Non-discriminatie by design

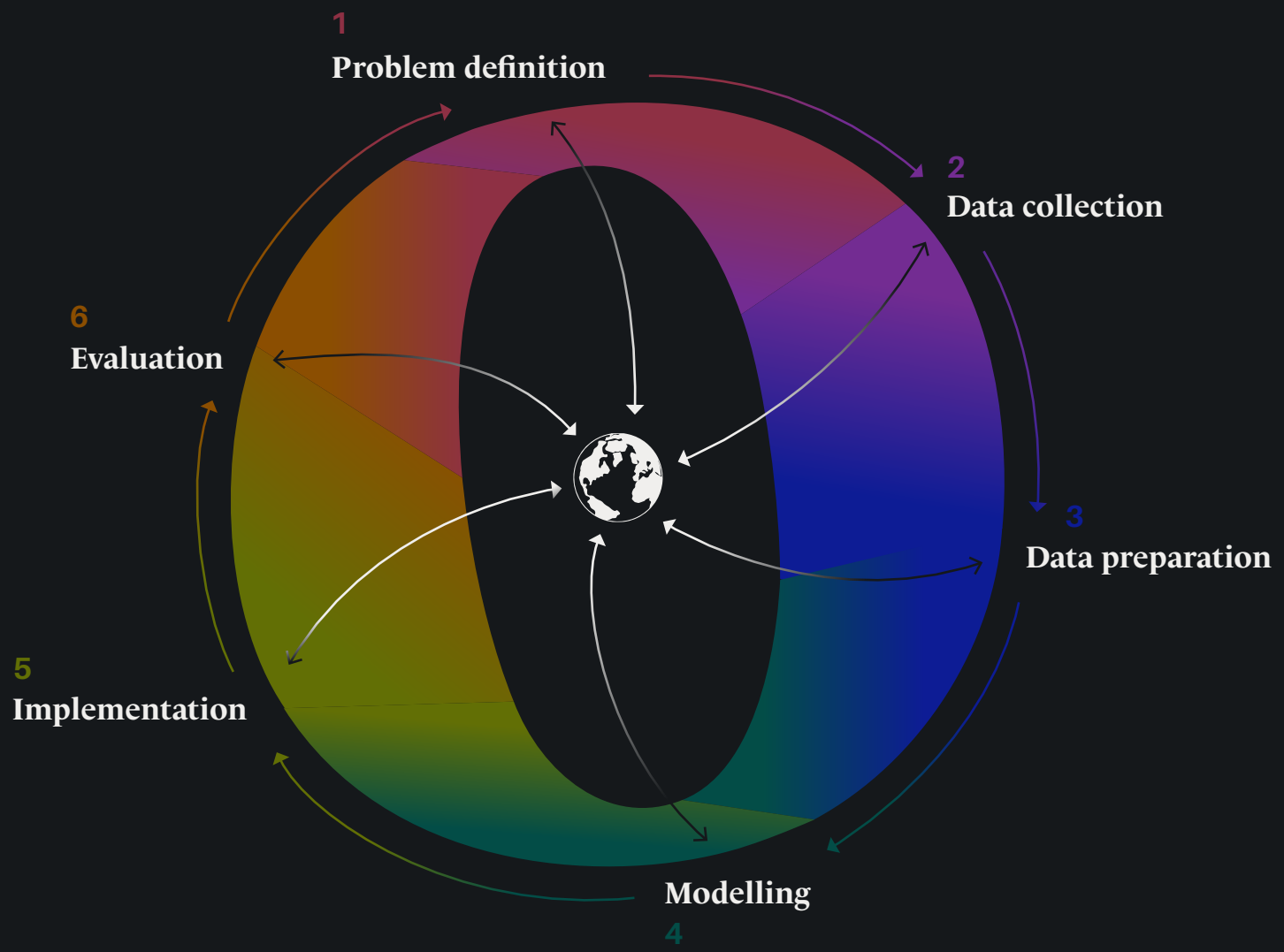
Text

- **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 Wagensveld**
  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.



Summary

This is the summary of the Handbook Non-Discrimination by Design. The handbook 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?

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 favors male candidates. How can we develop systems that are designed to minimise the risk of producing unintended and unjust distinctions between groups of people?

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.

This document distinguishes six steps that can help structure the development of an AI system. Three clusters of questions have been formulated for each step. Per cluster, technical, legal and organisational aspects are discussed. Importantly, non-discrimination law provides basic principles, but no absolute prohibitions. The law provides a standard or a starting point, but there are always exceptions. In addition, a judge will always consider the context, or what lawyers call "the circumstances of the case." Consequently, this handbook will not provide absolute rules or definitive answers on what to do or not to do. Rather, the most important thing is to be aware of the danger of discrimination and to assess whether the distinction between groups is necessary and fair.

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. An example might be a vacancy text holding 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 example may be an automated advertising system for furnished rental homes in the Netherlands which holds 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.

Direct discrimination is often quite overt. As mentioned, it is prohibited in most, but not all, circumstances. In the case of indirect discrimination, it is more difficult to determine whether such discrimination occurs and, if so, whether it is justified. It involves the use of characteristics that are indirectly related to the protected grounds: proxies. Because this indirect form of discrimination is most prevalent in AI systems, some examples will be presented below to provide a clear picture of what indirect discrimination based on different grounds may look like.

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.

This also holds true for direct discrimination. Discriminating between people based on sex is prohibited, for example, unless iti s 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.

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? 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. 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.

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 under-valued. I have to correct for this.

Good reason

↓

Legal exceptions

↓

Objective justification: legitimate aim

↓

Is the means that I use to reach my goal:

→ **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?

Purpose & necessity

1. What is the problem and how will AI help solve it?

2. Is the use of AI necessary, or could the problem also be addressed without using an AI system?

3. Which groups are differentiated in the problem definition(s) and why?

4. Based on which assumptions about the various groups were the problem definition and the purpose of the system formulated?

5. Have the various stakeholders been heard?

Impact

6. Does this project require the collection of more data than currently available within the organisation, and what consequences would this have for citizens?

7. What impact does the system have on citizens and society, both positive and negative?

8. Does 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. Which procedures are available to stakeholders to oppose to a decision?

10. What 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. What are the financial, computational and organisational costs of this system, and what would the costs be of a non AI-driven alternative?

12. When 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. What percentage of false negatives and false positives is acceptable, and why?

14. Which fairness definition is chosen and why?

15. What do the various success criteria mean for different groups?

Purpose & necessity

1. Which data are required for this project and why?

2. To what extent are these data already available within the organisation, and to what extent are externally collected data needed?

3. Is it permitted to collect and process these data for this project and if so, on what ground?

Data quality

3. Do the data contain bias and if so, what are the consequences?

4. In what context were the data generated, and what are the assumptions that underly the representations?

5. Are the data representative and are all relevant groups represented equally?

6. If 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. Can the merging of datasets lead to proxies and *"disparate impact"*?

Data storage

9. How long will the data be stored and how?

10. Will 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. Will 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?

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 is 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?

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?

Practical test

1. What is the application strategy?

2. What clearly defined and demarcated test case is representative and easy to monitor?

3. How does the model function, and is this in line with expectations?

Model alterations

4. What alterations are needed to improve functionality?

5. What alterations are needed to increase the model's fairness?

6. What alterations are need to reduce the error rates?

Application

7. What limitations arise from the previous steps with respect to the model's application potential and the implementation process?

8. What should be the key points of attention when deploying the AI application, and how can these be monitored in the implementation process?

9. How will stakeholders and others be informed and involved?

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?