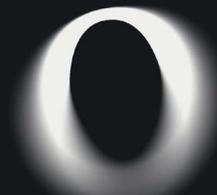
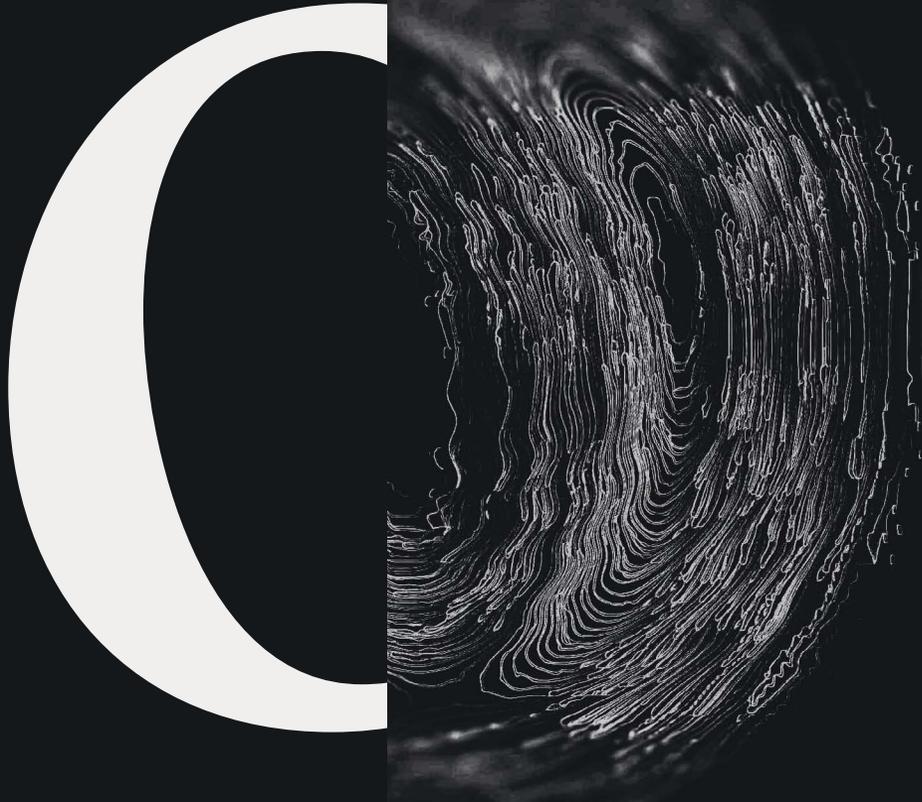


non-discrimination by design





About this handbook

This document explains which questions and principles should be leading in the development and implementation of an AI system in light of the prohibition on discrimination, taking into account legal, technical, and organisational perspectives. It is intended for project leaders managing a team of system builders, data analysts, and AI experts. Suppose you want to ensure that an AI system is as non-discriminatory as possible, which questions should be top of mind and which discussions should take place within your team?

About the design of this handbook

The visual concept of this document was developed to optimise the readability and recognisability of its contents. The design of the 'o' guides you through this document. It changes colour and shape. The colour tells you which phase you are in, and the different shades and shapes of the O point to the type of principles relevant to that phase. On the next page (Index), you will first be familiarized with the concept of the O. On page 18 and 19, the design choices and the concept will be elaborated further.

4 Introduction

- 5 Non-discrimination principles
- 6 Discriminatory grounds
- 10 Scheme of discrimination law

11 Purpose of this handbook

- 12 Diversity
- 13 Context sensitivity
- Verifiability
- 14 Evaluation

15 Let's get started!

17 System overview

18 Art & design explanation

20 1 - Problem definition

- 21 Purpose & necessity
- Impact
- Success criteria

Examples

- 22 Labour market context
- 23 Criminal justice context
- 24 Medical context

Principles

- 25 Legal
- 26 Technical
- 27 Organisational

28 2 - Data collection

- 29 Purpose & necessity
- Data quality
- Data storage

Examples

- 30 Labour market context
- 31 Criminal justice context
- 32 Medical context

Principles

- 33 Legal
- 33 Technical
- 35 Organisational

36 3 - Data preparation

- 37 Inclusion & exclusion
- Integration & aggregation
- Labelling

Examples

- 38 Labour market context
- 39 Criminal justice context
- 40 Medical context

Principles

- 41 Legal
- 42 Technical
- 43 Organisational

44 4 - Modelling

- 45 Pre-modelling
- Model(selection)
- Test

Examples

- 46 Labour market context
- 47 Criminal justice context
- 48 Medical context

Principles

- 49 Legal
- 50 Technical
- 52 Organisational

53 5 - Implementation

- 54 Practical test
- Model alterations
- Application

Examples

- 55 Labour market context
- 56 Criminal justice context
- 57 Medical context

Principles

- 58 Legal
- 58 Technical
- 60 Organisational

61 6 - Evaluation

- 62 Evaluation preparation
- Evaluation
- Points of action

Examples

- 63 Labour market context
- 64 Criminal justice context
- 65 Medical context

Principles

- 66 Legal
- 67 Technical
- 68 Organisational

69 Colophon

Article 14 of the *European Convention on Human Rights* prohibits discrimination: 'The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.' Non-discrimination is one of the pillars of liberal democracies and the rule of law.

In recent years, it has become abundantly clear that AI systems can have discriminatory effects. Examples include, but are not limited to, a facial recognition system that fails to accurately recognise people with a dark skin, translation tools that generate stereotyping texts, or a résumé screening system that unfairly favours male candidates. How can we develop systems that are designed to minimise the risk of producing unintended and unjust distinctions between groups of people?

Non-discrimination law typically provides principles and questions as opposed to requirements or prohibitions. Lawyers are often expected to simply spell out what is permitted and what is not, but that is generally not how the legal domain works. The law provides a standard or a principle, but there are always exceptions. In addition, a judge will always consider the context – or what lawyers refer to as “the circumstances of the case.”

Discriminating between people based on sex is prohibited, for example, unless it is a relevant factor. When a casting agency is looking for an actress to portray a female character, it is obviously permitted to exclude male candidates for that role. In certain circumstances, positive discrimination may even be permitted – an organisation consisting of mainly male employees may decide to favor women in the recruitment process.

Likewise, what is or is not “discriminating” cannot be answered in a straightforward manner. The legal principle is that equal cases in circumstances should be treated equally, and unequal cases

Non-discrimination principles

Direct discrimination

Indirect discrimination

in circumstances should be treated unequally. But what is equal and what is not? Which factors are relevant, and which are not? This guide cannot provide general answers to those questions, because they depend on the context and the way an AI system functions, what purpose it serves, and which safeguards are in place.

Importantly, the law not only prohibits discrimination based on a limited number of grounds – such as race, sex or sexual orientation – but on “any other status.” This requires system developers to have a good understanding of how a system makes distinctions, and whether those distinctions are justifiable. Is a distinction intended or unintended; is it relevant or irrelevant?

Non-discrimination law not only prohibits “direct discrimination,” but also “indirect discrimination.” Direct discrimination occurs when a person is treated differently than another person in a similar context.

A vacancy text states that only women can apply for a job. This is a form of direct discrimination based on gender, because men are denied the possibility to apply.

Indirect discrimination occurs when a seemingly neutral policy or practice affects one group of people more than others.

An automated advertising system for furnished rental homes in the Netherlands states that only “expats” are eligible for tenancy. Experience has learned that they are relatively trouble-free tenants. If an expat is defined as a person who is living abroad for a defined period of time, this criterion indirectly discriminates based on nationality: although there may be Dutch candidates who meet the criterion, expats are likely to have a non-Dutch nationality. Dutch nationals, then, are especially affected by such a criterion.

Ground

Marital status

Objective

Problematic proxy

Potential discriminatory effect

Gender

Objective

Problematic proxy

Potential discriminatory effect

Gender

Objective

Problematic proxy

Potential discriminatory effect

Religion

Objective

Problematic proxy

Potential discriminatory effect

List of examples

I want to ensure a fair distribution of scarce housing.

I program the system in such a way that, for the same housing type, an individual forming a two-person household pays less rent as compared to an individual in a single-person household.

Persons who fall into the “married/registered partnership” category are more likely to be part of a two-person household than those falling into the “not married/no registered partnership” category. The latter group will be negatively affected by this measure.

I am looking for strong employees for a physically demanding job.

A list of current employees shows that everyone is taller than 1.70 meters. I tell the algorithm to select résumés based on this criterion.

Women are generally shorter than men, making female candidates less likely to be selected by the algorithm.

As an insurance company, I want to minimise the financial risks of our unemployment insurance.

The algorithm gives points to low-risk persons, such as those who were continuously employed during the previous five years.

Women are more likely to pause their career, for example due to pregnancy, making them less eligible for the unemployment insurance.

I want to deploy a security system based on facial recognition.

I put a strict dress code in place, requiring the head and face to be uncovered, to ensure that the system can analyse faces.

Persons who wear head or face coverings for religious purposes are especially affected by the dress code.

*Disability /
chronic illness*

Objective

Problematic proxy

Potential discriminatory effect

I only want to provide a loan to people who will repay it.

I exclude persons who receive welfare from applying for a loan, because they often do not have a stable income.

Persons with a disability/chronic illness are generally highly represented in welfare programs. They will be especially affected.

Sexual orientation

Objective

Problematic proxy

Potential discriminatory effect

I sell erotic products and want to offer a discount during the PRIDE festival.

I offer a €5 discount to all customers who bought homoerotic products from me in the past.

Persons with a heterosexual orientation are less likely to buy homoerotic products, making them less eligible for the discount.

Age

Objective

Problematic proxy

Potential discriminatory effect

I want to ensure that I am viewed as an attractive employer by offering reimbursement for moving expenses.

The reimbursement is higher for employees who maintain a household than for those who do not maintain a household.

Persons under 30 are less likely to maintain their own household. Overall, younger people will receive lower reimbursements.

Nationality

Objective

Problematic proxy

Potential discriminatory effect

I only want to rent my furnished homes in the Netherlands to reliable tenants for short periods of time.

I instruct my algorithm to exclusively select "expats" for consideration as tenants.

Persons with the Dutch nationality are less likely to meet the "expat" criterion, affecting them negatively compared to other groups.

Nationality

Objective

Problematic proxy

Potential discriminatory effect

I want to attract good employees.

I let the algorithm select résumés based on diplomas from Dutch universities.

Most persons who hold a diploma from a Dutch university are Dutch, making non-Dutch persons less likely to be considered for the job.

Political opinion

Objective

Problematic proxy

Potential discriminatory effect

I want to achieve a harmonious work environment.

In recruitment processes, I let an algorithm scrape the internet, looking for data revealing whether a candidate has attended public demonstrations. If this is the case, the candidate is not selected for an interview.

Persons with strong political beliefs are more likely to be excluded from recruitment than others.

Race/ethnicity

Objective

Problematic proxy

Potential discriminatory effect

I want to attract candidates with a spontaneous personality for the vacancies I advertise.

I let an algorithm assess the spontaneity of candidates based on their application videos. I use my current employees to train the algorithm. My current employees are predominantly white.

The algorithm is less capable of recognising the desired qualities in non-white candidates.

It is important to note that indirect discrimination is not always prohibited – it is permitted when it can be “objectively justified.” This entails that there must be a legitimate reason to discriminate between groups of people, that such a distinction is proportional, and that there are no less discriminatory alternatives available to achieve the same objective.

The inclusion of a language requirement in a job offer, such as a good command of the native language, can be indirectly discriminatory. While the average native person will meet this requirement, it excludes many non-natives. But a language requirement may be justified, for example if the job involves intensive contact with customers.

Some of the previously mentioned points can be schematically summarised as follows:

1 - Awareness

Could the goal, design or outcome involve any of possible "suspicious" distinction?

- Marital status
- Disability/chronic illness
- Gender (incl. gender identity)
- Religion
- Age
- Philosophical beliefs
- Nationality
- Political affiliation
- Race/ethnicity
- Sexual orientation

→ The algorithm I use to test acceptance conditions gives a lower rating to people who are or have been long-term incapacitated.

→ My algorithm is trained on successful CVs. Only men work for me and nobody is under 18.

→ I want to build an AI system that filters out people with dual nationality in my data and mark them for extra verification.

2 - Distinction?

Does it lead to disadvantage?

→ Persons with disabilities/chronic illnesses may be excluded from my service.

→ The recruiter may not get to see CVs from women or people under 18.

→ The group of persons with dual nationality will be subjected to additional scrutiny and negative consequences.

3 - Can I justify my choice?

Do I have a good reason for the distinction made?

Example:

Labour market context

I am looking for persons to perform a dangerous job. Persons under the age of 18 are not legally allowed to do this work. I am allowed to use an age criterion.

However, I am not allowed to reject women for the job, while the algorithm may cause these CVs to be systematically undervalued. I have to correct for this.



1. Appropriate

- *Suitable* for achieving the legitimate aim (does it contribute to the achievement of the objective)?
- *Consistent* (free from internal contradictions?)
- *Coherent* (seen in the context in which measures have an effect)?

2. Necessary

Subsidiarity principle: Are there less intrusive means available that are equally effective in achieving the objective?

3. Proportionality

Proportionality principle: Are the aims pursued proportionate to the interests likely to be affected by application of the algorithm?

This handbook aims to assist people in choosing a data management architecture, building an algorithm and structuring processes in order to arrive at automated decisions. This means that the focus of this handbook is on the process that precedes the decision, while the focus in discrimination law is on concrete decisions and their effects. In this legal domain, the decision is then assessed based on three questions:

- Firstly: (1) was the decision made based on prohibited grounds (direct discrimination) or;
- Secondly: (2) does the decision have a disproportional negative effect on persons sharing specific characteristics that are covered by the prohibited grounds (indirect discrimination); and
- Thirdly: in case of (1) or (2), is this legitimate?

This handbook translates this ex post assessment (subsequent to a decision) into ex ante precautionary norms (preceding a decision). Two limitations arise from this approach. Firstly, adhering to the principles of this handbook does not rule out the possibility of a discriminatory effect; additional ex post assessments are always required. Secondly, the reverse also does not hold true: if the principles outlined in this guide are not respected, this will not necessarily result in a breach of non-discrimination law. If an AI system is used to predict astrological processes, for example, it is unlikely that discriminatory effects will be an issue.

It is of course important to realise that many existing decision-making processes, which do not involve the use of AI, are also biased, but that knowledge and data on this type of bias is currently lacking. Working with AI consequently does not necessarily entail a risk for discrimination – rather, it creates an opportunity to make processes more neutral and fair, and to achieve a better understanding of which groups are affected by specific practices, policies or decisions. Still, AI does carry the danger of engraining bias and discrimination into systemic structures, of which the consequences can be grave.

1. Diversity

Because non-discrimination law itself includes very few *ex ante* rules, this handbook also considers privacy and data protection law, statistical principles, organisational standards and technological best practices. Taken together, these domains provide array of principles needed to develop a non-discriminatory AI system. This handbook divides the relevant questions and principles into different phases, which are based on – but not identical to – the Cross-Industry Standard Process for Data Mining (CRISP-DM).

Four principles are leading throughout the handbook and must be taken into account during each phase (Diversity, Context, Verifiability, Evaluation).

The questions raised in this document cannot be answered in isolation. The document aims to serve as a bridge between the legal world and the technical world. While legal professionals are used to working with very abstract principles, such as “be transparent” and “avoid discrimination,” developing an AI system is about making practical choices. Which parts of the process must be transparent, for instance, and to whom exactly must the process be transparent: the system developer, a judge, or a non-expert? The same goes for non-discrimination: the choice to avoid one bias will often result in another bias.

Therefore, it is important that the team working on an AI project is as diverse as possible. Diverse in terms of expertise and professional backgrounds, but also in terms of personal backgrounds. Think of ethnicity, gender, sexual orientation, cultural and religious background, age and other aspects that may be relevant to the functioning of the AI system in practice.

2. Context sensitivity

AI systems ultimately have a real and concrete impact on society, often within a specific context. Therefore, the team must have specific domain knowledge. If AI-systems are not founded on real-world knowledge, a mismatch with reality may occur. This not only negatively impacts the system's effectiveness, but it can also cause or increase discrimination – for instance, such may occur when the developers of an AI system fail to recognise that a datapoint is likely to be a proxy for one of the protected grounds in discrimination law.

In addition, it is important to involve stakeholders in an early stage of the design process. Imagine that an AI system can correctly diagnose a disease in 80% of the cases, while doctors have a 60% accuracy rate. However, because the system was trained using male patient data, it has a much higher error rate in the diagnosis of female patients. This does not necessarily mean that the AI system cannot be used in clinical practice. But it does raise the importance of starting a conversation with patient organisations early on about the way the development and decision-making process is designed, and to discuss the mitigation of potential problems. Is it possible to collect additional patient data to complete the data set? Should the AI system only be used in the diagnosis of men? Could some of the saved time and resources be spent on extra medical staff for the diagnosis of women without the use of an AI system? Etc.

3. Verifiability

The process and each step within it must be clear, systematic and testable. Therefore, it is important to thoroughly document and explain each of the choices that are made. This makes it possible to check for errors and to update the system at a later stage. Ideally, documentation should be so detailed as to make each step repeatable and verifiable. Also keep in mind that the process must be transparent for different actors, who may want different types of information, provided in different forms: citizens/stakeholders, supervisory authorities, and fellow system developers carrying out a second opinion.

4. Evaluation

It is important to ensure internal and external quality control throughout the process. Such could be done by requesting a second opinion by external experts, doing tests on the same data with another algorithm, or going through the process twice using two different fairness definitions.

This handbook differentiates between six steps that are presented in a linear sequence, while building an AI system in fact is an iterative process. In reality, it may be necessary to jump from step 4 to step 2 and back to step 3, or you may start with components of step 5 before addressing the questions of step 1. Moreover, situations may occur in which the data have already been collected, in which the AI system is acquired from an external party, or in which the problem definition is already established.

The final of the six steps outlined in this document is the evaluation step. However, evaluating the AI system is something that needs to be done continuously. And because the AI system is a learning system that is always in flux, it needs to be evaluated constantly whether all requirements and non-discrimination principles are met.

Importantly, then, this guide is not intended as a checklist that can simply be ticked off. Instead, it aims to explain how you can make systematic and conscious choices when embedding non-discrimination principles in AI processes. Going through all the steps once does not mean the work is done – because the system changes once it is up and running, it is necessary to constantly evaluate how the AI system functions.

Let's get started!

Some final remarks. This guide is intended for an AI system project leader in making the right distinctions between different stages of the process, bringing together the right people at the right time, and letting them discuss the right questions. When technical experts and data analysts sit down with legal experts and the data protection officer, joined by the relevant stakeholders, domain experts and data stewards, the questions outlined in this document can guide their discussion. This document should be viewed mainly as a way to facilitate that discussion, and to ensure that all the relevant questions are posed at the right moment.

In addition, this guide can be of help to an organisation commissioning the building of an AI system – either in a preliminary phase to ask potential developers how they will address the issues outlined in this document, or during the development process to observe and think along, or after completion of the project to check whether the final product meets all the relevant requirements. Still, this document is written primarily for the teams that build AI systems.

The three most important aspects of the project – the legal, the technical, and the organisational aspect – will be discussed separately in this handbook. By keeping them separate, the reader will get a clear view of important matters in connection to each of these key aspects. These three aspects, however, do not exist in isolation of each other; they are complementary and must be regarded in a holistic manner. In other words, it is not a “pick-and-choose” model. All three components must be integrated into the AI project, because they complement and reinforce each other.

For each of the three primary aspects, this handbook contains questions and practical tips on how to minimise the risk of discrimination. Depending on the context of the AI project, some questions and remarks may be more relevant than others. Since this is a general guide that

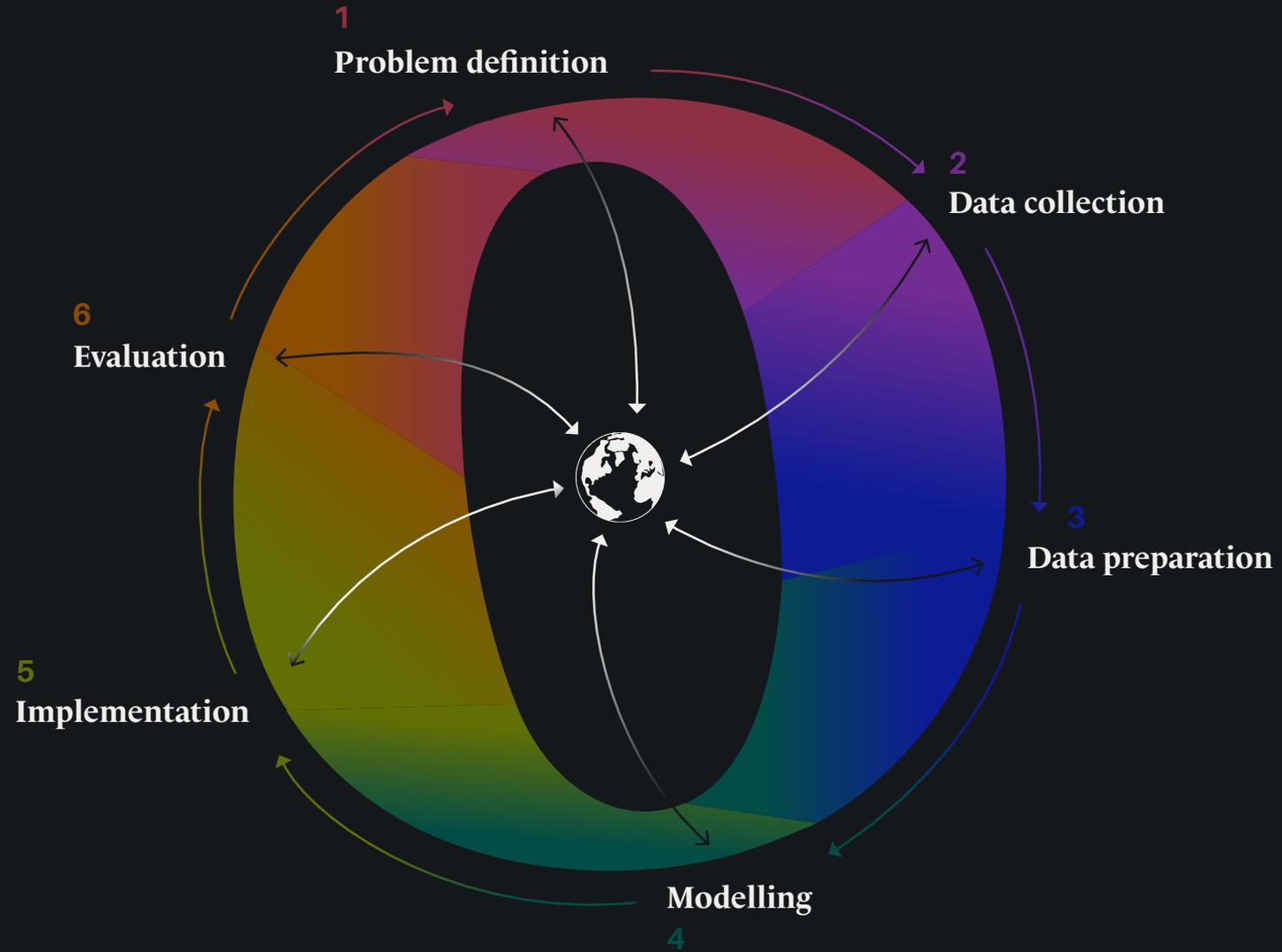


aims to be useful in different contexts, it is up to the reader to keep that in mind. However, to provide some insight into the ways in which different applications of AI result in different approaches, each phase is illustrated by hypothetical examples from three different fields: the labour market, the criminal justice system, and the medical field. Some but not all of the relevant questions are explored in three hypothetical cases. These cases serve as an illustration; they concern applications that speak to the imagination in a way that makes it easier to grasp the moral, legal and technological issues raised by the use of AI. They do not concern the most common applications of AI. The examples illustrate which questions an organisation might be trying to answer, and how – they are not best practices.

In short, this guide will take you through six different phases, with for each phase:

- The most important questions your team should discuss;
- An illustration of how you might address those questions in three hypothetical cases;
- And an elaboration of the most important questions related to the legal, technical and organisational aspects of the AI system

Non-discrimination by design



Phase

Strategy

1 - Problem definition

 Purpose & necessity

 Impact

 Success criteria

2 - Data collection

 Purpose & necessity

 Data quality

 Data storage

3 - Data preparation

 Inclusion & exclusion

 Integration & aggregation

 Labelling

4 - Modelling

 Pre-modelling

 Model(selection)

 Test

5 - Implementation

 Practical test

 Model alterations

 Application

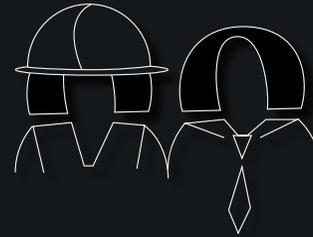
6 - Evaluation

 Evaluation preparation

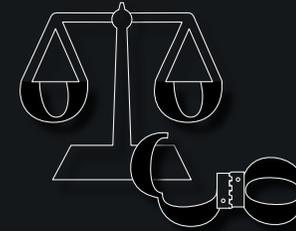
 Evaluation

 Points of action

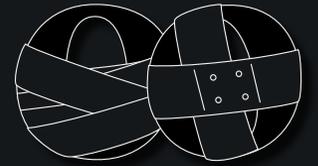
Example



Labour market



Criminal justice



Medical

Principle



Legal

Technical

Organisational

The starting point for the design of this document is based on three different manipulations of the letter "o". The "o" represents the subject that is the victim of bias. The design have been developed on the basis of a free interpretation of prominent (but non-exclusive) bottlenecks in the various domains in which the principles are classified.

The visual language has been developed to support the text and is used both functionally and aesthetically.

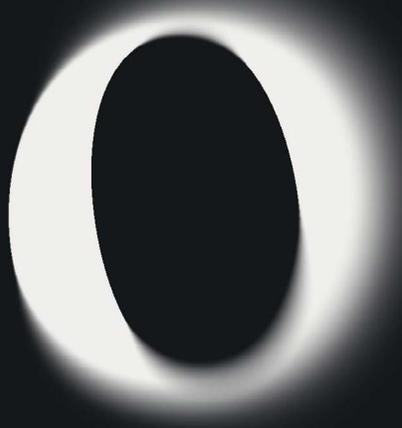
This concept is applied according to the following manipulations:

Design principles; a visual translation of the form on the basis of:

Legal: A one-sided or out of focus perspective.

Technological: An incomplete or selective data set.

Organisational: A guiding definition of success.



Legal



Technical



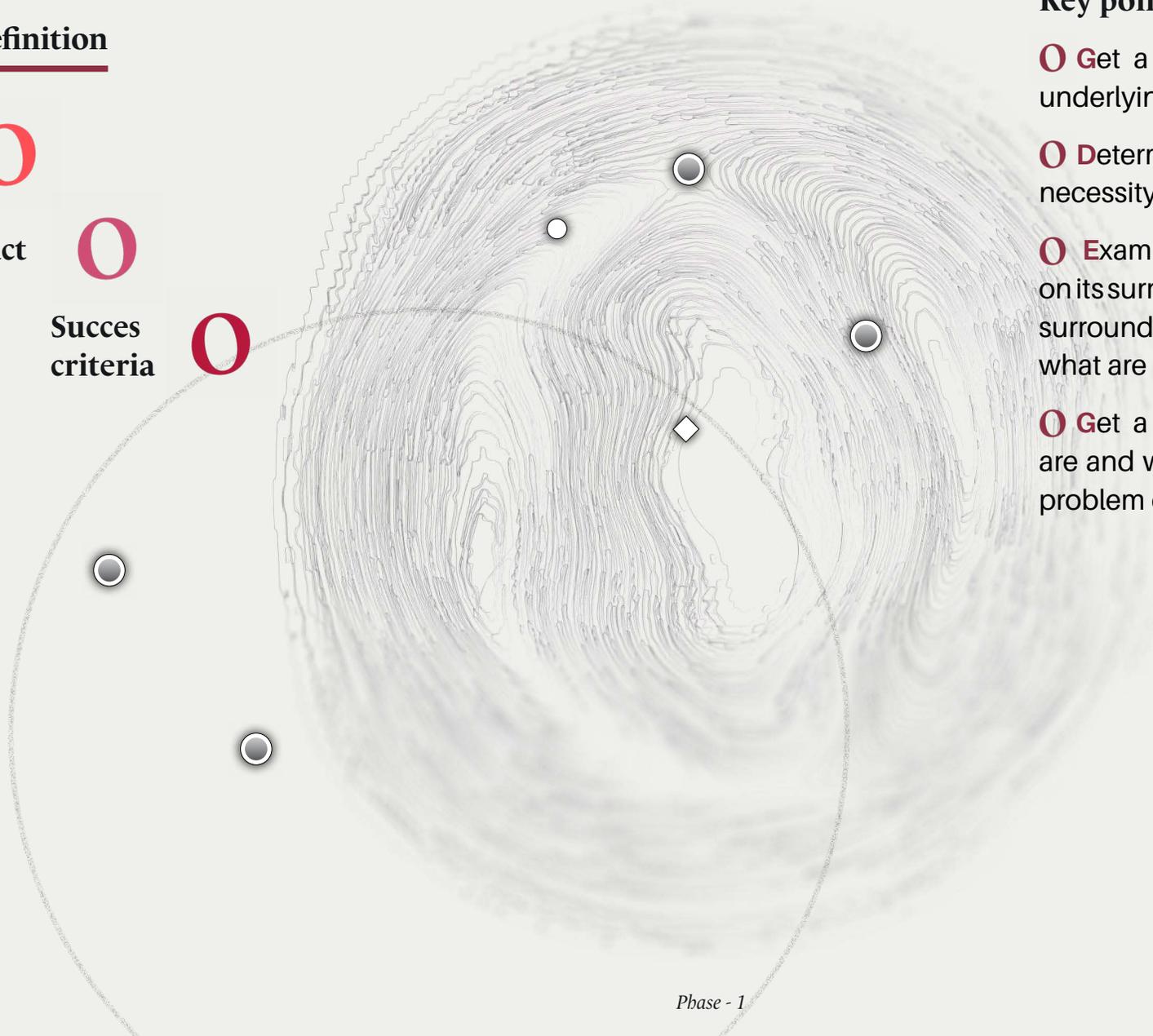
Organisational

Phase 1 - Problem definition

Purpose
& necessity ○

Impact ○

Success
criteria ○



Key points in this phase

- **G**et a clear view of the problem and the underlying assumptions.
- **D**etermine and formulate the purpose and necessity of the AI system.
- **E**xamine the influence the system will have on its surroundings and the people within those surroundings (what will change for whom, and what are the consequences for people?).
- **G**et a clear idea of who the stakeholders are and which groups are differentiated in the problem definition.

Purpose & necessity

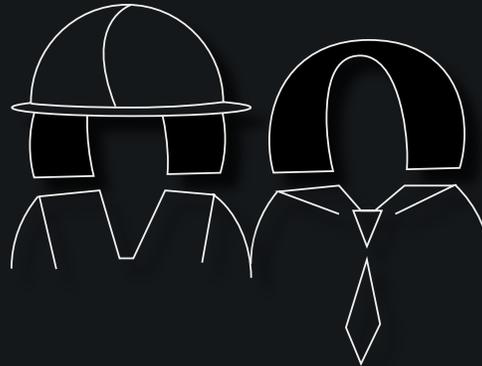
- 1. W**hat is the problem and how will AI help solve it?
- 2. I**s the use of AI necessary, or could the problem also be addressed without using an AI system?
- 3. W**hich groups are differentiated in the problem definition(s) and why?
- 4. B**ased on which assumptions about the various groups were the problem definition and the purpose of the system formulated?
- 5. H**ave the various stakeholders been heard?

Impact

- 6. D**oes this project require the collection of more data than currently available within the organisation, and what consequences would this have for citizens?
- 7. W**hat impact does the system have on citizens and society, both positive and negative?
- 8. D**oes the system serve to gain information, to aid in the preparation of decisions, or to make decisions autonomously? And what consequences does this have for the extent to which AI will be a determining factor in practice?
- 9. W**hich procedures are available to stakeholders to oppose to a decision?
- 10. W**hat is known about the occurrence of discrimination/bias in the existing processes? Can the implementation of the AI system have a positive impact in this respect, even if it is only by decreasing bias?

Success criteria

- 11. W**hat are the financial, computational and organisational costs of this system, and what would the costs be of a non AI-driven alternative?
- 12. W**hen is the AI system considered a success (for example, at which effectiveness percentage), and when must this benchmark be reached (for example, in 1 month or 2 years)?
- 13. W**hat percentage of false negatives and false positives is acceptable, and why?
- 14. W**hich fairness definition is chosen and why?
- 15. W**hat do the various success criteria mean for different groups?



The problem is that the manual assessment of letters of application is very time-consuming, and manual selection is biased (human bias). This system serves to make a pre-selection of eligible candidates based on application letters. Its purpose is to make the process more efficient and less biased by prioritising (ranking) the letters based on a number of pre-defined categories.

To train the system, existing letters from previous application procedures are used. The system makes automated decisions. False positives would entail a less efficient system; false negatives entail missed job opportunities for qualified applicants.

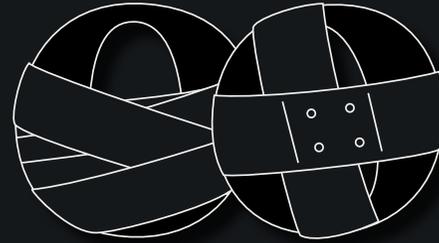
Success means 20% cost reduction and 5% less unjustified rejections as compared to the existing process. This benchmark must be achieved within 1 year. Maximum false positive rate: 40% (of candidates who are invited for an interview, compared to the existing process), maximum false negative rate: 2% difference (compared to the number of candidates who would be invited for an interview in a manual selection process).



The problem is that the police often does not arrive at the scene until after the fact. This system serves to predict where and when an offense will take place, and to conduct preventive patrol accordingly. Its purpose is to make the process more cost-efficient, more accurate and less biased.

Use of existing databases, complemented by data scraped from social media. The system makes predictions and informs the heads of the police force. False positives entail a loss of effectiveness, but potentially also a deterioration of support by the local community or the undermining of public trust in the police. False negatives entail a loss of effectiveness, as well as the undermining of the trust in the AI system of those who use it.

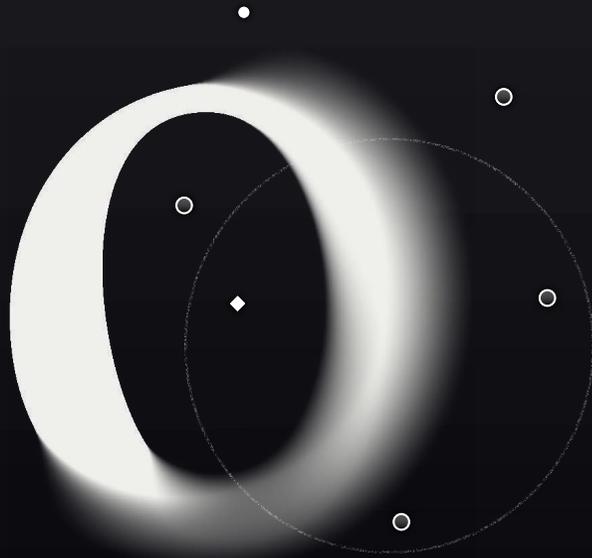
Success means a 10% increase of effectiveness of police patrols, measured by the number of arrests and charges. Must be achieved in 3 years. During a 3-year period, a number of police patrols are conducted based on the existing process, while for the planning of others, the AI system is used. The latter should not result in a higher false positive rate or false negative rate. To monitor false negatives, a random allocation policy is implemented.



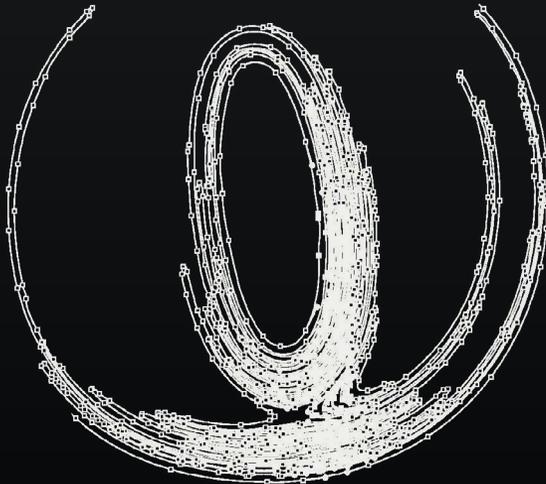
The problem is that the manual assessment of body scans is very time-consuming, expensive and not always accurate. This system serves to make better diagnoses of a specific cancer type. Its purpose is to make the process more cost-efficient and accurate.

Data collection is conducted during the trial phase of this system at various institutions across the world. The system makes automated decisions in clear cases, and leaves the decision with the doctor in case of high uncertainty. False positives cause fear; false negatives may, in extremis, cause death.

Success means a 10% higher accuracy rate of diagnoses as compared to the current false negative rate. The costs may not exceed the costs of the existing process. No end date has been set for this benchmark to be reached, but the system will be terminated immediately if an increase in the false negative rate is observed. Doctors continue to diagnose patients; in addition, an AI system makes diagnoses. An independent doctor compares and monitors the results.



- **A** clear and concrete goal for the AI-system must be defined beforehand.
- **S**o the predefined goals and/or parameters to differentiate between groups have potentially discriminatory consequences?
- **C**heck whether personal data are processed; personal data concerns any information that can be traced back to an identifiable individual. Click here for [further explanation](#).
- **A**n Impact Assessment must be conducted to determine the impact of the AI-system on discrimination and privacy, and how to mitigate those effects.
- If high privacy risks continue to exist after the implementation of additional measures, the Data Protection Authority must be consulted. Click here for [a model](#).
- **F**rom a legal perspective, AI systems must be necessary, proportional and respect the subsidiarity principle. This means that the means must be proportionate to the end, and there are no lower-impact alternatives available to achieve the same goal. Click here for [further explanation](#).
- **N**avigate through the diagram on page 10. Does discrimination occur, and if so, is there a legitimate reason to discriminate? Click here for [further explanation](#).



○ Explain and document the choice for AI in relation to the context.

For example: selection of target variable, classification task, performance goals, etc.

○ Explain and document the logic and the whys behind the AI system.

○ Motivate and document the choice for:

- **Evaluation metrics:** Do the metrics represent the interests of all stakeholders, including those outside the organisation? Take into account the impact of false positives and false negatives on different groups. Also consider incorporating a fairness metric.

- **Target variable.** Is the target variable a good measure of the concept that is predicted, or is there a measurement bias? In many cases, the concept is difficult to measure and a proxy is used. Instead of measuring crime, we measure arrests; instead of measuring the quality of personnel, we measure evaluation scores by their manager. The difference between the target and the proxy is a form of measurement bias. When this bias varies for different subgroups, it may result in a discriminatory model.

- **Explainability.** Is it necessary to use complex technologies such as deep learning, or does it suffice to use a model that is easier to explain, such as a decision tree or linear regression? To be able to discuss the fairness of the model, it helps to understand how the model comes to a decision.

○ Motivate the choice for a specific kind of AI system: is it a simple decision tree, a self-learning system or deep learning?

○ What is the purpose of the AI system: to gain insights, to aid in the preparation of decisions, or to make decisions autonomously.

○ Determine the success criteria, taking into account the false positive and false negative rates, the effectiveness and the duration of the project.

○ Conduct a technical evaluation. The Ethics Canvas, developed by the Open Data Institute, can serve as an example. Click here for [*the Canvas*](#).



🕒 **G**o through the following questions:

- Does the team have access to the necessary resources?
- Does the team have the necessary authorisations?
- Does the team have all the relevant expertise?
- Is the team as diverse as possible?
- Is a person with domain knowledge involved in the project?

🕒 **D**etermine within which domain the system will play a role and create a picture of the societal context.

🕒 **D**etermine who the stakeholders are, and who will gain the benefits and who will bear the negative effects of the system.

🕒 **I**nvolve the relevant stakeholders or representative (civil rights) organisations at an early stage.

🕒 **D**ocument all steps taken and choices made throughout the process, and discuss these choices both within the team and with stakeholders.

🕒 **I**n case of doubt, ask an external expert for a second opinion.

🕒 **S**et an end date for the process and determine the exit strategy. Determine which persons within the organisation will have which tasks and responsibilities.

Phase 2 - Data collection

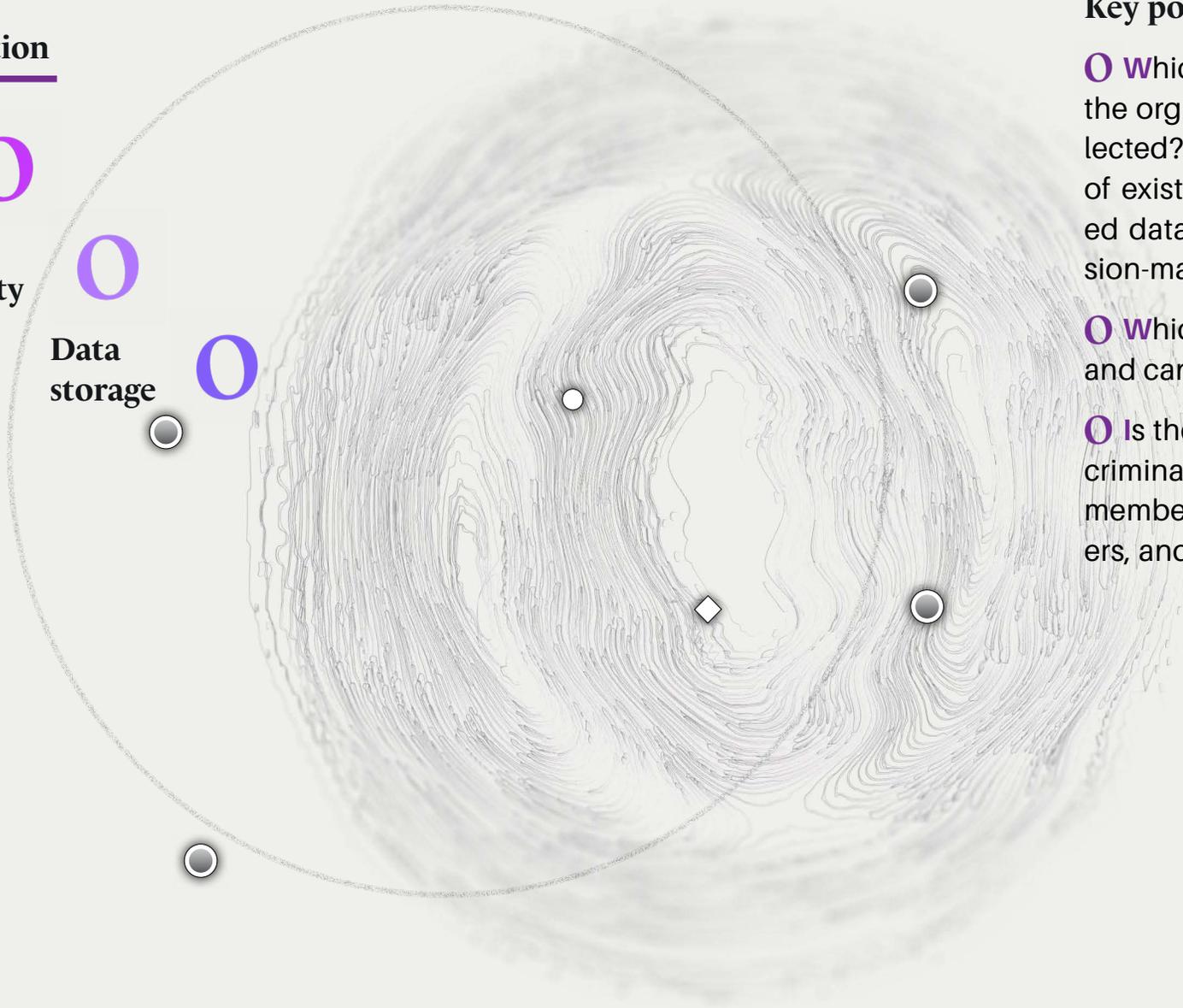
Purpose
& necessity



Data
quality



Data
storage



Key points in this phase

○ Which data are already available within the organisation, and what data must be collected? Is there a legal basis for the re-use of existing data or the use of newly collected data for profiling and/or automated decision-making?

○ Which bias can be found in the data, and can this bias be mitigated?

○ Is there a risk of misuse of the data for discriminatory purposes, either by internal staff members or external organisations or hackers, and if so, can this risk be mitigated?

Purpose & necessity

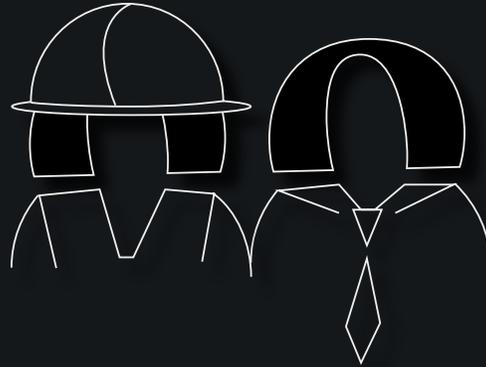
1. **W**hich data are required for this project and why?
2. **T**o what extent are these data already available within the organisation, and to what extent are externally collected data needed?
3. **I**s it permitted to collect and process these data for this project and if so, on what ground?

Data quality

3. **D**o the data contain bias and if so, what are the consequences?
4. **I**n what context were the data generated, and what are the assumptions that underly the representations?
5. **A**re the data representative and are all relevant groups represented equally?
6. **I**f multiple data sources are used, how is it ensured that these data are compatible and comparable? Is the methodology for gathering the data the same and if not, what will be the impact?
7. **C**an the merging of datasets lead to proxies and "disparate impact"?

Data storage

9. **H**ow long will the data be stored and how?
10. **W**ill the data be treated safely and confidentially; what consequences does a data leak have for specific groups or categories of persons represented in the data?
11. **W**ill data be shared with other parties, and what are the risks of misuse of the data resulting in negative consequences for groups or categories of persons?



Data from 30 random previous job application procedures will be assessed based on actual outcomes and the outcomes suggested by the AI system. Data are already available within the organisation. It concerns secondary use of data, which is legitimate because of the public interest in preventing biased selection procedures due to implicit or explicit human bias and prejudice.

The existing data set is significantly biased, inter alia, relative to the ethnicity and cultural background of candidates. Bias in the existing procedure is the exact reason for introducing this system. Therefore, the success of the AI system will not be determined by the extent to which the outcomes are similar to the outcomes of current process. An independent team will assess which selection method produces the fairest outcomes.

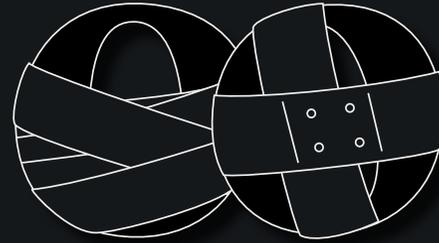
The data are kept for the duration of this project, after which they will be deleted. During the duration of the project, an internal cloud is used for the storage of the data. Only employees of the HR department and the developers of the AI system have access to the internal cloud.



Data on the effectiveness of previous surveillance and crime rates (available within the organisation). These data are linked to open data harvested from social media. The data are anonymised and aggregated to group level data. Further data processing is conducted based on statistical data as opposed to personal data.

The existing dataset available to the police contains a historical bias towards certain neighbourhoods, people with a migration background, low social-economic status, male gender, and other attributes. Because younger people are highly represented on social media, there will be an age bias in the social media data. The data are not representative or neutral. The two data sources are assessed separately. Corrections will be made when weighing datapoints to counter the bias.

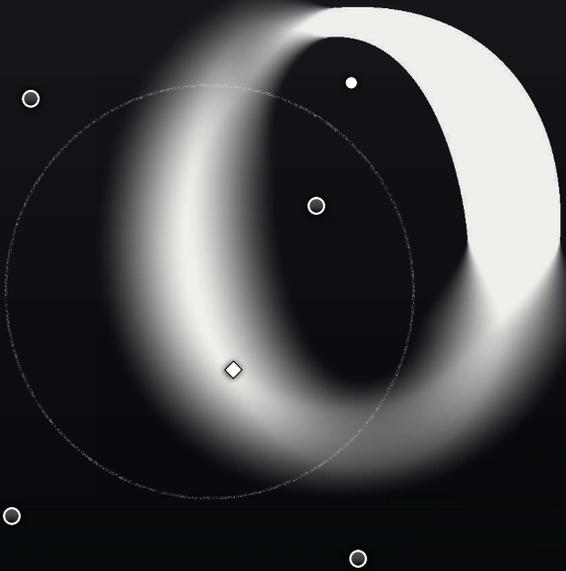
All relevant data are retained anonymously for an indefinite period; such is necessary because historical patterns may be valuable for future processes. The data are stored in a secured environment. Only members of a special unit can access these data, after having been given clearance.



As much data as possible are collected on diagnoses by doctors and second opinions by the AI system in order to assess how well the AI system performs compared to a doctor. Only data provided by patients are used. Patients give informed consent for the use of these data.

The patient population is skewed towards women aged 50+ with high BMI rates. Therefore, the AI system will learn to make accurate predictions for these groups in particular. Depending on the accuracy rates, this could mean that the AI system will be used to diagnose this particular patient group in the future, while other groups will be diagnosed by doctors. The data sources are assessed by a special committee of experts on the disease that is diagnosed, who will check for irregularities in the data.

Data are retained for a period of 20 years so that quality checks and assessments are possible. Only researchers and data analysts have access to the data. The data from different institutions will be linked and shared using a secured private cloud.



- Determine whether there is a legal basis for the processing of personal data, such as a legal obligation, informed consent, or public interest. Click here for [*further explanation*](#).
- The processing of personal data on race, sexual orientation, health, criminal history or other sensitive information requires explicit consent, a legal basis or an important public interest. Click here for [*further explanation*](#).
- If data are derived from external sources, it must be verified that these were gathered legitimately.
- Data must be accurate and kept up-to-date, and data subjects must be given the opportunity to provide additional data.
- Check your data for potential bias on the basis of the aforementioned grounds. You can do so using [*Aequitas*](#), for example.
- When data cross EU borders, for example when they are exported to the US or China, the organisation that receives the data must comply with the data protection rules applicable in the European Union.
- Click here for [*further explanation*](#).



- Ensure that data are collected in such a way that they can be used by the AI system.
- In doing so, determine the technical requirements on beforehand.
- Assess data based on:
 - Distribution of attributes (for example, the target attribute of a prediction task)
 - Relationships between pairs or small numbers of attributes
 - Results of simple aggregations
 - Properties of significant sub-populations
- Assess the potential presence of bias based on, inter alia:
 - Distribution of the target variable in sub-groups. Inequalities in present-day society are often reflected in the data, even if the data were generated in a neutral manner.
For example, when the relative number of “positives” is different for women as compared to men, this may point to historical bias.
 - Relationships between pairs or small numbers of attributes (features): does this lead to proxies?
For example, postal code may be a proxy for ethnicity.
 - Distribution of attributes and representation of relevant sub-populations. Representation bias occurs when particular parts of the input space are under- or overrepresented.
For example, when a dataset to train facial recognition software contains few (low representation) pictures of dark-toned faces (different distribution of attributes), the risk is that the system will produce suboptimal results for these groups.
 - The assumptions behind the data: are we measuring what we intend to measure?
For example, are sales numbers a good proxy for the sales skills of employees?
- Ensure the technical security and confidentiality of the data. Ensure that the data are encrypted and compartmentalised.

Organisational



- ① **E**nsure that the data collection is as neutral and objective as possible, and that it is conducted in the most transparent and verifiable way possible. The manner in which data are collected – how, by whom, where and using which techniques – can determine the neutrality and reliability of the generated data.
- ① **D**ocument this process meticulously, and keep information on how the dataset was built to ensure the testability and repeatability of the processes. Do the same for data obtained through third parties.
- ① **S**creen the data for missing values, accuracy and representativeness, and check whether the distribution is the same across subgroups. Explain differences and consider mitigating measures.
- ① **C**onsider collecting new data or revising the goal of the project: if you choose to do so, return to phase 1 (problem definition).
- ① **F**acilitate a conversation on the screening of missing values, accuracy and representativeness between the various parties involved (for example, data scientists, AI experts, product owner, project leader, supervisory authority).
- ① **E**nsure the organisational security and confidentiality of the data. Ensure that a limited number of employees can access the data, using a double-factor authentication procedure.

Phase 3 - Data preparation

Inclusion
& exclusion ○

Integration
& aggregation ○

Labelling ○

Key points in this phase

- Ensure that the criteria for data selection and the underlying considerations are clear and well documented.
- Examine how the process of data selection differentiates between different groups.
- Check whether the linking of data leads to proxies.
- When labelling the data, check for the presence of sensitive labels, such as those referring to ethnicity, sexual orientation or gender, and labels that may indirectly refer to these attributes, such as postal code. If present, is there a logical and legitimate reason for using such sensitive labels?

Inclusion & exclusion

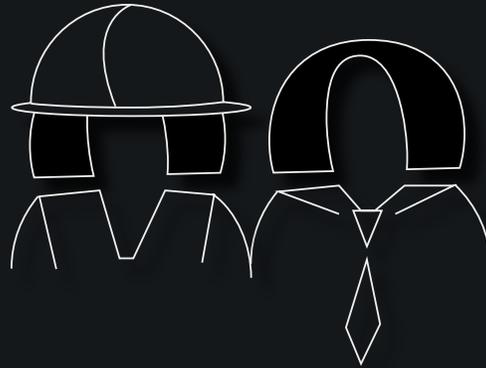
1. Which of the collected data are relevant for the model and why?
2. What happens with the data that are not used?
3. Which criteria are used for data selection and how do they reflect distinctions made between groups?
4. Does the selection of specific data or processes influence the problem definition?
5. Which aspects of the problem are not taken into consideration?

Integration & aggregation

6. How is it ensured that historical data and newly collected data fit together: are the data comparable, and what assumptions about groups and categories are inherent to the existing data and the data that is to be collected?
7. How are the data aggregated, and what consequences does this have for the representativeness of the data?
8. What does this mean for the representation of the problem and the stakeholders? For example, does this entail a reformulation of a group or category?
9. Does combining different data lead to proxies, and if so, which?

Labelling

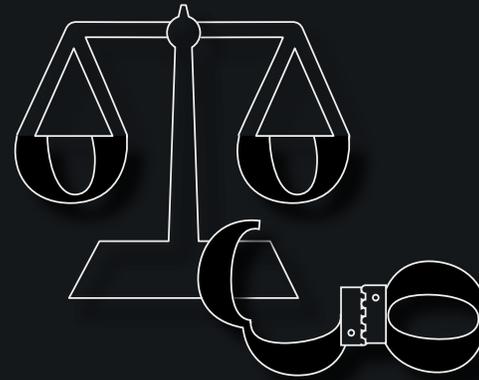
10. How are data labelled and why?
11. Is this in line with the way other organisations label data and use datasets on which the algorithm has been trained?
12. Is this in line with the way other stakeholders/citizens and domain experts would label data?
13. Does the dataset contain sensitive labels, such as those referring to ethnicity, sexual orientation or sex, or labels that indirectly refer to these attributes. If so, why?



Those data are selected that are relevant in the assessment of candidates: education, work experience and other activities. The job ad will list a set of minimal requirements that a candidate must meet. These requirements serve as the criteria used by the AI system. The data that are not used are stored so that, at a later stage, tests can be done to assess whether the system performs better with additional or different data elements.

New data will not map onto historical data, because the historical data contains significant bias. Data are aggregated based on education level and experience (junior, medior, senior), and the models are trained on these categories. It will be examined whether potential bias, either in the existing data or in the performance of the AI system, varies for junior, medior and senior levels.

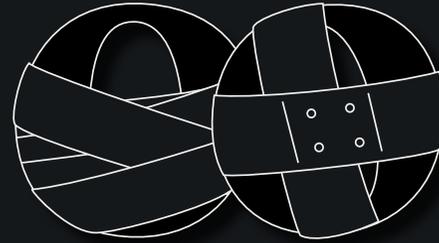
Data are labelled based on the aforementioned qualities of candidates, so that a candidate can be selected objectively. The way data are labelled is consistent with data labelling processes at other organisations. Sensitive data are present within the dataset but do not play a role in the labelling process. However, the categories may indirectly refer to group attributes; this will be tested.



Data are incorporated that indicate where most criminal activity takes place: location data, information about alleged or confirmed offenses, time data, suspects' background information. The selection of these criteria is based on reported burglaries/offenses and filed charge sheets. Non-used data are deleted.

All data that are used are integrated into a single format, to allow the comparison of specific aspects in the data. For example, historical arrest data contain data on time and place, nature of the offense and previous convictions of the offender. Uniformity in the data is created with regards to these dimensions, allowing for comparisons. Data are aggregated based on time and place, nature of the offense, nature and number of previous convictions of the offender.

Data are categorised based on neighbourhoods/postal districts, in order to assess which neighbourhoods require more attention than others. Data are also categorised based on period: time of day, day of the week, period of year. This way, data analysis can reveal, for example, that burglaries are likely to occur at night, on weekends or during holiday periods. There are sensitive labels present, especially postal codes. These can indirectly refer to specific ethnic groups.



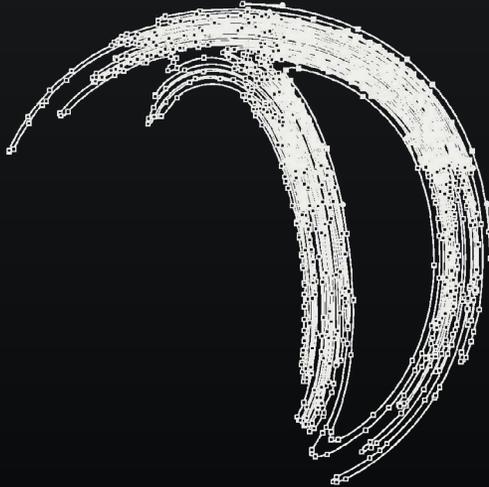
Data are selected that indicate an increased risk of this form of cancer: sex, age, social-economic status, lifestyle and others diseases. The selection of these criteria is based on historical data, which show that these datapoints are most relevant. Non-used data are deleted, in line with the GDPR, unless consent was given for data storage or re-use for the purpose of medical scientific research. Based on risk groups, further data selection is carried out and sub-groups are differentiated.

Historical data are assigned a certain weight. These historical data are incorporated in new cases. Data are aggregated based on age, sex and social-economic status. It has been agreed with other organisations to use the same medical and technical terms, so that there is consistency in data from different sources.

Data are structured based on categories such as age, ethnicity and sex, in order to examine whether this form of cancer is more likely to occur in specific groups. Sensitive categories are present within the dataset. This is necessary to determine among which groups high percentages of the disease are prevalent.

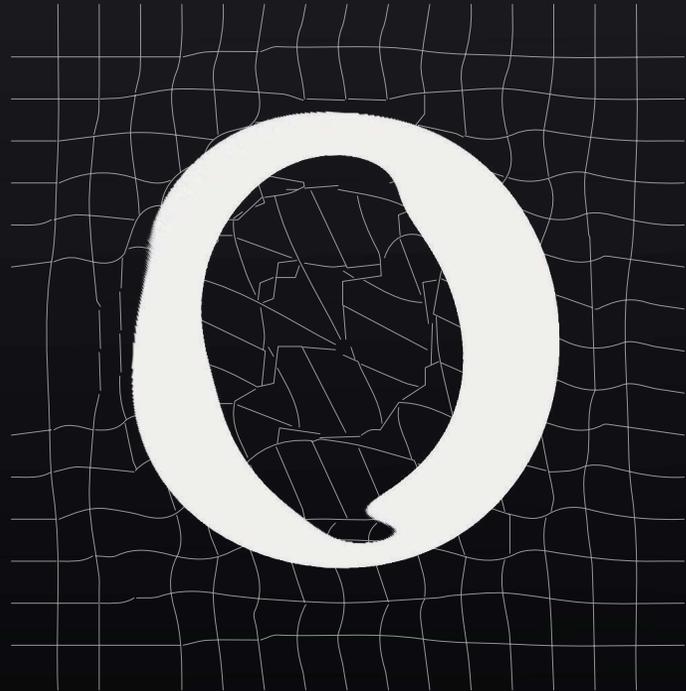
Legal

- When including and excluding data, consider how the data selection affects representation, especially for categories that indirectly refer to protected grounds such as marital status, sex, religion, sexual orientation, nationality, political opinion, race/ethnicity.
- When integrating and aggregating data, determine what consequences the integration and aggregation of data has for categories that directly or indirectly refer to protected grounds, such as marital status, sex, religion, sexual orientation, nationality, political opinion, race/ethnicity. Even the merging of two neutral databases can result in a significantly biased database.
- Determine what consequences the labelling of data has for categories that directly or indirectly refer to protected grounds, such as marital status, sex, religion, sexual orientation, nationality, political opinion, race/ethnicity.



- **W**hat are the technical limitations of the data? Make adjustments to the technical system if the quality of the data requires so. Does the data quality have consequences for phase 1 (problem definition)?
- **H**ave the various data that are used been collected in the same way?
- **H**ow do differences in the data collection (methodology, time, place, etc.) influence the integration and comparability of the data?
- **W**hich categories are selected in the structuring of the data, and why?
- **I**s it likely that the relationship between the features and the target variable varies across different groups? In this case, avoid a “one-size-fits-all” model, as it will not work well for any of the groups or only for the majority group. **Example:** the relationship between haemoglobin levels and diabetes varies across genders and ethnicities. If this is not taken into consideration in the modelling process, the predictions will not be accurate.
- **C**heck for measurement bias. Does the target variable (such as an evaluation by a manager) correlate with a certain construct (such as the quality of an employee)? The quality of the target variable can vary across groups, which can result in biased data.
- **A**lso keep in mind that the granularity and quality of data can differ between groups, and that each classification constitutes a simplification of reality.

Organisational



- **M**eticulously document the choices that are made concerning inclusion and exclusion. What rationale was adopted to include or exclude data, and why?
- **W**hat consequences do the quality and representativeness of the data have for the functionality of the system and the success criteria that have been set? If specific groups are over- or under-represented, consider collecting more data; in this case, return to phase 2.
- **D**atasets, method of analysis and decisions are selected with the goal of objectivity in mind; errors are documented and corrected immediately. If these errors are of relevance to other organisations, they are notified immediately.
- **M**ethods, procedures, definitions and classifications are applied in a consistent manner; they are standardised as much as possible to allow for comparison and verification.
- **E**xamine the influence of data labeling on the quality and representativeness and functionality of the system.
- **I**n case certain groups are over- or underrepresented, consider collecting more data: return to phase 2.

Phase 4 - Modelling

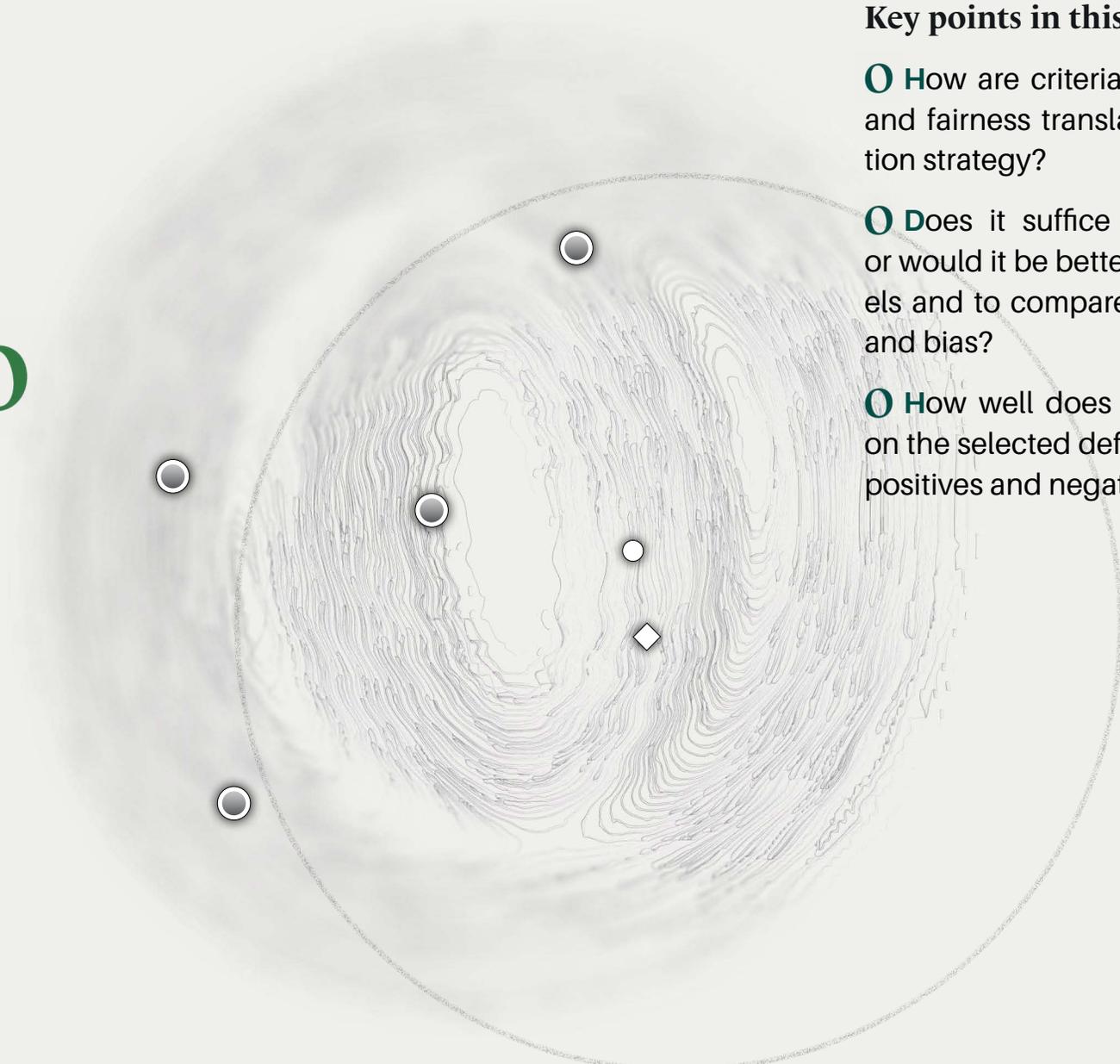
Pre-
modelling



Model
(selection)



Test



Key points in this phase

- How are criteria concerning explainability and fairness translated into the model selection strategy?
- Does it suffice to build a single model, or would it be better to develop multiple models and to compare them in terms of fairness and bias?
- How well does the model perform based on the selected definition of fairness and false positives and negatives?

Pre-modelling

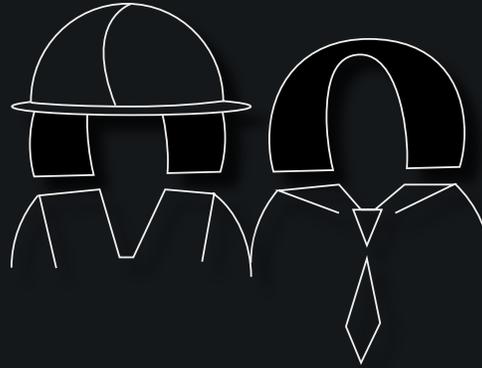
1. Which algorithm is selected and why?
2. What type of model will be built and why?
3. How are criteria concerning explainability and fairness translated into a model selection strategy?

Model(selection)

4. What parameters are chosen for the model and why?
5. Does it suffice to build a single model, or would it be better to build multiple models and compare them?
6. Is the model based on existing models and why (not)?

Test

7. How does the model perform on effectiveness?
8. How does the model perform on the selected definition(s) of fairness?
9. How does the model perform on the predetermined success criteria in terms of false positives and false negatives?



Supervised algorithms will be used to select a candidate based on rules that are incorporated and the scores of previous successful candidates. The choice is made to use a decision tree, because this is the most simple and efficient model to achieve the aim.

A single model is used, namely an existing model built by an external party, which has already proven its effectiveness in similar applications.

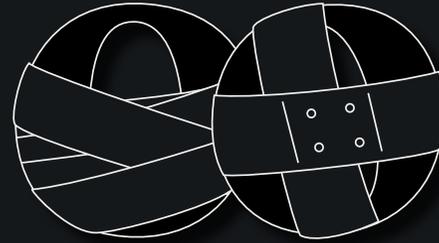
The effectiveness of the model is determined by the ratio of candidates who are categorised as “qualified” by the system against candidates who are also categorised as such by an external panel. The same applies to an a-select sample of candidates who were rejected by the system.



A supervised learning algorithm will be used to predict, based on time and location data, when a specific offense will take place. **T**he model produces a fair outcome if none of the suspects are disproportionately affected based on discriminatory grounds, either directly or indirectly.

Multiple models will be compared, due to the sensitive context. **T**he model is developed internally, considering the project's sensitive nature and the area of application.

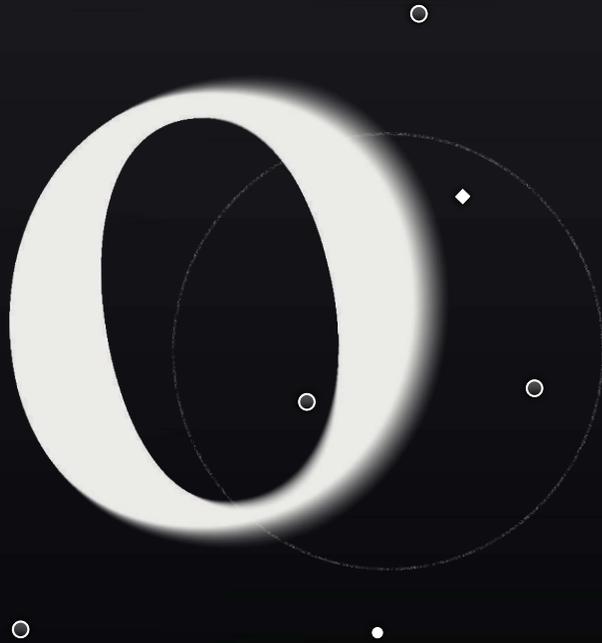
The model has a 70% effectiveness. **T**he model performs in accordance with the adopted notion of fairness. **T**he model performs within the acceptable range of false positives and false negatives.



A self-learning algorithm will be used, which can recognise certain patterns in historical data based on rules that are incorporated. The effectiveness and accuracy of predictions are ultimately more important than the explainability and testability of the AI system. However, providing patients with (understandable) information is of particular importance in the medical domain. Therefore, it is a minimum requirement for this model that the patient is able to understand it.

Three models will be developed and compared to assess effectiveness, accuracy and bias. The models are developed by three different research teams, operating within the various organisations involved. An external panel validates the functionality of the models.

The effectiveness of the model is determined by the ratio of patients categorised as “true positive” by the system against confirmed positive cases, and the ratio of patients categorised as “true negative” against confirmed negative cases. This is assessed by an independent team of doctors.



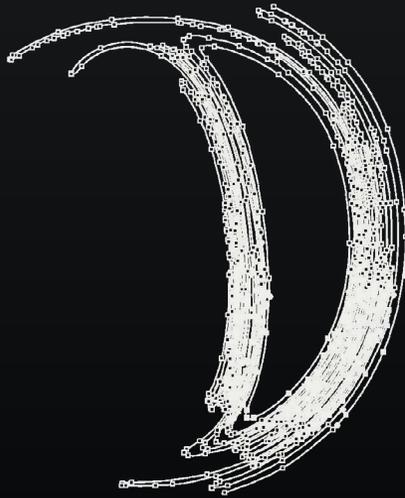
○ Is the model based on causation or correlation? Keep in mind that the legal domain is based on causality: legal explainability and legal justification can often not be found in statistical correlations. Therefore, assess whether the outcomes of a system based on correlation can be linked to causal explanations at an individual level. The use of deep-learning systems with black box elements in decision-making processes are nearly always prohibited if those decisions affect individuals.

○ AI systems based on statistical models can be evaluated on principles such as:

- Reliability
- Neutrality
- Objectivity
- Comparability
- Consistency

○ Find [more information](#) here.

○ Assess the performance of the model with respect to categories that directly or indirectly refer to an individual's gender, race, skin colour, language, religion, political or other opinion, nationality, cultural background or belonging to a national minority.



○ Algorithm selection:

- Explain and document the choices for an ML-algorithm, taking into account implementability and explainability.

○ Model selection:

- Consider the previously defined fairness principles when selecting a model.
- Check whether relationships found are consistent with existing domain expertise, and not random.
- Use the simplest model possible to achieve the performance objectives.

○ Make a selection of potential AI algorithms based on the requirements of fairness, interpretability and explainability. **Example:** when fairness criteria have been set, consider using in-processing or post-processing bias mitigation methods to optimise for fairness.

○ Consider using *unfairness mitigation* technologies to optimise for a fairness metric.

Click [here](#) and [here](#) for tools. **Example:** when causality is a condition, ensure that the algorithm candidate pool contains immediately interpretable algorithms, such as linear regression.

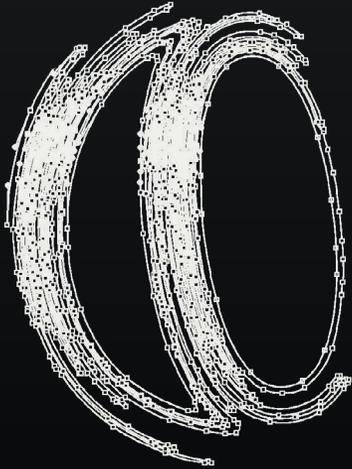
Example: when predictions serve to support human decision-making, ensure that the algorithm candidate pool contains methods that are explainable to people without a technical background, such as a simple decision tree.

○ Translate the success criteria into technical standards, such as standards for accuracy, false positives, false negatives and fairness. **Example:** to avoid the model working less accurate for minorities, the fairness metric equalised odds will be included in the model selection.

○ If existing models are used, include these in the model selection strategy. **Example:** in a NLP application, the use of pre-trained word embeddings is considered. Therefore, evaluating the embeddings with respect to the success criteria is a part of the model selection strategy.

○ Bias reduction:

- Consider using bias mitigation methods.



- If necessary, apply post-processing techniques to reduce bias after training the classification model. White-box methods alter the model; black-box methods alter the predictions.

Example: When predictions serve to support human decision-making, ensure that the algorithm candidate pool contains methods that are explainable to people without a technical background, such as a simple decision tree.

○ **Assess** the performance of the models with respect to categories that directly or indirectly refer to an individual's gender, race, skin color, language, religion, political or other opinion, nationality, cultural background or belonging to a national minority.

○ **Select** the simplest model possible to achieve the specified performance objectives.

Example: since both a logistic regression model and a random forest model are sufficiently accurate, the logistic regression model is selected.

○ **How** does the model perform on effectiveness, the selected definition of fairness, and the success criteria for false positives and false negatives?

○ **How** would the system function if a different model, fairness definition and/or algorithm were chosen? Adjust the model based on the outcomes.

○ **Check** for evaluation bias; this can occur during the testing of the model. Make sure the success criteria used to assess the system match the target group.

○ **If** the model uses personal data, incorporate a comparison of the performance of the model and the performance of the current decision-making process in the evaluation strategy.

Example: the performance of the model is assessed in a pilot study in a small group of users.

○ **Evaluate** the model based on:

- Effectiveness
- Fairness
- False positives and false negatives

Technical

Organisational



- Implement improvements.
 - Determine what the context of the selected test case tells you about the general functioning of the system, and consider to what extent different application domains come with different context sensitivities. Take suitable measures accordingly.
-

- The model selection strategy is formulated, documented and made public. Ideally, the design is universal so that the models can be compared more easily in terms of outcomes and fairness.
- Involve stakeholders, such as end users and decision-makers, in the selection of the algorithm candidate pool to ensure that the interpretability and explainability of the models matches the background of the users.

- **Explanation:**

- Metadata are stored and documented;
- Data are made available to third parties when possible;
- The model must be explainable and understandable to stakeholders;
- What form of explainability is offered by the system?
- For whom is this explanation understandable?

- Ask an independent team of experts, with diverse personal and professional backgrounds, for a second opinion.

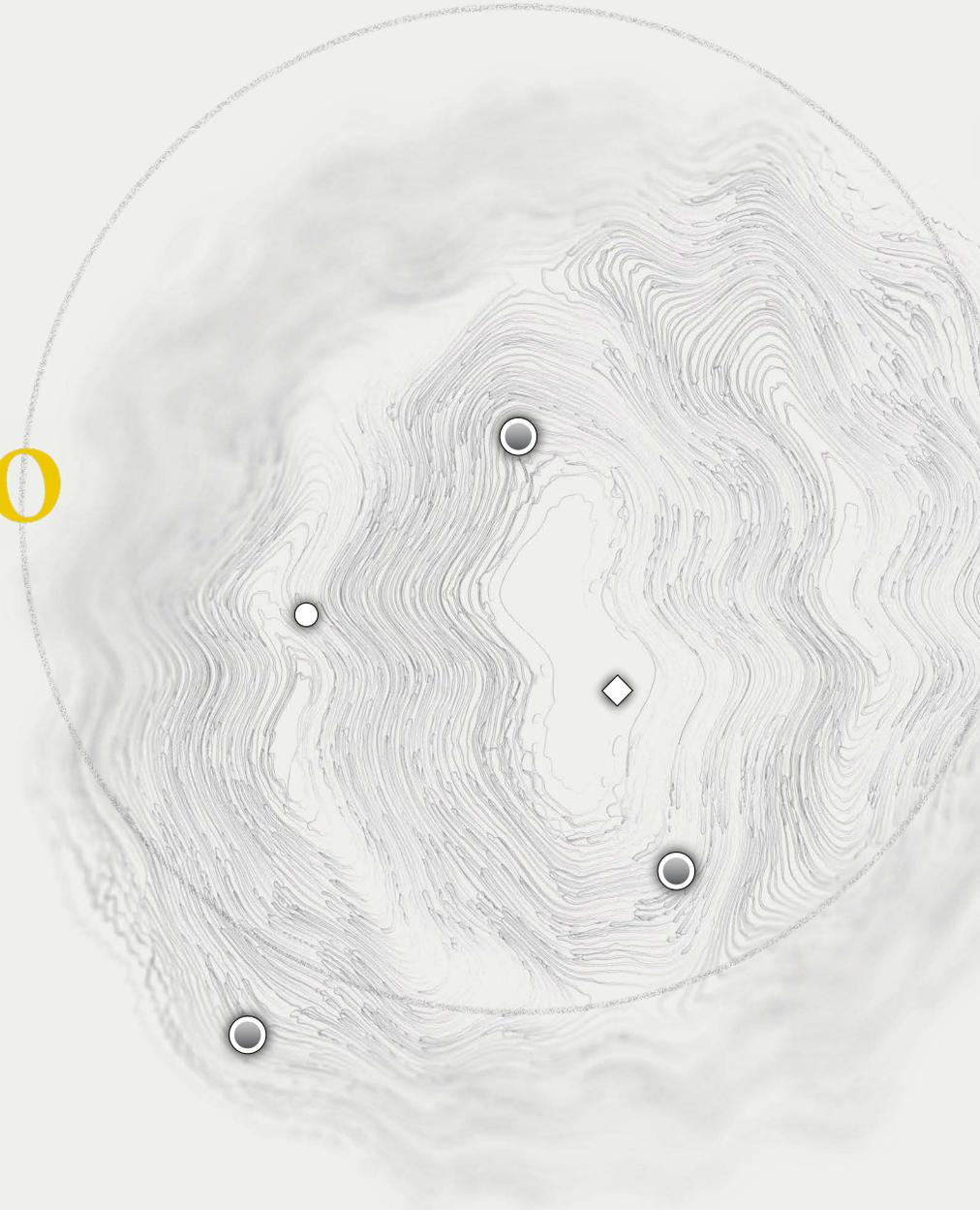
- Document the model selection and the results of the evaluation – for example, by using a model card.

Phase 5 - Implementation

Practical test **O**

Model alterations **O**

Application **O**



Key points in this phase

- O** Choose a specific application to test the system; ensure that this clearly defined application is representative of the entire domain in which the AI system will eventually be used.
- O** Adjust the model based on the results of the test case.
- O** Adjust the expectations regarding the applicability of the system based on the test case; for instance, which potential applications are found to be not feasible?
- O** Document the possibilities and limitations of the system, and inform users about the conditions under which the system might be deployed.

Practical test

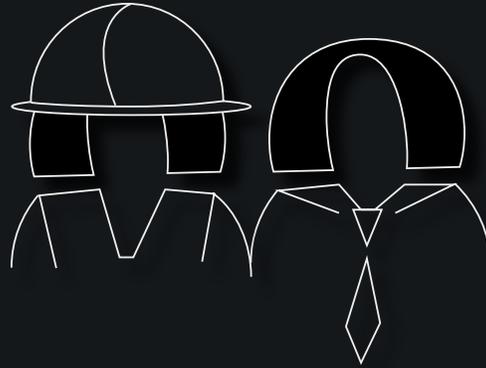
1. **W**hat is the application strategy?
2. **W**hat clearly defined and demarcated test case is representative and easy to monitor?
3. **H**ow does the model function, and is this in line with expectations?

Model alterations

4. **W**hat alterations are needed to improve functionality?
5. **W**hat alterations are needed to increase the model's fairness?
6. **W**hat alterations are need to reduce the error rates?

Application

7. **W**hat limitations arise from the previous steps with respect to the model's application potential and the implementation process?
8. **W**hat should be the key points of attention when deploying the AI application, and how can these be monitored in the implementation process?
9. **H**ow will stakeholders and others be informed and involved?



First, the letters of application and résumés are made quantifiable. Subsequently, the relevant variables are identified. The candidates are then compared based on these variables, only using those variables that do not lead to undesirable bias. As a representative test case, an evaluation is done of current employees who were selected by the system, assessing whether the employer is satisfied with the true positives.

The model is adjusted as time progresses and more data can be derived from application procedures. Due to the increase of information that becomes available, the performance of the model can be improved. As more data are obtained from recruitment processes, the fairness of the model should also increase. The variables that lead to bias are filtered out, excluding them from the process.

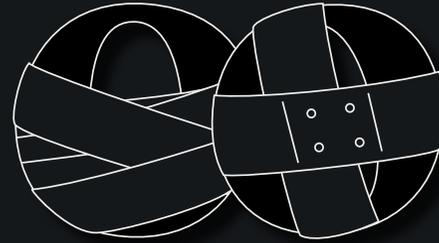
A résumé contains limited information, and quantifying a cover letter is difficult and can be done in more than one way. Therefore, the number of variables is limited and the weighing of these variables is subjective. A second test will be performed, in which the data are enriched with candidate information from open data sources. Whether this model performs better will be assessed subsequently. Stakeholders are informed about their rights and the use of an AI system prior to the process.



Since the model predicts where and when a crime is most likely to take place, police will patrol at those times and places. The algorithm is not trusted completely, in the sense that the outcomes are acted upon with caution. Other neighbourhoods will not be ignored. After some time, an evaluation will be conducted of how accurately or poorly the model predicts crimes. As a test case, the number of burglaries are monitored during a specific period in a demarcated area.

More data are needed to increase the predictive power. Ideally, these would be non-sensitive data. An audit of the system should indicate which data do not have predictive value. Subsequently, these data should be taken out.

The test showed that the model was good at predicting common criminal offenses, such as theft, but not at predicting heavier crimes, such as homicide. Therefore, the decision has been made to use the system, at least initially, only for the most prevalent criminal offenses. After three years, another evaluation should indicate whether the system has become better at predicting heavier crimes.



The risk groups are determined based on historical data. As a representative test case, an assessment is done of current cases at a hospital in Amsterdam, which has a varied patient population.

The model does not require much adjustment, because many historical data are available from earlier cases at the hospital. More detailed categorisation can enrich the data, which may lead to improved prediction. This also applies to the error rates.

Because the historical data are exhaustive, few limitations arise. However, this also increases the risk of overfitting. Overfitting occurs when too many rules are incorporated in the model, leading to incorrect judgements. The risk of overfitting will be reduced by the use of pruning. Stakeholders are informed about their rights and the use of an AI system prior to the process.

Legal

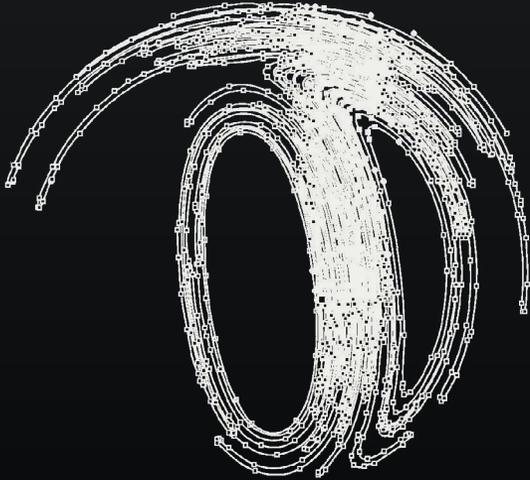
○ Inform stakeholders about their rights:

- Right to information, including information on the algorithm
- Right to challenge the decision
- Right to provide additional information
- Right not to be subjected to automated decision-making

○ Click here for [more information](#).

○ Assess the performance of the model vis-à-vis categories that directly or indirectly refer to protected grounds, such as marital status, sex, religion, sexual orientation, nationality, political opinion, race/ethnicity.

○ If the model significantly discriminates, directly or indirectly, based on one of these grounds, is there a justification for this? If so, which? Consult a lawyer about this.



○ Answer the following questions:

- What is the application strategy?
- What clearly defined and demarcated test case is representative and easy to monitor?
- How does the model function and is this in line with expectations?
- What is the exit strategy?

Consider user testing to assess the ease of use and accessibility for people with a disability.

○ Monitoring and maintenance

○ If the model is updated regularly to include new data, ensure that these new models are also thoroughly evaluated.

○ Plan for the monitoring of changes in the data distribution, such as concept drift and shifts in the demographics of data subjects.

○ Evaluate how the model performs on the benchmarks concerning:

- Effectiveness
- Fairness
- False positives and false negatives

○ Implement improvements.

○ Determine what the context of the chosen test case tells you about the general functioning of the system, and consider to what extent different application domains come with different context sensitivities. Take suitable measures accordingly.

○ *Monitoring and evaluation:*

- Schedule evaluations periodically
- Document the results of use
- Check exit criteria
- Make APIs available to external auditors
- Make data sheets and model cards publicly available to the extent possible

Organisational



- ① **I** Inform relevant parties and stakeholders that an AI system is implemented in a test setting. If possible, do this prior to their first dealings with the system and its consequences.
- ① **D** Discuss the effectiveness, fairness and accuracy of the model with both stakeholders and external experts.
- ① **I** Implement their advice and suggestions to the extent possible.
- ① **D** Document alterations made to the system.
- ① **D** Document which model has been chosen and what the outcomes of the evaluation are.
- ① **I** Inform all parties when the AI system is implemented outside of the initial test setting, and establish a procedure for complaints.
- ① **R** Request an independent team of experts, with diverse personal and professional backgrounds, to conduct a second opinion.

Phase
6 - Evaluation

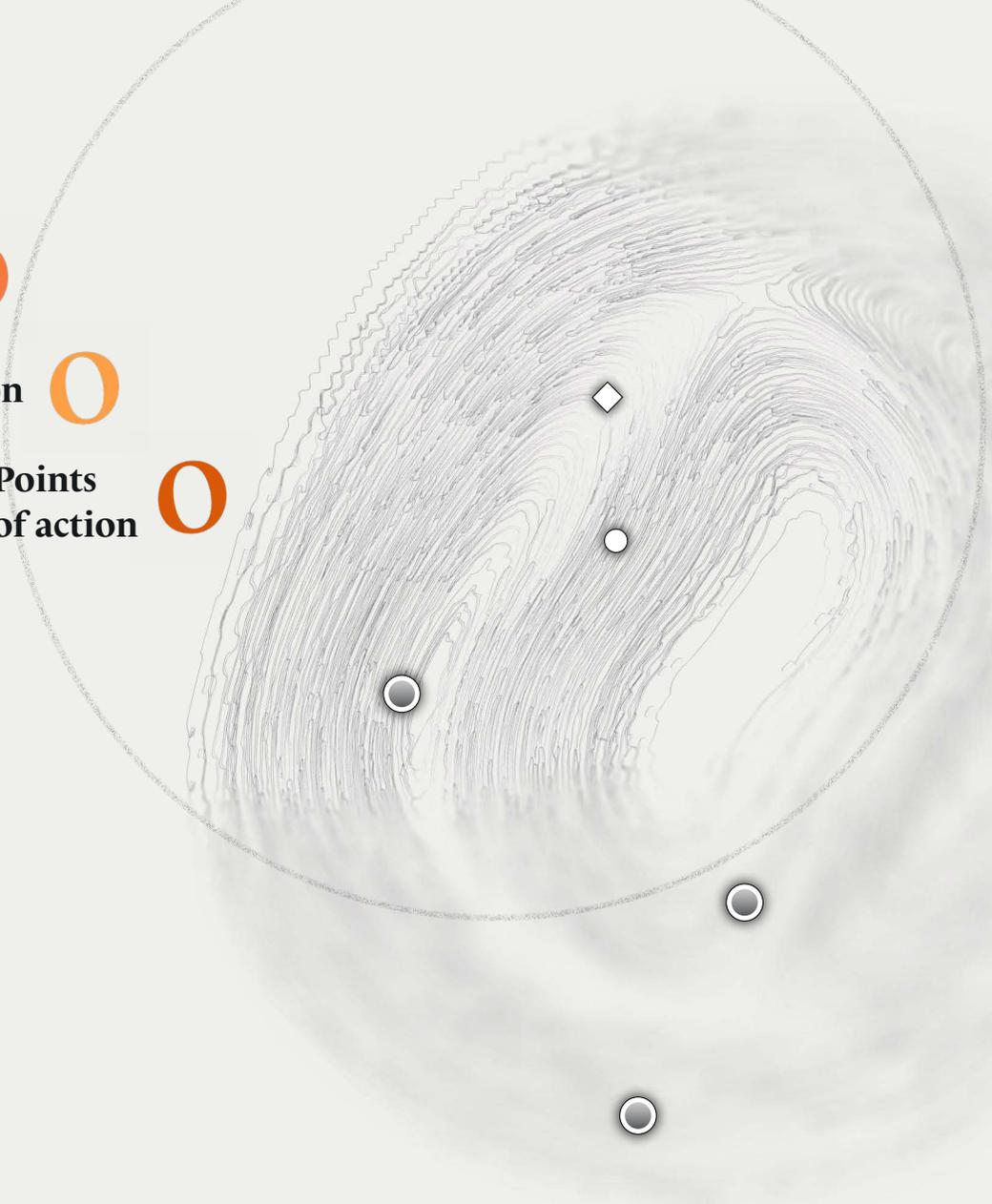
**Evaluation
preparation**



Evaluation



**Points
of action**



Key points in this phase

- O** Pick an implementation strategy and formulate an evaluation strategy. Preferably, involve external experts in the evaluation process.
- O** Assess how the system would function in case a different model, fairness definition and/or algorithm had been chosen.
- O** Determine whether, based on the evaluation, the system should be put on hold, improved or implemented.

Evaluation preparation

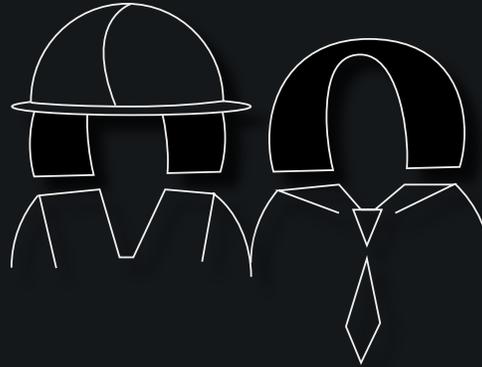
1. Will evaluation take place continuously, periodically or both?
2. Will evaluations be conducted internally, externally or both?
3. How will the evaluation be assessed, and based on which measurement points?

Evaluation

4. How does the system perform with respect to the success criteria?
5. Which improvements are needed with respect to the protected categories?
6. How would the system perform if another model, fairness definition and/or algorithm would be adopted?

Points of action

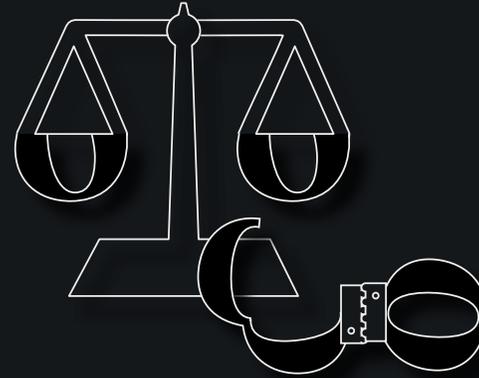
7. Should the system be (temporarily) put on hold?
8. Can observed problems and obstacles be solved?
9. How are the evaluation results perceived by stakeholders and external experts?



Evaluation will take place periodically at specific moments, each time after a vacancy has been filled. At a later point in time, an independent team will review the candidates with the highest scores and filter out errors.

The previous and the current situation are compared to each other, zooming in on how satisfied the management is with the selected candidates and the differences between the old and the new candidates. The protected grounds are incorporated as restrictions in the model, and therefore left out of consideration. However, these data will be stored separately, so that it is possible to check for indirect discrimination.

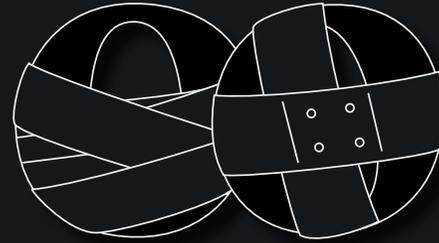
The system will be (temporarily) put on hold if unwanted bias is detected.



A team of specialists will continuously evaluate the system, its outcomes and potential complaints. This team consists of two lawyers, two data analysts and two former police officers.

Parallel tests were conducted with different models, different algorithms, and different fairness definitions. These tests show that, in general, the selected definitions and models contain the least bias. However, the other definitions and models do perform better on some aspects. Therefore, the system is slightly adjusted.

Evaluating the perspectives of stakeholders is difficult in this case. However, two civil rights organisations will be asked to give their critical opinion. Internal evaluation showed that the system can be used in its current form, but that it is necessary to monitor proceedings permanently.



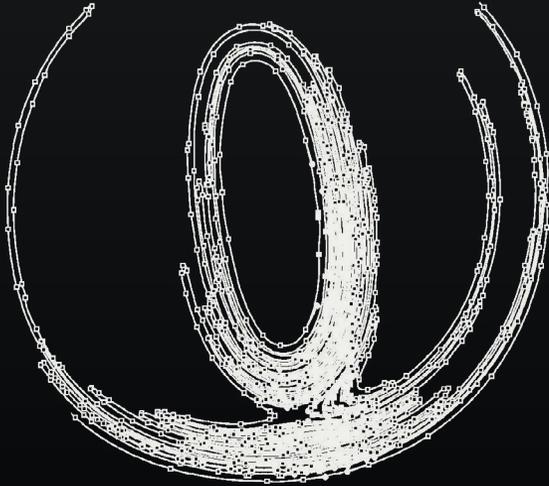
Because of the constant stream of new patient data, it was decided to evaluate continuously. **A**t a later point in time, an internal evaluation will take place, during which doctors and statisticians will check the system for unjustified bias.

It is assessed whether identifying the risk groups results in more patients receiving successful treatment. Based on the results, the possibility of extrapolating the predictions of the system to prevention policy will be assessed

The system will be (temporarily) put on hold if it appears that the system makes poorer judgments than the current average, after which the variables that cause the errors in predictions will be identified. **D**etected problems and obstacles can be addressed by incorporating more restrictions.

Legal

- If personal data are processed, the AI system must be clearly more effective than the status quo in order to meet the requirements of necessity, proportionality and subsidiarity.
- Explain why this is the case.
- Assess the performance of the model vis-à-vis categories that directly or indirectly refer to protected grounds, such as marital status, sex, religious conviction, sexual orientation, nationality, political opinion, race/ethnicity.
- If the model discriminates, directly or indirectly, based on one of these grounds, is there a justification for this? Consider whether or not the system should be (temporarily) put on hold, or whether it is necessary to return to one of the earlier phases of the process to make adjustments.
- Ask an external lawyer for a second opinion and advice on the legal requirements of the system.



○ **M**eticulously document the following:

- How is the evaluation conducted, and why in this manner?
- Who is responsible for the evaluation, and why?
- Which measurements points are selected, and why?

○ **E**valuate the model in terms of:

- Effectiveness
- Fairness
- False positives and false negatives

○ **D**oes the performance of the model meet the proposed success criteria? If not, (temporarily) put the project on hold immediately. If so, are there ways to further improve the system in terms of its effectiveness, fairness or accurateness?

○ **H**ow does the system perform:

- On a different data set?
- Using a different algorithm?
- Using a different definition of fairness?
- Using a different model?

○ **A**sk an external data analyst for a second opinion and advice on the technical system.

Organisational



- ① **I** Involve stakeholders in the evaluation. **Example:** conduct a survey or facilitate a focus group discussion with stakeholders to acquire their experiences.
- ① **E** Explore whether the system, the model, the data and/or the evaluation can be made public, either in anonymised form or not.
- ① **D** Document which evaluation methods are used, what the motivations behind the various choices are, and where the responsibility for the evaluation lies.
- ① **A** Ask an organisational expert for a second opinion and advice on the procedural organisation of the system and the team.



Colophon

Text:

- **Bart van der Sloot** (Tilburg University)
- **Esther Keymolen** (Tilburg University)
- **Merel Noorman** (Tilburg University)
- **The Netherlands Institute for Human Rights**
- **Hilde Weerts** (Eindhoven University of Technology)
- **Yvette Wagenveld** (Tilburg University)
- **Bram Visser** (Vrije Universiteit Brussel)

Design:

- **Julia Janssen** (artist)
- **Suzan Slinger** (production manager) of Studio Julia Janssen

This handbook was commissioned by
the Dutch Ministry of Internal Affairs.