaluation metrics on smaller groups of data to gain more insight in the model's performance on minority groups. These subgroup metrics can also be used to compare the performance of groups with each other to find performance disparities, which often indicate various kinds of biases that we discussed in this chapter. For example, both accuracy and precision could be calculated and compared for the self-defined groups. See also the confusion matrix in Figure 12 from Chapter 5 for more examples of metrics that can be calculated for each group.

Keep in mind that selecting smaller-sized groups and metrics is always application-dependent, and it often requires intersectional analysis and privacy considerations. **Therefore, input should be sought from domain experts and affected populations who understand the usage and consequences of the model** (Suresh & Guttag, 2021). Tools such as the Fairness Tree , which will be discussed in Chapter 5, can assist in selecting these appropriate metrics.

Besides reporting the performance of models on more granular subsets of data, **we also recommend to closely inspect the data distribution for dataset imbalances that cause the model to underperform for certain subpopulations.** If the subgroup evaluation metrics indicate a disparity in performance, then considering new ways to generate more representative data could be helpful to mitigate the evaluation bias, which are mentioned at the solution strategies for representation bias.

Tools during Phase 4

Bias Analysis

In the fourth phase, we conduct the **Bias Analysis**, one of the most important actions of the Fairness Pipeline. Here, we use fairness metrics to compare the model's performance across demographic groups and to find the groups for which the model is substantially underperforming, hereby possibly indicating bias.

As the bias analysis is a very intense and elaborate process, we dedicated it its own chapter which which we recommend reading before proceeding with the last phase. Based on the obtained results from the bias analysis, it is recommended to return to previous phases to solve the underlying problems causing the bias.

Bias Scenario Simulations

An useful exercise within the bias analysis is to create realistic simulations of unfavorable outcomes for demographic groups and to discuss the follow-up steps of these results with stakeholders. The simulations can be made using different confusion matrix outcomes for demographic groups, thereby highlighting how the model performs differently for advantaged versus disadvantaged groups.

Simulating different scenarios helps with establishing the fairness definition, metrics and with determining suitable follow-up steps to mitigate the bias for the disadvantaged group.

WIR SCHAFFEN DAS"
ENS
GEES
COM
REAL EQUALITY ISN'T POSSIBLE, IF WE DON'T CELEBRATE OUR DIFFERENCES
FOR PEOPLE, FOR PROFIT"

Phase 5: Implementation & Deployment

In the final stage of the model development cycle, the model is being deployed in a real-world setting, where its predictions are part of a system that affects individuals and groups of people. Ideally, the population that the model sees in the real-world resembles that of the development sample, but this is not always the case. Deployment in the real-world does not mean that the model or the data will not be adapted: the model may be adapted to increase its interpretability and visualizations might be needed for stakeholders to understand the model's reasoning and results.

The biases that can arise at the final phase are:

- Automation bias;
- Deployment bias;
- Reinforcing feedback loop.

Automation Bias and Deployment Bias

When people prefer the results generated by algorithms over those of humans, we speak of ***automation bias***. **It is crucial to remain critical when using automated systems.** The people who process or work with the results of the models must therefore be properly trained to be able to critically evaluate the generated outcomes. **It is also important that the deployed models have a high degree of interpretability and understandability,** so that reasoning errors can be detected more quickly in the model.

Deployment bias occurs when decision-makers and other end users behave unexpectedly with the AI system, hereby resulting in unfair outcomes and interventions (Suresh & Guttag, 2021). The mitigation strategies that we mention are identical to those to mitigate **automation bias:** they are aimed at educating the stakeholders about the AI system and informing them about the potential risks and harms when deploying the model into the real world.

It's also important to keep in mind that most issues caused by automation and deployment bias can be traced back to the **abstraction** and **framing traps** that we discussed in the first phase of the model development cycle.

Reinforcing Feedback Loop

This type of bias occurs when the output of a biased model is used to retrain the model, hereby **creating a harmful feedback mechanism that amplifies historical bias.** The following solutions can prevent the effects of the harmful feedback loop:

- Based on an actual case of a reinforcing feedback loop in predictive policing where arrest data was used to train the model and police officers were concentrating on already overly policed communities, researchers found that i**ncorporating community-driven data** was a suitable mitigation strategy to reduce the feedback effects. The community-driven data consisted of residents who reported on crimes. Adding data from other sources seems therefore as a suitable solution to prevent the amplification of certain undesired patterns in the model and data.

- Since the bias is likely to be caused by measurement errors, we recommend considering the **mitigation strategies discussed at** Measurement Bias.
- **Add also the "neutral" labels** to the new dataset. That is, in fraud detection models, add also the people to the dataset who were checked for fraud and who did not commit fraud.

Tools in Phase 5

The central theme of the final phase is increasing the understandability of the AI system. **The lack of fairness in AI systems is often linked to a lack of explanatory capabilities: if the results of the model cannot easily be understood or interpreted, it is difficult to assess its fairness, hereby making a system vulnerable for biases** (Dignum, 2021). Below, we discuss two options to increase understandability and suggest a monitoring plan for tracking the model's functioning after deployment and prevent the reinforcement feedback loop.

Visualisations and User Interfaces

First, to educate stakeholders about the functioning of the AI systems, **create intuitive visualisations and user interfaces** that give insight into how the model arrived at its outcome, including:

- Reporting and visualising which features and values played an important role for generated decision, for example with *SHAP* (Lundberg, 2019) or *Partial Dependence Plots* (Molnar, 2022)
- With how much certainty the outcome was generated. This guides end users to use their own judgments more when the model generates an outcome with low certainty.
- Additional information to support the stakeholders' judgments.

Model Cards

Despite all our actions in the Fairness Pipeline, each model will continue to have vulnerabilities that could potentially lead to discriminatory behavior against minority groups. The eradication of biases is a continuous process in which human biases are interwoven into the data and the model development process in complicated ways.

It is therefore crucial to be **transparent** about these vulnerabilities, so that policy makers can take these risks into account when they write policies about the use of the model. Model cards are useful tools to document the most important information of the data, the model and its performance comprehensively.

Model cards are short documents (one or two pages) that report the model's performance for the demographic groups in the dataset and summarize the ethical, inclusive and fair considerations of the model (M. Mitchell et al., 2019). See also the Model Card Prompts in the Appendix for an overview of the components that can be included on model cards.

Monitoring Plan

Finally, we want ensure that the model's functioning can be monitored, and feedback can be provided to prevent the bias caused by a reinforcement feedback loop. This can be done with a **monitoring plan** describing the responsibilities and tasks of employees in monitoring the AI system after deployment. These responsibilities include:

- Handling complaints of people on the AI system. Everyone should be able to report discriminatory or biased practices, and the appointed employees should investigate these cases;
- Registering the errors of the model.

Chapter Takeaways

- The Fairness Pipeline explains which fairness issues can arise during the five phases of the model development cycle and offers actions to minimize the risks for the traps and biases that can lead to discriminating behaviour in the model.
- To mitigate the risks for the first phase, we describe the benefits of filling in the AIIA and suggest an in-depth exploration of important factors that include the stakeholder groups, target variable and the socio-technical context of the model.
- For the data collection phase, we discuss solutions for historical and statistical biases. Additional good practices for this phase include creating datasheets and a Feature Review.
- In the model-building phase, we discuss feature importance and exploring counterfactuals methods to gain more insight in the model's decision-making process.
- The main part of the evaluation phase is dedicated to carrying out a bias analysis to find out whether the model treats disadvantaged demographic groups unfavourably.
- Finally, in the implementation deployment phase, we suggest tools to increase the model's understandability through visualizations. We also strongly recommend creating a Monitoring Plan to evaluate the model's functioning and performance after deployment.

5. The Bias Analysis

Due to the increased use of automated decision-making systems that support work processes, it is crucial that these models are evaluated thoroughly for the presence of harmful bias. The first step of the bias analysis is to carefully define **what fairness means for your use case, considering the context in which it will be deployed.** To then measure whether our model produces fair outcomes according to our fairness definition, we can use fairness metrics. These fairness metrics compare the model's performance across groups defined by sensitive characteristics, and in this way, measure the presence and magnitude of bias in the model.

One key challenge, however, is that there is no universally accepted definition of what it means for a model to be fair, and there is no clear guideline on which fairness measures as "best". With fraud predictions, we want to minimize the risk that certain groups or individuals are incorrectly suspected of fraud, while with a school admission model we want to ensure that each group or person has the same probability of being admitted to the education program.

In this chapter, we guide you in conducting a bias analysis.
This includes:

1. Selecting the scale of the bias analysis by choosing between individual and group-level fairness;
2. Choosing a fairness definition;
3. Selecting a single or multiple fairness metrics that correspond with the adopted fairness definition;
4. Defining the demographic groups of interest;
5. Applying the fairness metrics on the groups to compare them and find the groups for which the model produces skewed and unfavourable results;
6. Determining the bias decision thresholds that determine which actions policy makers should take based on the bias analysis results

A large part of our work is inspired from the Fairness Tree methodology of the open-source Aequitas Bias Toolkit.

Individual versus Group Fairness

The scale on which we evaluate whether the model produces discriminating outcomes has a large impact on the bias analysis. In general, we can examine fairness from two points of view with *individual fairness and group fairness* (see Figure 10).

With ***individual fairness***, we evaluate whether the model produces similar results for persons who share similar characteristics. Suppose that two people with the same bank account history apply for a loan at their bank, and one of them receives his/her loan, while the other person's application is denied. If we would *assess why* the model is producing different outcomes for these persons, we should evaluate the fairness on *individual level*. However, in practice, it is difficult to find a similarity metric that measures the degree of similarity between individuals (Fleisher, n.d.). If you are interested in learning more about individual

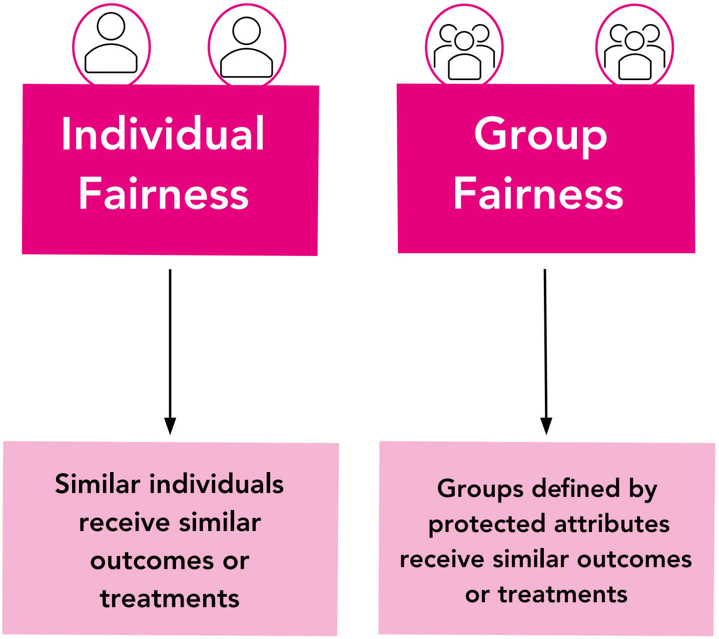

Individual Fairness
Group Fairness
Similar individuals receive similar outcomes or treatments
Groups defined by protected attributes receive similar outcomes or treatments

Figure 10: The difference between individual fairness and group fairness

fairness, we recommend the capstone paper by Dwork et al. (2012) about individual fairness and the *Lipschitz condition*.

With ***group fairness***, we evaluate whether the model treats *groups of persons defined by sensitive characteristics* worse when compared with other groups (Weerts, 2021a). In other words, we measure the extent to which a particular group statistic differs across groups.

In this handbook, we will focus on group fairness, since this level allows a wider range of fairness metrics that can be used.

Requirements for the Bias Analysis

Now that we established our focus on group fairness, we can select a fairness definition and metric(s) for our model. In this section, we explore the main components that are needed for choosing and applying a fairness definition and metric. The following components are needed for fairness metrics on group level:

The **results** of the model

- The performance metrics of the model are summarized on a **confusion matrix**. This confusion matrix is a table that compares the model's predictions with the *ground truth* of the data. The confusion matrix not only shows the model's performance, but also displays the type of errors that the model made, which are the False Positives and False Negatives. Based on the four categories of the confusion matrix, we can calculate several ratios that provide insights about the model's performance on the protected groups. Figure 11 shows a confusion matrix and a selection of metrics that can be distilled from this table.

The **groups** within the dataset that will be compared with each other.

- These groups are characterized by a single or combination of attribute(s) from the dataset that link to sensitive information (see Figure 1). Having a feature such as "sex" may lead you to split the data into a group of men and a group of women to compare the model's performance on these groups and seek large performance discrepancies that indicate bias.

In practice, it is not a straightforward job to establish the groups of interest, because most features present in the dataset link to sensitive information in many complex ways through *indirect* **bias**. Some suggestions for establishing the groups in the dataset can be found in section Forming Groups from Datasets.

Confusion Matrix

		Actual Values			
		Admitted to school Y = 1	Not Admitted Y = 0	P (Y=1 | D)	P (Y=0 | D)
Model Predictions	Admitted to school *D = 1*	True Positives	False Positives	P(Y = 1| D = 1): *Positive Predictive Value*	P(Y = 0| D = 0): *False Discovery Rate*
	Not admitted *D = 0*	False Negatives	True Negatives	P(Y = 1| D = 0): *False Omission Rate*	P(Y = 0| D = 0): *Negative Predictive Value*
	P (D = 1|Y)	P(D = 1| Y = 1): *True Positive Rate*	P(D = 1| Y = 0): *False Positive Rate*		
	P (D = 0|Y)	P(D = 0| Y = 1): *False Negative Rate*	P(D = 0| Y = 0): *True Negative Rate*		P (D=Y) Accuracy

- **True Positives** (TP) are individuals for whom both the model prediction and actual outcome are positive labels.
- **False Positives** (FP) are individuals for whom both the model predicts a positive label, but the actual outcome is a negative label.
- **True Negatives** (TN) are individuals for whom both the model prediction and actual outcome are negative labels.
- **False Negatives** (FN) are individuals for whom both the model predicts a negative label, but the actual outcome is a positive label.

Figure 13: In this Confusion Matrix, we use the same example as our school admission model from Chapter 2.

Selecting a Fairness Definition

One of the main challenges in the bias analysis is determining how fairness should be defined for your use case. Each stakeholder may have a different understanding of fairness, which can be difficult when selecting a fairness metric. Here are some recommendations on selecting a suitable definition:

- **Organise a meeting with stakeholders** to discuss their perspectives on what fairness means for them. These stakeholders involve data scientists, decision makers, and a representative sample of the persons who may be affected by the application of the model. Consider together with them how the different types of errors can harm individuals, groups of persons and society (Rodolfa et al., n.d.).
- **Make simulations of the model outcomes in different scenarios,** for example by using different confusion matrix results. By simulating the disparities across groups, we make fairness issues more tangible and visual, which often helps with establishing the fairness definition.
- **Dive into the fairness metric selection process discussed in the sections below.** Choosing a fairness definition may go hand-in-hand with selecting a metric, with each metric having its own goals, opportunities and constraints. Specifically, the differences between the metrics will be helpful in defining the fairness definition.

Note that **the bias analysis also offers options to adopt multiple fairness definitions for different purposes.** For example, you might select different scenarios or focus groups and translate back what a fair model would look like for these persons. However, when conducting the bias analysis for multiple objectives, **it is important to make explicit how each objective links to a fairness definition and metric, and to keep the results of each bias analysis separate from each other.** This is because the same results do not necessarily apply to all objectives.

Selecting a Fairness Metric

Fairness metrics can be used to compare the model's performance across groups. Since there is an abundance of metrics available to measure the group disparities, we use the Fairness Tree displayed in Figure 13 to navigate through the most important considerations when choosing a metric.

The Fairness Tree

The Aequitas Fairness tree developed by University of Chicago (Saleiro et al., 2018) can be used as a starting point to decide the scope of your bias analysis and find the right fairness metric(s). The fairness tree is part of the Aequitas open source bias audit toolkit containing tools to audit the predictions of AI systems and find biased outcomes. The main advantage of the fairness tree is that it allows both policy makers and data scientists to make an informed decision about the fairness definition and metric.

In the following sections, we zoom in a bit further on the concepts and competing options of the Fairness Tree. Figure 13 contains the numbers that represent the sections below.

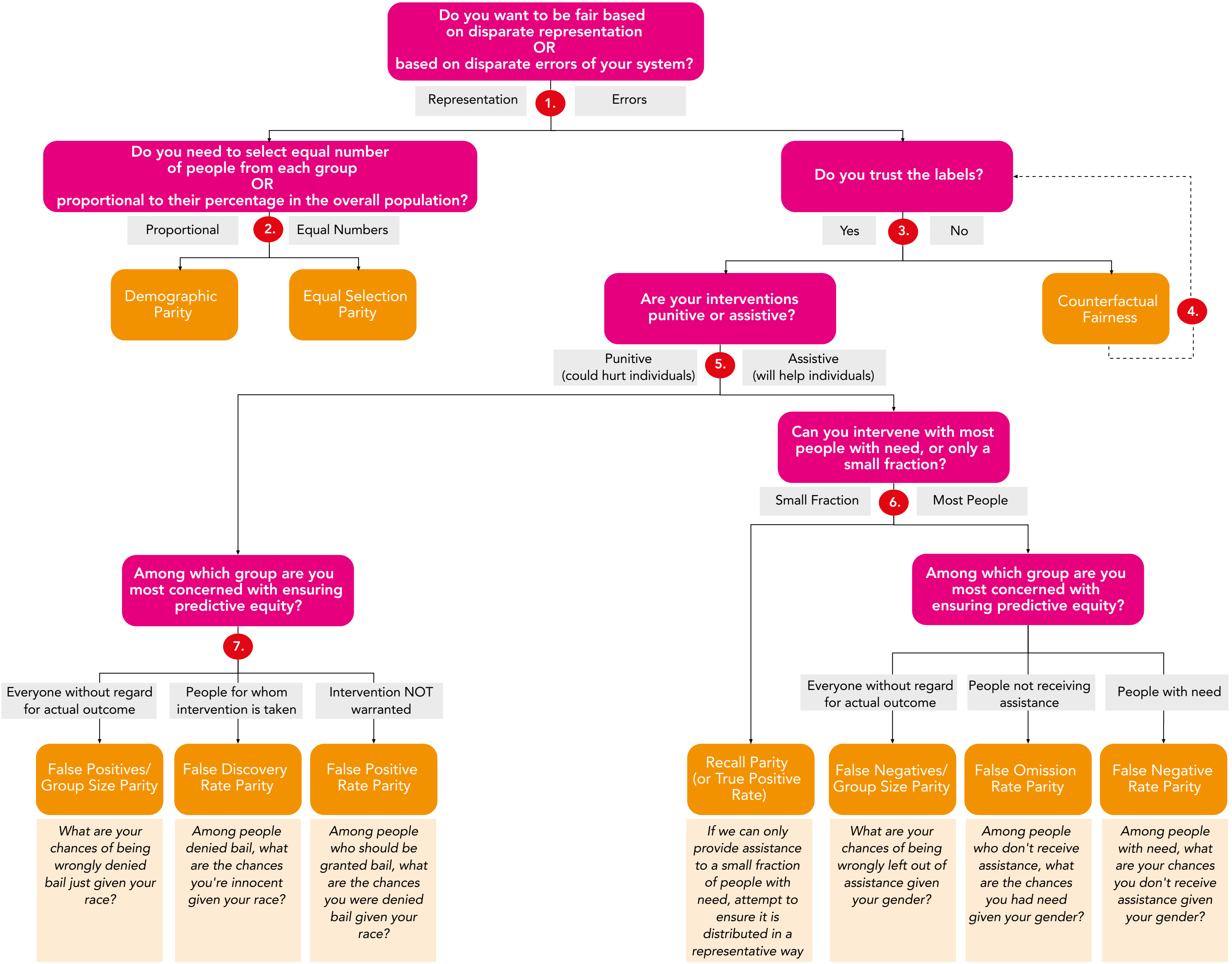

Do you want to be fair based on disparate representation OR based on disparate errors of your system?
Representation
1.
Errors
Do you need to select equal number of people from each group OR proportional to their percentage in the overall population?
Proportional
2.
Equal Numbers
Demographic Parity
Equal Selection Parity
Do you trust the labels?
Yes
3.
No
Are your interventions punitive or assistive?
Counterfactual Fairness
4.
Punitive (could hurt individuals)
5.
Assistive (will help individuals)
Can you intervene with most people with need, or only a small fraction?
Small Fraction
6.
Most People
Among which group are you most concerned with ensuring predictive equity?
7.
Among which group are you most concerned with ensuring predictive equity?
Everyone without regard for actual outcome
People for whom intervention is taken
Intervention NOT warranted
Everyone without regard for actual outcome
People not receiving assistance
People with need
False Positives/ Group Size Parity
False Discovery Rate Parity
False Positive Rate Parity
Recall Parity (or True Positive Rate)
False Negatives/ Group Size Parity
False Omission Rate Parity
False Negative Rate Parity
What are your chances of being wrongly denied bail just given your race?
Among people denied bail, what are the chances you're innocent given your race?
Among people who should be granted bail, what are the chances you were denied bail given your race?
If we can only provide assistance to a small fraction of people with need, attempt to ensure it is distributed in a representative way
What are your chances of being wrongly left out of assistance given your gender?
Among people who don't receive assistance, what are the chances you had need given your gender?
Among people with need, what are your chances you don't receive assistance given your gender?

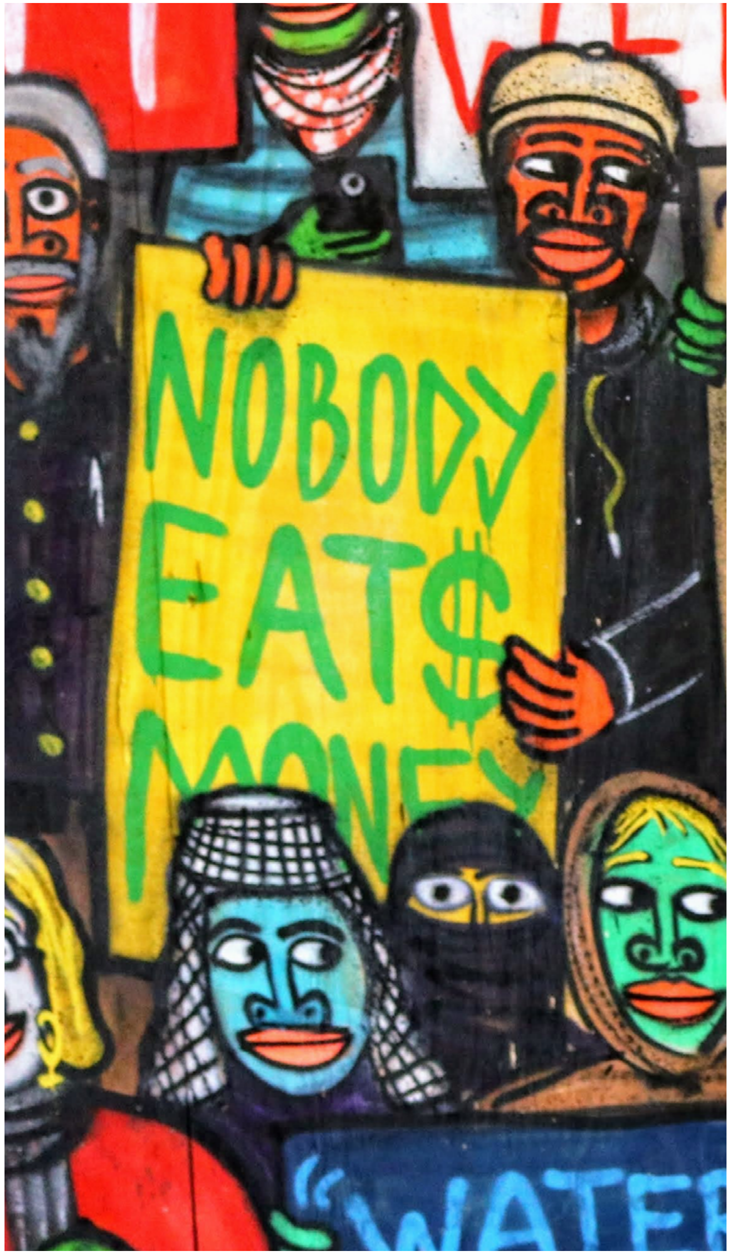
Figure 13: Fairness Tree developed by Aequitas

1. Disparate Representation versus Disparate Errors

The first decision concerns choosing between metrics to evaluate whether the model is fair based on disparate representation or on disparate errors.

With **representation-based metrics**, we compare whether persons from both advantaged and disadvantaged groups have equal probability of being selected by the model, also named the selection rate. This method is often used to evaluate whether persons from different groups have equal access to be selected by the model for a *desired service or good*, such as a loan, insurance or admission to a school programme. These metrics are ***Demographic Parity*** and ***Equal Selection Rate Parity***, which will be discussed in the next section.

With **error-based metrics**, we evaluate the difference in error rates across groups. Suppose we have a model that either denies or approves of a loan and we are curious about if there is a bias against women, e.g., women are more often wrongly denied a loan. Using error-based fairness metrics, we can compare the False Negatives Rate (FNR) between men and women and determine whether women have a substantially higher False Negative Rate than men. The error-based metrics include False Positives Rate Parity, ***False Negatives Rate Parity*** and ***False Omission Rate Parity.*** These metrics will be discussed in later sections.



2. Demographic Parity versus Equal Selection Parity

In this section, we will look at ***Demographic Parity*** and ***Equal Selection Rate Parity.*** These metrics look at fairness as a problem of *disparate representation*.

Demographic Parity

According to ***Demographic Parity***, a model is fair when the selection rate (also named *acceptance rate*) is equal for all the groups that we investigate for presence of bias. If a large discrepancy is found between the selection rates of the groups, we mitigate this bias by picking a threshold such that the fraction of the members that qualify for a service becomes the same (Hardt et al., 2016).

Demographic parity can be used to mitigate *allocation harm* and *quality-of-service harm,* as a low selection rate for a group means that the model less often assigns a favourable outcome for the persons in this group, which can result in, for example, more often denying applications for social benefits or loans for women when compared with men.

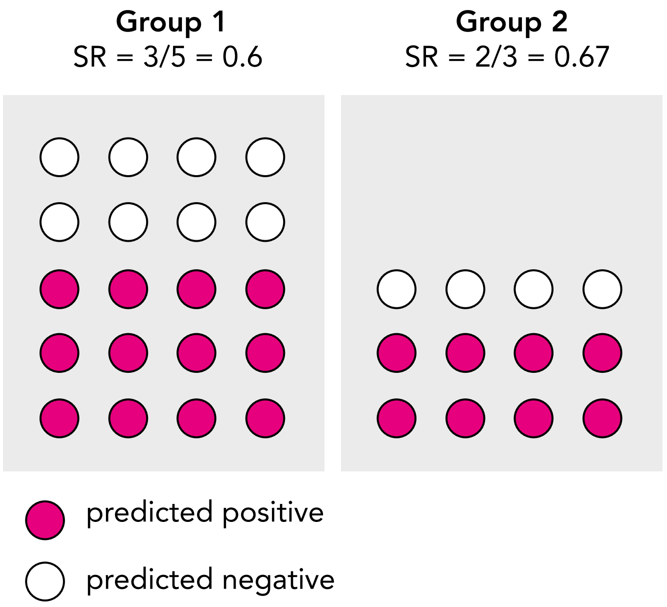


Figure 14: With Demographic Parity, we analyse the difference in selection rate across groups

Pros and Cons

Pros:

- **This definition is not affected by measurement errors within the data,** because this fairness measure does not consider the actual model outcomes, such as the errors and correct predictions. Additionally, this metric aligns most with our human perception of what fairness entails, which makes this metric relatively easy to explain and communicate with stakeholders (Srivastava et al., 2019)

Cons:

- By enforcing demographic parity, we treat groups differently to achieve the same selection rates. This can cause otherwise similar people to be treated differently, which can result in ***procedural harm***, since some decisions from the model may not make sense anymore for individuals. Suppose we use the school admission model and calculate the selection rate for male students and for female students (0.6). Here, we could also decrease the selection rate of male students to 0.6 to achieve an equal selection rate, but this often makes less sense than increasing the women's selection rate to 0.8.
- The fairness metric **requires equal base rates for the different groups.** A base rate is the selection rate observed in the ground truth data. For example, in a fraud prediction model, the base rate represents the probability that the persons from a group defined by sensitive attributes commit fraud. This base rate might be different across groups in reality, but *demographic parity* assumes these base rates to be equal.
- This method rules out any possible correlations between the sensitive attribute and the target.

The dataset can reflect human decisions or second-order effects of societal or historical inequities.

Assumptions behind Demographic Parity

Using demographic parity as the fairness definition, we have the following assumptions:

Regardless of what a measured target variable says,

1- Everybody **is** equal

For example, we may believe that traits relevant for a job are independent of somebody's gender. However, due to social biases in historical hiring decisions, this may not be represented as such in the data.

2- Everybody **should be** equal

For example, we may believe that different genders are not equally suitable for the job, but this is due to factors outside of the individual's control, such as lacking opportunities due to social gender norms.

When to Use Demographic Parity?

Demographic parity is a suitable fairness definition when:

- **We want to use AI systems to change the state of the current world by supporting unprivileged groups.** This support includes, for example, admitting more students from underrepresented ethnicities when compared to other ethnicities.
- **We are aware of historical biases the quality of our data**. For example, we might use Demographic Parity when our model is trained to hire software engineers at a company where nearly no women were hired before.

Calculating Demographic Parity

The difference in demographic parity between groups can be calculated using the metric ***Positive Predictive Value***.

$$PPV = \frac{True\ Positives + False\ Positives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

Equation 1: Predictive Positive Value can be used to quantify demographic parity

Disparate Impact: The Four-Fifths Rule as Threshold

An application of demographic parity can be found in the American legislation named as *disparate Impact*, where the *four-fifths rule* is added to the equal selection rate criterion. We speak of disparate impact when a model *indirectly* discriminates against individuals or groups of people through proxy variables. In U.S. Law, the degree of disparate impact can be measured with the selection rate and the four-fifths rule: **if the selection rate for a vulnerable group is less than 80% of that of the group with the highest selection rate, we speak of discrimination against this vulnerable group** (Mondragon, 2018). The four-fifths rule is an example of a threshold used to determine when demographic parity has been seriously violated. However, we recommend setting up your own threshold for each use case that is appropriate for the context of the model.

Disparate Impact can be calculated using the Disparate Impact function on AIF360.

Equal Selection Parity

This fairness metric is almost similar to Demographic Parity. ***Equal Selection Parity*** compares whether *equal numbers of people* from each group were selected, independent of their group sizes. Hence, the only difference between Equal Selection Parity and Demographic Parity is that the first metric compares *equal numbers*, while the latter compares *equal proportions*.

In our school admission example, this fairness metric would be satisfied if the exact same numbers of people were selected from each group, even if one group contains more students than the other.

3. Trusting the Labels

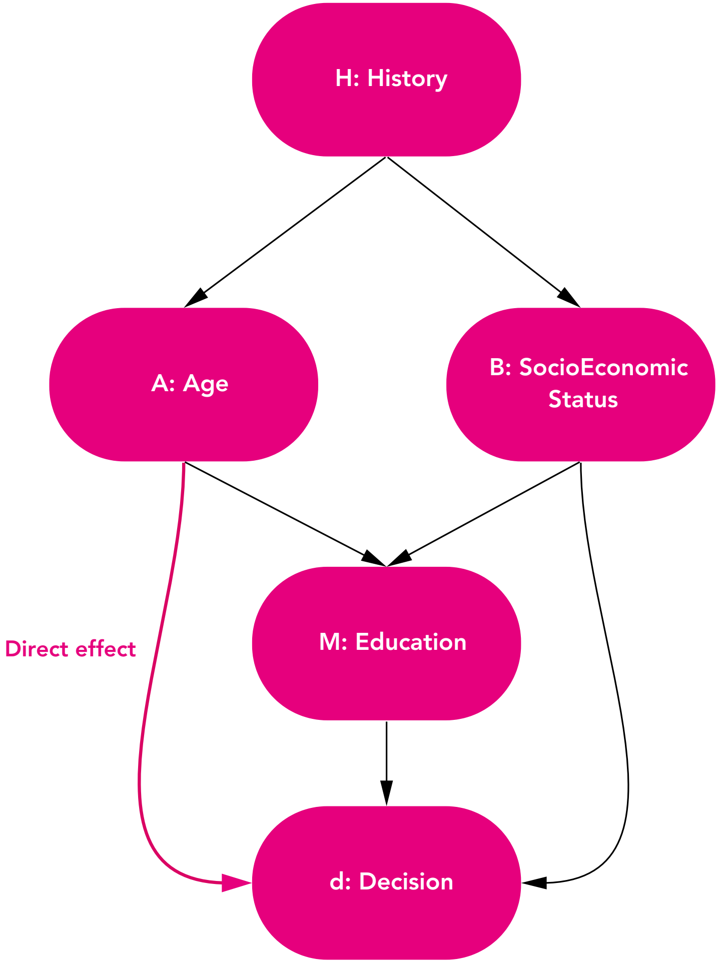
Returning to the Fairness Tree, we see that there is a choice to be made that depends on whether we **trust the labels**.
Suppose we use a fraud prediction model. The dataset contains features, such as 'transactions' and 'average spend', and a label for each instance, such as 'high risk' or 'low risk'. As Machine Learning models are trained on the dataset, the labels should be trustworthy to ensure that the correct patterns are learned. Generally, of most datasets we can say that we can trust the labels. If not, then other, more creative options should be sought to work with the faulty dataset, such as ***Counterfactual Fairness***.

4. Counterfactual Fairness

An alternative way of approaching fairness is to focus more on causality to create causal pathways from sensitive attributes to the model's decisions (see Figure 15). **The causal pathways allow us to gain more insight in the judgments made by the model to evaluate whether these judgments are reasonable** (S. Mitchell et al., 2021).

Additionally, we can translate these causal statements into counterfactuals to find out how a different feature value affects the model's decision, which would still allow us to evaluate a model's fairness for individuals sharing largely similar characteristics. For example, we could find out what the difference in outcome would be if a person would have a young versus old age for him to be hired by a company (S. Mitchell, n.d.). Causal reasoning can be used instead to design interventions to reduce disparities and improve overall outcomes, rather than to define fairness. Particularly, causal graphs can be used to develop interventions at earlier points, prior to decision-making.

However, there is a debate on whether the counterfactuals are well-defined. In practice, it is hard to reach a consensus in terms of what the causal graph should look like and it is even harder to decide which features to use even if we have such a graph, as we may suffer large loss on accuracy if we eliminate all the correlated features (Zhong, 2018).



Example of a causal graph which can be used to investigate causal statements. Suppose we use a model to hire new personnel where the hiring decisions are mainly based on the applicant's educational background. If the model's outcome is affected by increasing the applicant's age while keeping the other features unchanged, we can conclude that the "Age" feature has a large influence on the applicant's probability of being hired. Source: Shira Mitchell

5. Punitive versus Assistive Interventions

The distinction between punitive and assistive interventions plays an important role in our bias analysis, as it helps with determining which type of errors (e.g., False Positives or False Negatives) are most harmful.

When a model has **interventions that are assistive** in nature, people might be harmed when the model fails to intervene on them when they have need (Rodolfa et al., n.d.). A high rate of **False Negatives** is therefore undesirable, as it would mean that the model wrongly withholds this intervention from people. For example, for a model that decides who should receive a governmental subsidy, we could compare the False Negatives Rate across groups to see whether there is a large discrepancy between advantaged and disadvantaged groups.

On the other hand, with **punitive models**, people are harmed by the intervention, which makes the **False Positives** more suitable to further explore during the bias analysis. Suppose we have a model that predicts which of the defendants who committed a crime are likely to reoffend. If our model produces substantially more **False Positives** for people with a non-Dutch ethnicity when compared with people with a Dutch ethnicity, we can say that the model discriminates against people with a non-Dutch ethnicity.

For both the assistive and the punitive models, there are different metrics available, each of which can be used to find error rate disparities across groups. The main differences between these metrics can be attributed to **the size of the intervention** and t**he type of groups that are compared with each other.**

In the sections below, we will further describe the metrics for assistive and punitive interventions. We will explain the metrics for assistive interventions using the example of a model that distributes a scarce financial subsidy amongst all applicants and use a fraud prediction model to elucidate the metrics of the punitive interventions.

6. Models with Assistive Interventions

The Fairness Tree shows the following four metrics applicable for models with assistive interventions:

- Group Size-Adjusted False Negatives (FN/GSP)
- False Omission Rate (FOR)
- False Negative Rate (FNR)
- Recall/True Positives Parity

Group Size-Adjusted False Negatives (FN/GSP)

As discussed in the previous sections, for models with assistive interventions the False Negatives are of particular interest. A natural starting point is therefore to count the number of False Negatives for each group and compare these numbers with each other. Counting False Negatives would result in statements such as:

> *Twice as many men from group A who qualify for the subsidy were wrongly denied the grant when compared with the men from Group B.*

However, if Group A would have twice as many men as Group B, then the difference in False Negatives is still deemed as fair. Thus, the method of comparing numbers of False Negatives neglects the group sizes of Groups A and B. For this reason, the Group Size-Adjusted False Negatives (FN/GSP) metric might be more applicable to compare the FN rate among differently sized groups.

Calculating FN/GSP

Metric FN/GSP can be calculated as follows:

$$\frac{False\ Negatives}{False\ Negatives + True\ Negatives + False\ Positives + True\ Positives}$$

Interpretation of FN/GSP

This fairness definition asks the following question:

> *Just by the virtue of fact that an individual is member of group X, what are the chances they will be falsely denied the subsidy?*

Finding no disparities using this metric implies that if we were to choose a random individual from a given group, we would have the same chance of picking out an eligible person who did not receive the subsidy across all groups.

The Group-Adjusted False Negatives metric therefore considers the following groups in his phrasing of fairness:

- The groups who do not receive the subsidy, including people who do not qualify for the subsidy (True Negatives);
- The groups of persons who qualify for the subsidy (True Positives).

False Omission Rate (FOR)

With the False Omission Rate, we focus on the individuals who did *not* receive the subsidy, regardless of whether they deserved it, which are both the False Negatives and the True Negatives. Our main interest with this metric is to evaluate the fraction of people who were wrongly denied the subsidy from all the people who did not receive the subsidy.

Calculating FOR

The FOR can be calculated for each group with the following formula:

$$\frac{False\ Negatives}{False\ Negatives + True\ Negatives}$$

Interpretation of FOR

When using FOR, we seek answers for questions like:

> *Given that Sarah was denied a subsidy, what are the chances she was actually eligible for it?*

The FOR is useful for assistive models because the False Negatives are relatively easy to track down, as these are the people who did not receive the service or good when they were entitled to it.

False Negative Rate (FNR)

The False Negative Rate measures fairness specifically for people who need the intervention. This metric gives an answer to the following question:

> *For individuals who needs the subsidy, what are the chances they will not receive the subsidy because they are a member of a given group?*

Calculating the FNR

The FNR can be calculated using the following formula:

$$\frac{False\ Negatives}{False\ Negatives +\ True\ Positives}$$

Interpretation of FNR

Finding no disparities in FNR implies that, if we were to choose a random individual *who qualifies for the subsidy* from a given group, we would have the same chance of picking out a person incorrectly denied this subsidy across groups.

Recall

Often, models are used to fairly allocate a **scarce resource** to serve a small fraction of individuals who might benefit. If this is the case for your model, then using recall as your fairness metric might be a good option for the bias analysis. This metric gives insight into how effective the organization is in distributing the subsidy fairly amongst groups.

Calculating Recall

Recall can be calculated as follows:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Alternatively, we can calculate recall by 1 – False Negative Rate:

$$Recall = 1 - \frac{False\ Negatives}{False\ Negatives + True\ Positives}$$

Interpretation of Recall

With recall, we ask the following question:

> *Given that the organization cannot provide the subsidy to all people who apply, is it at least supplying the subsidy to individuals from all groups in a manner that reflects their level of need?*

Suppose we compare the recall rates of men and women. Here, the recall represents the probability of an applicant who deserves the subsidy to be correctly supplied this financial good.

Recall is also named True Positives Parity or Equal Opportunity.

7. Models with Punitive Interventions

Three metrics fall under the scope of the punitive interventions, these are:

- Group-Size Adjusted False Positives (FP/GSP)
- False Discovery Rate Parity (FDR)
- False Positive Rate Parity

As discussed in section Punitive versus Assistive Interventions, for models with punitive interventions we often investigate the False Positives during the bias analysis to prevent that innocent people are picked out by the model for an undesired intervention.

Group-Size Adjusted False Positives (FP/GSP)

Similar to FN/GSP, this metric considers the group sizes when computing the differences in False Positives across groups. Thus, with group-size adjusted False Positives, attempt to seek answers for questions like:

> *Just due to the fact that a person is member of a given group, what are the chances they'll be wrongly classified as fraudulent?*

Calculating FP/GSP

For each demographic group, the FP/GSP can be computed as follows:

$$FP/GSP = \frac{False\ Positives}{False\ Positives + True\ Positives + False\ Negatives + True\ Negatives}$$

Interpreting FP/GSP

Finding no disparities between the groups when using this metric implies that, **if we were to choose a random person from a given group (regardless of whether they are innocent or the group-level fraud rates), we would have the same chance of picking out a wrongly convicted individual across all groups.**

This metric might be useful when there is no sufficient information available about the True Positives, False Negatives or True Negatives, as the denominator is just the sum of all predictions and therefore does not require you to specify how they are distributed across the confusion matrix categories.

False Discovery Rate Parity (FDR)

The False Discovery Rate describes the proportion of positively classified instances which were falsely identified as such (Ruf & Detyniecki, 2021). Thus, it focuses specifically on *the people who receive the (undesired)* intervention and calculates the fraction of *wrong* fraud convictions from *all persons who were classified as fraudulent.*

Calculating FDR

The False Discovery Rate can be calculated as follows:

$$FDR = \frac{False\ Positives}{False\ Positives + True\ Positives}$$

Interpreting FDR

The FDR is a simple and well-scoped metric for the bias analysis, as it only focuses on the people who received the undesired intervention. This metric can also be used to compare the biases of the model with the biases of the decision makers and other people who intervene, because we only need the positive predictions for this metric. It is often more difficult to find the data for the people who committed fraud but were not recognized as such by the model, e.g., the False Negatives.

False Positive Rate Parity

As opposed to FDR where we investigated the people who received the intervention, the False Positive Rate focuses on the people who should not receive the intervention, which are the innocent individuals. Thus, the FPR provides an answer for the following question:

> *For an innocent person, what are the chances they will be wrongly classified as fraudulent due to their membership of a given group?*

Calculating FPR

The False Positive Rate can be calculated as follows:

$$FDR = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

Interpreting FPR

Finding no disparities for FPR implies that, if we would choose a random *innocent* individual from a given group, we would find the same chance of picking out a wrongly convicted person across all groups.

Additional Metric: Equalized Odds

Although the Fairness Tree covered almost all well-known fairness metrics, we discuss an unmentioned metric worth discussing: Equalized Odds.

Equalized Odds

Equalized Odds extends the Recall/True Positive Rate Parity metric with the addition of the False Positive Rate (FPR) (Hardt et al., 2016). Here, we evaluate for all groups whether they have an equal FPR and TPR across groups, which is suitable for models with both punitive and assistive interventions. Using this metric, we provide an answer for the following question:

> *Due to the fact that a person is member of a given group, what are the chances that a person who qualifies for the subsidy receives this financial grant, and a person who does not qualify is denied the subsidy?*

Calculating Equalized Odds

Equalized Odds is calculated by dividing the True Positives by all positive predictions:

$$Equalized\ Odds = \frac{True\ Positives}{False\ Positives + \ True\ Positives}$$

Interpretation of Equalized Odds

With Equalized Odds we primarily want to find out whether the model makes more mistakes for some groups than for other groups (Weerts, 2021b). This metric gives insight in whether the accuracy of the model is equally high in all groups and is able to highlight both allocation and quality-of-service harms.

Forming Groups from Datasets

After choosing the fairness metrics, we compute the performance disparity across groups. However, establishing which groups will be compared and investigated for bias is often a difficult task. Some people belong to multiple vulnerable groups, e.g., women with a black skin colour or older people with a disability status, which increases the chance for **intersectional bias**. This type of bias occurs when a model produces more errors or assigns more undesired outcomes for people who belong to multiple disadvantaged demographic groups.

Below, we offer some suggestions for splitting the features into groups.

1. Create groups based on top-down knowledge of sociotechnical context

Our most important advice is to consult the domain experts to gain insight into the population for which the AI system is designed.

2. Create groups based on feature and error distributions

For the simpler features, it can help to create feature distribution plot to check the distribution of the feature values and to see which groups often helps to see if any 'natural' groups stand out.

For the more complex features, the error distribution can also be informative. **First, apply a feature importance method** on the model to find out which features have the highest predictive value. Then, **create error distribution plots for these features** where the number or magnitude of errors is placed on the Y-axis and the feature values are placed on the X-axis (see Figure 16).

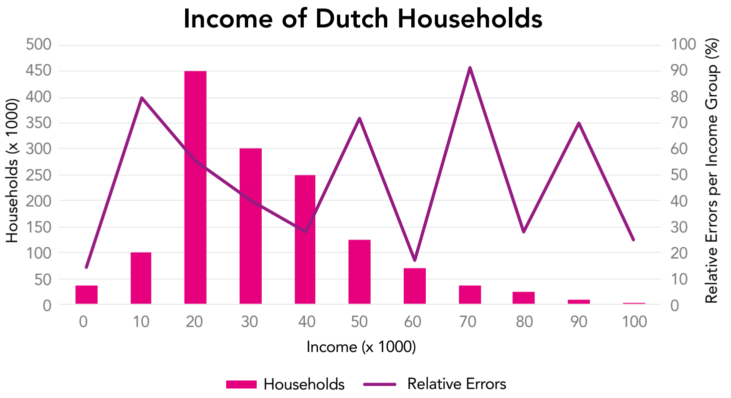


Figure 16: An example of a feature-error distribution plot. Here, we see that the model particularly underperforms for households earning 10, 70 or 90K, which may indicate the presence of bias for these groups.

The visualisations give insight in the error distribution amongst the features and can visualize which categories or values the model is underperforming for, hereby potentially hinting to a demographic group associated with these values.

Determining the Bias Threshold

After applying the fairness metrics on the demographic groups, we analyse the results with stakeholders, with whom we establish **the thresholds for the fairness metrics**. These thresholds are needed to draw conclusions about the presence and magnitude of bias in the algorithm and to determine which actions should be taken to mitigate the bias. Determining the thresholds can be a complicated task, as the impact of the bias can vary by metric, by feature, and by demographic group.

At this time, there are not yet widely accepted and implemented approaches available for determining this threshold. Instead, it will require a tailored approach in collaboration with your stakeholders

Conducting the Bias Analysis

In this chapter, we discussed the components required for the bias analysis. In summary, this is the step-by-step plan:

1. Determine which **definition of fairness** is appropriate for your AI system. Establish with stakeholders the consequences when the model shows bias against when demographic groups.
2. Choose (a) **fairness metric(s)** that fits the fairness definition, and formulate to which question on fairness this metric can provide an answer.
3. Select the features from the **Feature Review** that will be investigated for bias, and **create demographic groups** on which the bias metric(s) will be applied.

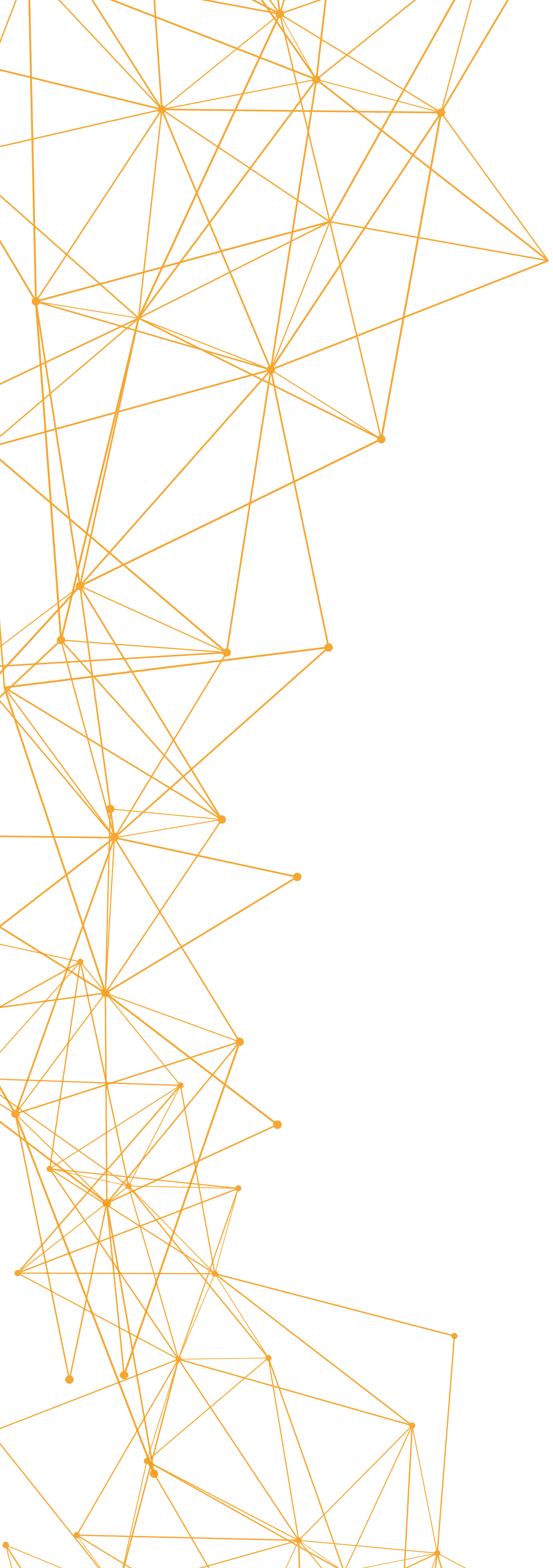
4. Establish **the thresholds** for the bias metric.
5. Carry out the bias analysis and make sure to **document the entire process**.
6. Discuss the bias analysis results with stakeholders. **When biases are found, identify the source of the bias** and return to previous stages of the Fairness Pipeline to mitigate the bias.

See the Appendix for additional useful sources that discuss fairness definitions, metrics and other concepts from this chapter in more detail.

There is no universally accepted definition of what it means for a model to be fair, and there is no clear guideline on which fairness measures as "best".

Chapter Takeaways

- We can investigate fairness from the *individual fairness and group fairness perspectives*, of which the first compares the model's outcomes for similar individuals, and the latter compares whether the model produces more harmful outcomes for persons due to their membership to a given group.
- As adopting a fairness definition for the model requires a deep understanding of the socio-technical context, we recommend closely collaborating with stakeholders in these undertakings.
- The Fairness Tree is a helpful guide for selecting an appropriate fairness metric.
- Creating demographic groups from the dataset can be done based on top-down knowledge of the AI system, and based on a bottom-up approach using feature and error distributions.



6. Conclusion

Whether AI systems will be beneficial for all people depends on the choices we make through the model development cycle: how do we represent complex and difficult to measure constructs? How do we choose the features to predict our target variable? Most importantly, can we trust the data that fuels the AI system? Is it representative and complete?

In this Fairness Handbook, we provided insights and practical tools about how harmful biases impact the fairness of models. If not taken care of these biases, they will become baked in and scaled by AI systems, thereby increasing their negative impact on vulnerable groups and individuals.

As you know by now, when we want to find and mitigate traces of discrimination, it is impossible to solely approach fairness from a technical-oriented perspective. Since biases in models origin from human decisions and societal inequalities, there is an interdisciplinary approach required, where domain experts and impacted groups of people play a crucial role in determining which factors the model development team needs to consider before implementing and deploying a model. These factors include investigating the target variable and sensitive attributes of the dataset and simulating different scenarios of possible biases in the model.

The purpose of this book is not to give you a ready-made answer immediately applicable to your own use case, as each model, dataset and context requires specific interventions. Instead, we aimed to provide guidance based on evidence-based methods and our own experiences during our bias analysis to help you choose which interventions are most suitable in your quest to create fairer models.

7. References

Abbasi, M., Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2019). Fairness in representation: Quantifying stereotyping as a representational harm. *SIAM International Conference on Data Mining, SDM 2019*, 801–809. https://doi.org/10.1137/1.9781611975673.90

Barocas, S., Hardt, M., & Narayanan, A. (2016). Chapter 1: Introduction. In *Fairness and Machine Learning*.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 1–6. https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, *2017-Decem*, 3993–4002.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/JAIR.953

Developers, G. (2019). *Fairness: Types of Bias | Machine Learning Crash Course*. https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias

Dignum, V. (2021). The Myth of Complete AI-Fairness. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12721 LNAI*, 3–8. https://doi.org/10.1007/978-3-030-77211-6_1

ECP | Platform voor de InformatieSamenleving. (2018). *AI Impact Assessment*. https://futurium.ec.europa.eu/en/european-ai-alliance/best-practices/ai-impact-assessment-code-conduct

Elhassan, A., Aljourf, M., Al-Mohanna, F., & Shoukri, M. (2016). Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. *Global Journal of Technology and Optimization*, *01*(S1), 111. https://doi.org/10.4172/2229-8711.s1111

Fairlearn. (n.d.). *1. Fairness in Machine Learning — Fairlearn 0.7.0 documentation*. Retrieved December 15, 2021, from https://fairlearn.org/v0.7.0/user_guide/fairness_in_machine_learning.html#types-of-harms

Fleisher, W. (n.d.). *What's Fair about Individual Fairness?*

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. In *Communications of the ACM* (Vol. 64, Issue 12, pp. 86–92). Association for Computing Machinery. https://doi.org/10.1145/3458723

Ghani, R., Rodolfa, K. T., Driscoll, A., Casey, P., & Amarsinghe, K. (2020). *Data Science Project Scoping Guide*. http://www.datasciencepublicpolicy.org/our-work/tools-guides/data-science-project-scoping-guide/

Giurca, A. (2021). *AI Fairness in Financial Services How to quantify and improve fairness in machine learning and AI applications?* www.probability.nl

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*, 3323–3331.

Hattori, K., & Takahashi, M. (2000). A new edited k-nearest neighbor rule in the pattern classification problem. *Pattern Recognition*, *33*(3), 521–528. https://doi.org/10.1016/S0031-3203(99)00068-0

Hayes, A. (2020). Stratified Random Sampling Definition. In *Investopedia.com* (pp. 1–1). https://www.investopedia.com/terms/stratified_random_sampling.asp

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the International Joint Conference on Neural Networks*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hu Zhang, B., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, *18*. https://doi.org/10.1145/3278721

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/S10115-011-0463-8

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 924–929. https://doi.org/10.1109/ICDM.2012.45

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7524 LNAI*(PART 2), 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*, 1–5. http://jmlr.org/papers/v18/16-365.html.

Lundberg, S. M. (2019). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, *32*(2), 1208–1217. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. In *arXiv*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. In *ACM Computing Surveys* (Vol. 54, Issue 6). https://doi.org/10.1145/3457607

Menardi, G., Torelli, N., Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Min Knowl Disc*, *28*, 92–122. https://doi.org/10.1007/s10618-012-0295-5

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, *Figure 2*, 220–229. https://doi.org/10.1145/3287560.3287596

Mitchell, S. (n.d.). *Reflections on Quantitative Fairness | Definitions: Causality*. Retrieved January 5, 2022, from https://shiraamitchell.github.io/fairness/#definitions-causality

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Molnar, C. (2020). *5.5 Permutation Feature Importance | Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/feature-importance.html

Molnar, C. (2022). *8.1 Partial Dependence Plot (PDP) | Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/pdp.html

Mondragon, B. N. (2018). *What is Adverse Impact? And Why Measuring It Matters*. https://www.hirevue.com/blog/hiring/what-is-adverse-impact-and-why-measuring-it-matters

Pessach, D., & Shmueli, E. (2020). Algorithmic Fairness. *AEA Papers and Proceedings.*

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems*, *2017-Decem*, 5681–5690.

Rodolfa, K. T., Saleiro, P., & Ghani, R. (n.d.). *Chapter 11 Bias and Fairness | Big Data and Social Science*. Retrieved April 5, 2022, from https://textbook.coleridgeinitiative.org/chap-bias.html#dealing-with-bias

Ruf, B., & Detyniecki, M. (2021). *Towards the Right Kind of Fairness in AI*. http://arxiv.org/abs/2102.08453

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). *Aequitas: A Bias and Fairness Audit Toolkit*. http://arxiv.org/abs/1811.05577

Saleiro, P., Rodolfa, K. T., & Ghani, R. (2020). Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 3513–3514. https://doi.org/10.1145/3394486.3406708

Silberg, J., & Manyika, J. (2019). Notes from the AI frontier: Tackling bias in artificial intelligence (and in humans). *Mckinsey Global Institute*, 1–8. https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans#

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. Human perception of fairness: A descriptive approach to fairness for machine learning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2459–2468. https://doi.org/10.1145/3292500.3330664

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3465416.3483305

Swee Kiat, L. (n.d.). *Understanding Bias I | Machines Gone Wrong*. Retrieved December 24, 2021, from https://machinesgonewrong.com/bias_i/

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings - International Conference on Software Engineering*, 1–7. https://doi.org/10.1145/3194770.3194776

Weerts, H. J. P. (2021a). *An Introduction to Algorithmic Fairness*. 1–18. http://arxiv.org/abs/2105.05595

Weerts, H. J. P. (2021b). *Fairness workshop*. PyLadies. https://github.com/pyladiesams/ml-fairness-beginner-nov2021/blob/master/workshop/measuringgroupfairness.ipynb

Wu, T., Ribeiro, M. T., Heer, J., & Weld, D. S. (n.d.). *Errudite: Scalable, Reproducible, and Testable Error Analysis*. Retrieved May 15, 2020, from https://youtu.be/Dil5i0AYyu8.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *30th International Conference on Machine Learning, ICML 2013*, *PART 2*, 1362–1370. https://proceedings.mlr.press/v28/zemel13.html

Zhong, Z. (2018). *A Tutorial on Fairness in Machine Learning*. Towards Data Science. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

8. Appendix

The Fairness Handbook appendices elaborate on the topics discussed in the Fairness Pipeline and Bias Analysis chapters.

A. Overview of Traps and Biases

In this chapter of the appendix, we discuss the traps and biases that may arise during the model development cycle. This overview further elaborates on the concepts encountered in Chapter 4: The Fairness Pipeline.

Sources of fairness issues in models can already be present before data collection and model training. Often, these sources can be found at the initial phases of the model development cycle, where the model's objectives are determined, and the real-world problem is translated to a problem that can be solved with predictive algorithms. We refer to these sources as ***design traps.***

When there are problems in the data, the model, or in the deployment of the model that lead to skewed outcomes, we speak of ***biases***. These biases can lead to discrimination against individuals or groups defined by protected attributes, such as *nationality* or *disability* status.

Solutionist Trap

Machine learning is not the solution for every task. Avoid falling in the ***solutionist tra***p, where a (new) technology's advantages are overestimated and its risks are underestimated (Weerts, 2021a). Particularly, the objectiveness of the data and algorithms is often overvalued. Therefore, rather than asking *"can we use machine learning"*, it is better to ask, *"how can we solve this problem?"* and then consider machine learning as one of the options.

Abstraction Trap

If something goes wrong in the translation of the real world to the model, we speak of an ***abstraction trap***, where we fail to capture all the relevant aspects and the sociotechnical context of the model. Let's say we are designing a model to measure employee quality. We would likely use a (combination of) proxy variable(s), but these may not fully capture the complexity of the target variable.

Omitted Variable Bias

The abstraction trap can lead to ***omitted variable bias***, a form of statistical bias which occurs when one or multiple relevant features are left out of a model. This type of bias can indicate *procedural harm* when the impact of the missing feature is attributed to the remaining features. Say, for example, that we create a regression-based model to determine whether a person is eligible for a sport scholarship, and that the model considers age as feature, which is heavily correlated with *health history*. If *health history* is not included as feature, the model might wrongly produce outcomes based on age, while in fact, these outcomes depend on *health history*. Eventually, the results of the model wrongly indicate direct discrimination and *procedural harm*, while this is not the case.

Aggregation Bias

Aggregation bias arises when a *one-size-fit-all* model is used for groups with different data distributions. Here, the wrong assumption is made that the labels hold the exact same meaning across groups: as datasets often represent people or groups with different backgrounds and norms, the given variables can also mean different things across them (Suresh & Guttag, 2021). Consequently, the model might not work well for any of the groups.

When the aggregation bias occurs simultaneously with ***representation bias***, the model will primarily work well for the majority population (Weerts, 2021a). This problem relates to relates to *underfitting*, as the model is unable to capture the more subtle differences in data distribution. Failing to recognize the different data distribution of the minority groups leads to a model that has a poor predictive performance for these underrepresented groups. This in turn leads to allocation harm and quality-of-service harm.

For example, in loan eligibility models, the set of actions that a model can often conduct is either approving or denying a loan, while in reality there is a much wider range of actions possible.

Ripple Effect Trap

The ***Ripple Effect Trap*** takes place when the introduction of the model into an existing social system changes the behaviours and values of the system in unforeseeable ways and potentially leading to *procedural harm* (Selbst et al., 2019). As the model produces outcomes that lead to interventions conducted by people and organizations, it is essential to understand how the interventions affect the context and the dynamics of the system.

For example, a model used for hiring new personnel may change the behaviour and power dynamics of the HR department, which can eventually lead away from the desired goals.

Historical Bias

Historical bias can encode human and real-world biases into AI models. This might inflict *representation harm* to particular groups that already experience structural discrimination. Historical bias can lead to construct validity bias, if the labels of the variables are based on human judgment. As humans are biased, this may lead to biased labelling.

If there's one thing you should remember from this handbook, it's that data is subjective. The features, categories and the measurement methods in which information is converted into data is a process carried out by people who unconsciously bring their own background, unintended biases and real-life prejudices into the end product.

Therefore, be critical of what features represent and keep asking yourself whether the target variable, feature or other piece of data actually represents what it should represent.

Historical bias can also occur when the data is a good representation of a biased reality. The variable is labelled correctly (so there is no **measurement error** nor **construct validity bias** present), but its values are different across groups due to structural inequalities in society. For example, if a hiring AI system is based on data that contain a pattern of favouritism towards men, it may lead to a system which has discrimination against women built into it.

The feedback mechanisms of a model can amplify the existing historical biases when new data is collected based on the output of a biased model (Weerts, 2021a). This **feedback loop** can lead to a **measurement error** like *predictive policing*, where the model continuously targets the same groups of people. For example, a model predicting in which areas more police should be deployed can result in higher arrest rates for this area, which in turn leads the model to deploy even more police to these regions. Over longer span of time, these patterns of overpolicing will be reflected in the data, which leads to even more biased data.

Construct Validity Bias

When a variable does not accurately measure the construct it is supposed to measure, we speak of **construct validity bias** (Weerts, 2021a). Construct validity bias stems from failing to translate an abstract real-world concept to a concept that can be measured, leading to representation harm, allocation harm and quality-of-service harm.

When a variable does not accurately measure the construct it is supposed to measure, we speak of ***construct validity bias*** (Weerts, 2021a). Construct validity bias stems from failing to translate an abstract real-world concept to a concept that can be measured, leading to *representation harm*, *allocation harm* and *quality-of-service harm.*

There is a specific risk for this when working with sensitive attributes which are made-up constructs, such as *socio-economic status*, or complex constructs, such as *religion*, as this creates a risk that these features mean different things for different individuals. For example, a feature regarding *race* can be created using self-reported racial identity, observed race based on appearance, or it can be constructed using observations of how individuals interact with each other. It is therefore important to report how sensitive attributes are constructed and whether this process was equal for all groups in the dataset (Weerts, 2021b).

Representation Bias

If there is a lack of geographical, social or ethnic diversity in the dataset, the model is unable to generalize well for these groups and will thus produce more errors for these groups. We refer to this type of **measurement error** as ***representation bias***. These errors can lead to *quality-of-service harm*, as the AI system will not work as well for minority groups as for the majority of the population.

Representation bias is mainly caused by ***selection bias***. Selection bias occurs when the data collection or selection results in a non-random sample of the population (Weerts, 2021a). There are several subtypes of selection bias:

- **Coverage bias:** Occurs when the population we want to make predictions about is not accurately represented in the dataset. For example, we have a model that predicts people's enjoyment of a movie and want to collect data to train the model. If we collect surveys data from people leaving a movie theatre viewing of the movie, we sample from a population that is likely to like the movie more than average (Developers, 2019).
- **Sampling bias:** Occurs when data is not collected *randomly* from the target group. For example, when collecting data to build a model that predicts how many people would be likely to apply for a job at a company based on their applications, it would not make sense to only pick the first 200 applicants for the dataset. By including proper randomization during data collection, all the instances have equal probability of being included in the sample, which increases the representativeness of the data.
- **Self-selection bias:** Occurs when people from certain groups more often opt-out of the data collection process. For example, people with strong opinions about a topic may respond more often to surveys than people with mild opinions. This inclination for a person to opt in for a survey is only problematic when this choice correlates with the features in the dataset.
- **Hawthorne Effect:** this bias takes place when data subjects behave differently when they know they are observed. The Hawthorne Effect is particularly prevalent when self-reports are used (e.g., surveys or questionnaires), due to people being influenced by social desirability.

Learning Bias

Representation bias may lead to ***learning bias***, which happens when modelling choices amplify performance disparities across different examples in the data, thereby disproportionately affecting underrepresented groups or individuals (Suresh & Guttag, 2021).

This bias is often caused by the prioritisation of some objective over another. For example, when we optimize for *compactness or privacy,* we often reduce the influence of underrepresented data on the model. Consequently, the model's overall performance can decrease for the minority groups, as the model has now mostly preserved information about the most frequent features which can often be found in the *majority* group. This causes the model to not generalize well for underrepresented groups.

Evaluation Bias

When the performance metrics and procedures are inappropriate for the model, we speak of ***evaluation bias***. A common occurrence of this bias is when a single measure is used to report the performance of all the groups in the dataset. This single measure often hides the underperformance of models on minority groups, which in turn is often caused by the lack of data on these groups.

When investigating the model for evaluation bias, it is important to closely inspect the assumptions behind the used performance metrics and to determine whether they are suitable for your model.

Deployment Bias

Deployment bias arises when there is a mismatch between the model's design and the context in which it will be applied. The role of the complex sociotechnical context in which the model is used is often overlooked, which means that a good performing model can be harmful due to it being used by decision makers and institutions in a wrong way, thereby leading to *quality-of-service harm*.

For example, the results of the model can be interpreted wrongly or differently for different groups by human decision-makers, who in turn might experience automation or confirmation bias. Confirmation bias is the tendency to primarily focus on information that confirms one's beliefs (Weerts, 2021a)

Automation bias

Automation bias refers to a tendency of people to favour results generated by automated systems over those generated by non-automated systems (e.g., human inspectors), regardless of the error rates of each system (Suresh & Guttag, 2021). This bias creates the risk that the errors of the models are overlooked and that the consequences are discovered a much later stage.

In addition, machine learning models are often reused in context that are different then what they were initially designed for. If this is not done in an intentional way, there is a risk of the *portability trap* emerging, where differences between the intended and actual contexts are neglected.

Another way in which deployment bias may manifest is through *temporal bias*. This bias arises from differences in populations and behaviours over time (Mehrabi et al., 2019). For example, many "patterns" of daily life have been radically changed by the Covid-19 pandemic, which has reduced the predictive value of much of the pre-pandemic data when applied to a post-pandemic world.

Both *deployment* and *automation bias* differ from other types of biases in that they are not linked to the data nor the model. Rather, these biases occur when decision-makers and other stakeholders behave unexpectedly based on the AI system's outcomes.

B. Mitigation Algorithms

This chapter of the appendix discusses the specifically designed fairness ***mitigation algorithms*** that can be used to detect, mitigate and prevent biases in models. We made this distinction to stimulate considering both policy-based and algorithm-based approaches for analysing models for biases, as the bias analysis is a multi-disciplinary pursuit.

Generally, we can distinguish three moments in which bias mitigation techniques can be applied

- Before creating the model, as pre-processing techniques
- After training the model: in-processing techniques
- After the model evaluation on the test: post-processing techniques

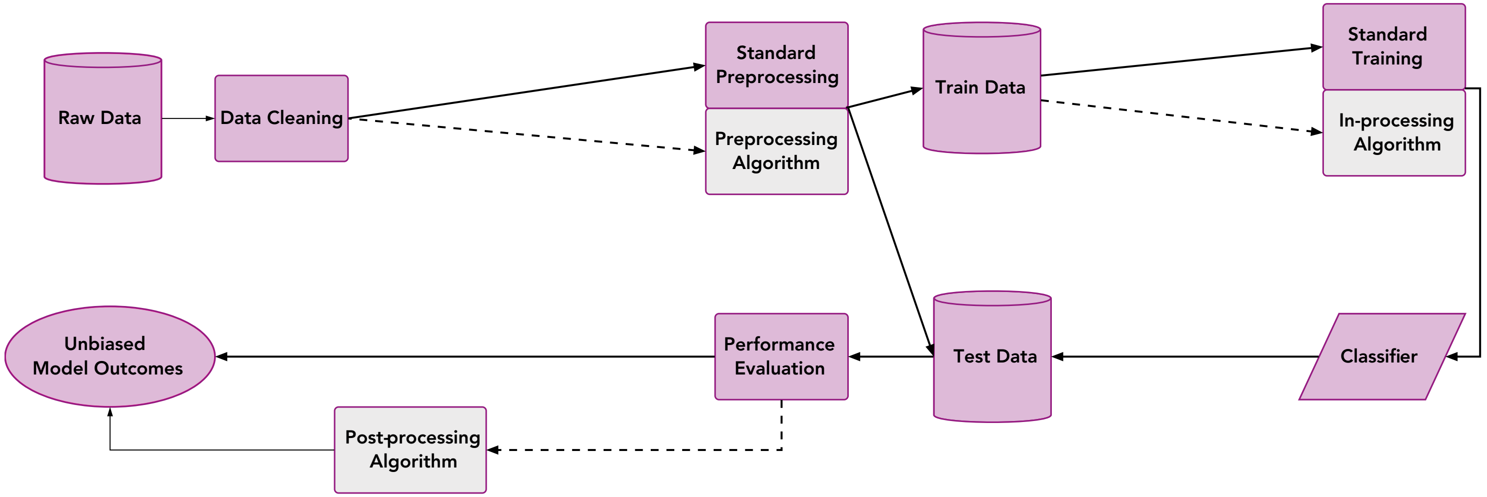


Figure 17: Bias Mitigation during model development cycle. Inspired from (Giurca, 2021)

Pre-processing Techniques

The first branch of techniques addresses bias by transforming the data before creating the model in order to remove the underlying discrimination (Mehrabi et al., 2021). These transformations may include modifying the labels, observed data and the weighting of the features. After de-biasing the training data, the model can be trained in the standard way.

Pre-processing techniques are preferred over in-processing and post-processing algorithms when the model is only available as a "black box" or when it originates from a third party, because pre-processing methods do not need to modify the model.

Examples of pre-processing techniques include ***optimized pre-processing*** (Calmon et al., 2017) and **reweighing instances** (Kamiran & Calders, 2012). The reweighing technique weighs the examples in each group and label combination differently to ensure fairness before classification. The observations of the disadvantaged group with a favourable label get higher weights, while observations of the privileged group with a favourable label get lower weights (Giurca, 2021).

In-processing Techniques

In-processing techniques aim to modify and change learning algorithms to remove discrimination during the training process of the model. This modification can either consist of incorporating changes into the objective function, or it can impose new fairness constraints (Mehrabi et al., 2021). For example, the cost function of the learning algorithm can be modified so that it now includes an extra discrimination-aware regularization term.

Generally, it is recommended to use in-processing techniques when the model is built in-house, as this branch of techniques offers the highest flexibility to choose the trade-off between model performance and fairness. A disadvantage, however, is that in-processing algorithms often depend on the type of AI model, which makes them less generalisable onto other AI models.

Some examples of in-processing techniques include:

- ***Adversarial Debiasing*** (Hu Zhang et al., 2018)
  Using generative adversarial networks, this method learns a classifier to maximize prediction accuracy while reducing an adversary's ability to determine the protected attribute from predictions. Eventually, this approach leads to a fair classifier, as the predictions cannot carry any discrimination information that the adversary can exploit.
- ***Prejudice Remover*** (Kamishima et al., 2012)
  This technique adds a discrimination-aware regularization term to the learning objective that penalizes unfair solutions. For instance, the cost function can take into account what the differences are between the learning algorithms' classification performance on protected versus non-protected classes, and it can penalize the total loss based on the magnitude of the difference.

Post-processing Techniques

Finally, post-processing techniques are applied on a holdout set on after the model has been trained. This branch of techniques is most suitable when we have a "black box" model of which the training data or the learning algorithm cannot be modified (Mehrabi et al., 2021).

The main idea behind post-processing techniques is to manipulate output predictions in such way that it minimizes a fairness metric. Using the outputs of the classifier, thresholds are sought for each group that eventually result in equal prediction distributions. Another branch of post-processing techniques is directly intervening in the validation dataset to choose the appropriate classification threshold that ensures fairness.

- ***Reject Option Classifier*** (Kamiran et al., 2012)against females Gives favourable outcomes to unprivileged groups and unfavourable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty
- ***Equalized Odds Post-processing*** (Hardt et al., 2016) Solves a linear program to find probabilities with which to change output labels to optimize equalized odds.
- ***Calibrated Equalized Odds*** (Pleiss et al., 2017) Optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective.

C. Further readings

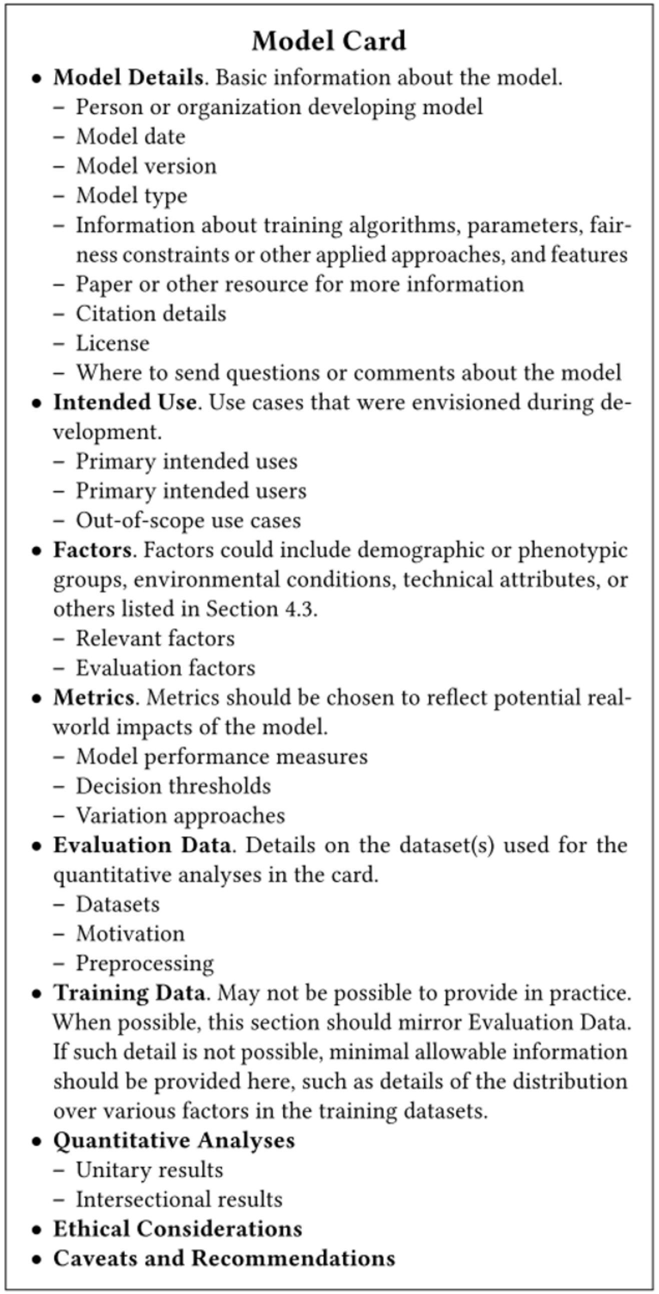
To read more about fairness definitions and metrics, we recommend the following sources:

- Verma, S., & Rubin, J. (2018). Fairness definitions explained. Proceedings - International Conference on Software Engineering, 1–7. https://doi.org/10.1145/3194770.3194776
- Weerts, H. J. P. (2021). An Introduction to Algorithmic Fairness. 1–18. http://arxiv.org/abs/2105.05595
- Suresh H, Guttag J (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle - ACM International Conference Proceeding Series. https://doi.org/10.1145/3465416.3483305
- Garg, P., Villasenor, J., & Foggo, V. (2020). Fairness Metrics: A Comparative Analysis. Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020, 3662–3666. https://doi.org/10.1109/BigData50022.2020.9378025
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. In arXiv. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Because conducting a bias analysis is a challenging job, we recommend following online workshops to gain practical knowledge. We found these workshops useful:

- Fairness Workshop by the developers of the Fairness Tree
- Fairness Workhop by Hilde Weerts
- Lecture on Fairness and Bias by MIT

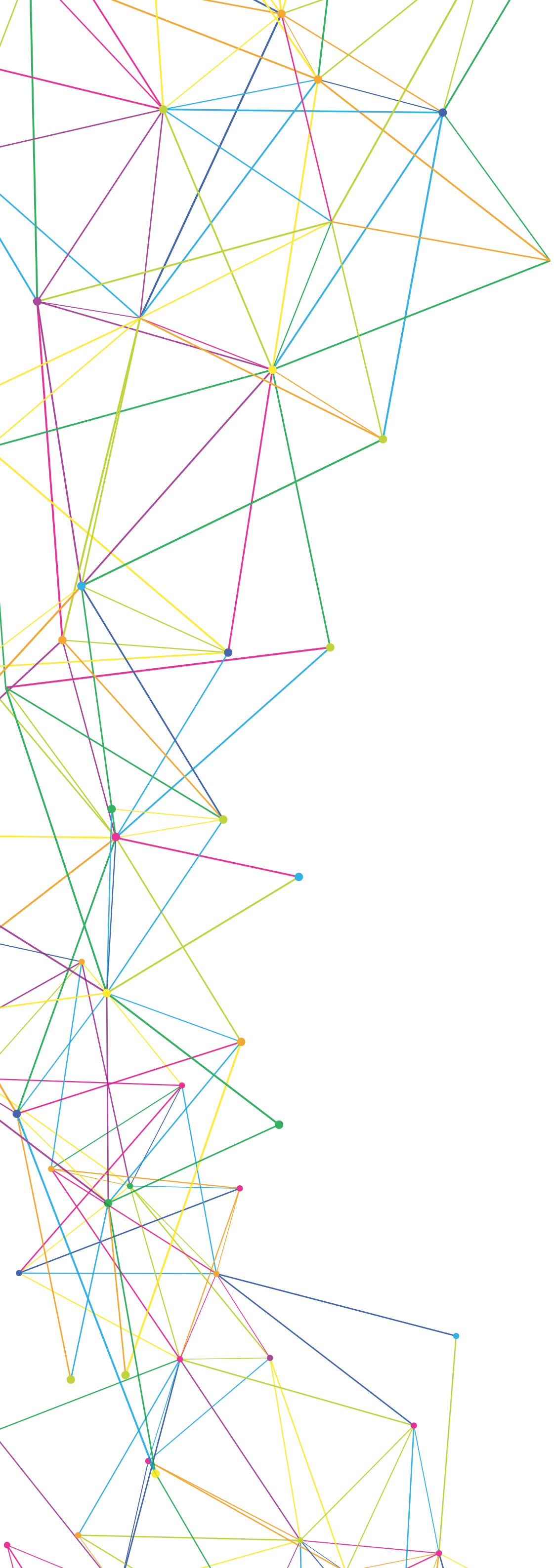
D. Example of a Model Card



Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Figure 18: Suggested Prompts for Model Cards. Source: (M. Mitchell et al., 2019)



Colofon

Redactie:
Selma Muhammad

Met dank aan:
Swaan Dekkers
Meeke Roet
Sofie Verhoeven
Sebastian Davrieux
Bart de Visser
Linda van de Fliert
Hilde Weerts
Maranke Wieringa
Mirthe Dankloff
Joris van Klingen

Vormgeving:
Vorm de Stad