Summary of the Report *Non-discrimination by design*

This document provides a scientific summary of the research project AI and Non-Discrimination by design. It includes a description of the research steps, the literature consulted, the methodology, the outcomes and findings. The intended readers of the report are scientists from various relevant disciplines (lawyers, AI specialists, ethicists, etc.), as well as policy makers who are interested to know more about how the Handbook Non-Discrimination by design came into being, why certain choices were made, and how those choices were motivated. The Handbook itself is a hands-on guidebook for people and organizations involved in the development of AI systems or their implementation, and may also provide commissioning parties with relevant criteria when outsourcing a project. The contents of the report are structured as follows.

<table>
<tr><th colspan="3">Chapter 0 (Introduction)</th></tr>
<tr><td colspan="3">The introductory chapter provides a:

- Description of the reason for this research project
- Description of the nature and purpose of this research project
- Analysis of the need for this research project
- Description of the definitions used in this study
- Overview of the selected structure and methodology of this study</td></tr>
<tr><th>Chapter 1 (Problem analysis)</th><th>Chapter 2 (Analysis of existing standards)</th><th>Chapter 3 (Mapping of problems related to existing standards)</th></tr>
<tr><td>Through a review of the literature, complemented by the outcomes of two surveys, an overview is given of which problems are described in relation to AI and non-discrimination.</td><td>Through a review of the literature, complemented by the outcomes of the surveys, an overview is given of the existing AI and non-discrimination standards in law and regulation, jurisprudence, directives and literature.</td><td>Using three test groups, complemented by the outcomes of the surveys, it is examined to what extent the existing standards are adequate, and which points require improvement or adjustment.</td></tr>
<tr><th>Chapter 4 (Developing concept handbook)</th><th>Chapter 5 (Validation of concept handbook)</th><th>Chapter 6 (Finalizing the handbook)</th></tr>
<tr><td>Based on the first three steps, the research team develops a first draft of the handbook.</td><td>This concept is tested through three workshops. The first workshop focuses on the use of AI in the criminal justice system, the second on the use of AI in health care. The third workshop is an interdisciplinary discussion between experts.</td><td>Based on step 5 and the input of external experts, the concept is adjusted and the design and the text of the handbook are finalized.</td></tr>
</table>

Chapter 0 – Introduction

Description of the reason for the research project

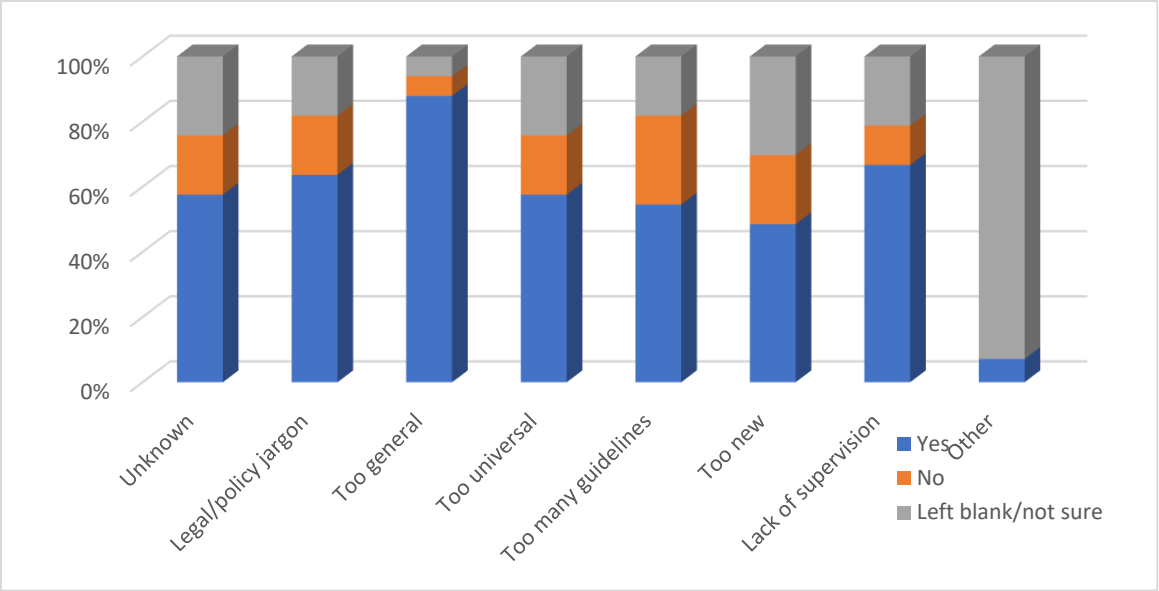
Algorithmic decision-making is increasingly embedded in organizational structures; such carries potential (legal) consequences for individuals and society. With the emergence of Big Data, giving rise to Artificial Intelligence (AI) and Machine Learning (ML), data methodologies and technologies advance rapidly. The increase in statistical possibilities and storage capacity allows for the analysis of large amounts of data and the discovery of patterns in those data, which can in turn be used in decision-making processes.

Description of the nature and purpose of the research project

While AI and ML can lead to discrimination in various ways, they can also be used to combat or decrease existing discrimination. In the policy letter "AI, public values and human rights" of October 8, 2019, the Dutch Minister of Interior expressed her intention to assess how public values and human rights can be operationalized in AI system, starting with non-discrimination. The present report follows from that letter, and analyses if and how AI systems can be developed to safeguard the right to non-discrimination in a society that will become increasingly interwoven with the use of automated decision-making. For an adequate enforcement of the prohibition of discrimination in AI systems, it is necessary to translate the legal framework into practically applicable system principles. In other words, a translation must be made from norms to concrete design strategies that can be used in the development of AI by governmental organisations, institutions and private organizations.

Analysis of the need for this study

Both nationally and internationally, the consensus is that current shortcomings in AI and ML development can lead to social stratification, societal inequalities and dire situations for individuals in the immediate and longer-term future. The starting point for this report is the observation that the existing guidelines have had insufficient practical effect. To examine whether there is indeed a gap between legal principles and policy documents on the hand and the practice of system developers on the other, and what might cause this gap, a question on this matter was included in the two surveys that were disseminated in the context on this study. The results show that, in all response categories, the "agree" answer option was selected more often that the "disagree" option. It is also clear that nearly all respondents see the fact that guidelines are currently of a too general nature as an important cause of the gap between policy and practice. Other factors that are commonly mentioned as causes for the hiatus between legal reality and technical practice are the too universal and abstract terms in which current guidelines are presented, the lack of supervision and enforcement, and the fact that the guidelines are riddled with legal and policy jargon.



Description of the definitions used in this study

To prevent discrimination by AI-based systems, a number of characteristics are of particular importance. Firstly, the new forms of uncertainty brought about by learning systems particularly raise concerns over (unintended) discrimination. If a task cannot be completely and explicitly defined beforehand and techniques are used that autonomously derive patterns and correlations from data, biases may occur that are not easily and immediately recognizable or detectable. Secondly, systems that aim to achieve a complex goal by various data sources can quickly become difficult to interpret and to control. Thirdly and lastly, many AI systems have a great level of independence (autonomy) in executing their tasks, without requiring direct human supervision or control. This autonomy can be another reason for new uncertainties and unpredictabilities to occur, which may amplify the risk of discrimination. Therefore, this report places emphasis on systems that:

- have a certain level of autonomy;
- have a certain level of complexity, which makes it difficult to oversee how the system comes to a certain result;
- and/or can adjust themselves (or models) based on an analysis of previous actions and the environment/new data.

Methodology for this study

Chapter 1	Chapter 2	Chapter 3
➢ Literature research ➢ Surveys	➢ Literature research ➢ Surveys	➢ 3 test groups ➢ Surveys
Chapter 4	**Chapter 5**	**Chapter 6**
➢ Research team's own reflection	➢ 3 workshops	➢ Research team's own reflection ➢ Creative design

Chapter 1 – Problem analysis

Building blocks

Some of the building blocks of both society and a data-driven way of working inherently require categorization and differentiation based on those categories.

Language	Data	Datafication
Language is a collection of categories and concepts derived from reality; linguistic concepts and categories, in turn, shape the way people perceive reality.	Like language, data are a representation of reality. By whom the data are collected, for what purpose, in which way, etc., significantly affects what the dataset will look like.	Personalization is a misleading concept within the context of AI. Categories and correlations are discovered, after which persons, objects or phenomena are placed in those categories.
Redlining	**Discrimination grounds**	**Trade-off**
All data, categories and correlations indirectly refer to the protected categories defined by the law (race, sexual orientation, etc.) – they do so by definition, the only question is how strong the derived correlation is.	The question is whether the existing categories as defined in anti-discrimination and equal treatment legislation and jurisprudence (race, sexual orientation, etc.) are the most relevant categories when assessing AI decision-making processes.	AI is about differentiating. Prohibiting predictions based on race or sexual orientation, and the other protected grounds, as well as on datapoints that indirectly refer to those grounds, will lead to less accurate predictions.

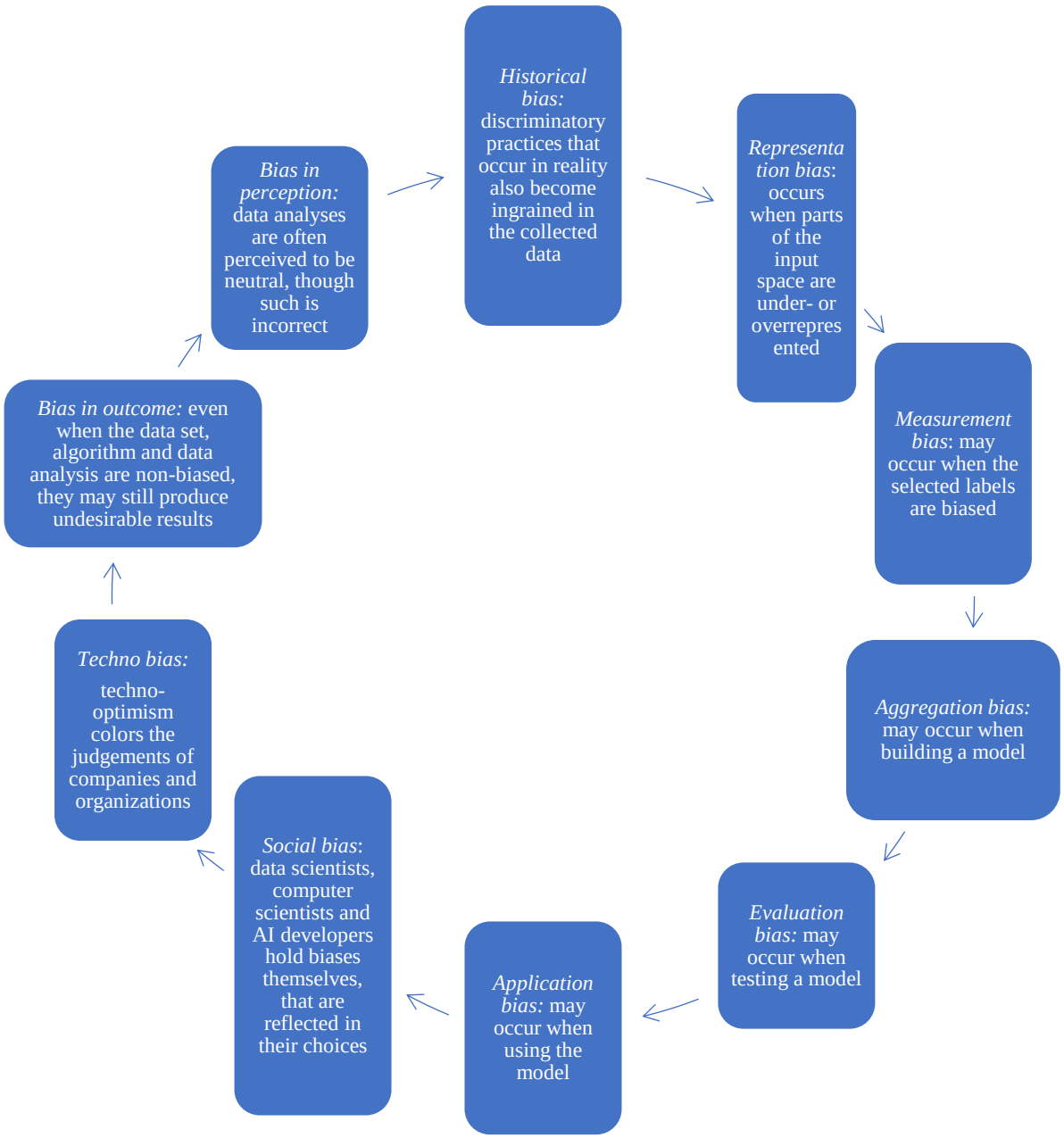
Explainability

The explainability of AI systems is necessary to allow for the evaluation of decision-making processes, to be able to conduct checks and to examine potentially discriminatory effects. There are a number of obstacles:

- Firstly, what lies between input and output is often unknown. Initiatives exist to develop glass box ML models that are inherently explainable, and post-hoc explanations for black box models are being developed. However, these solutions have limitations.
- A second problem is that explainability does not necessarily result in a meaningful explanation. To increase the robustness of an AI system and to further develop and improve it, certain tools provide an explanation to technical experts. However, such an explanation may not be meaningful to people without a technical background.
- Thirdly, financial costs are an obstacle to the explainability and transparency of AI systems.

Bias

In the ML and AI literature, discrimination is predominantly associated with the presence of "bias." Bias is a broad concept, which can refer to a number of specific problems that may occur throughout the lifecycle of an AI-based system. Bias necessarily occurs in machine learning systems. Different types of bias can be differentiated, such as:



Fairness

To ensure that algorithms produce fair outcomes, technologists have proposed several formal definitions of fairness to compare the distribution of outcomes across groups or individuals. There are different definitions of fairness pertaining to different levels, such as group parity or individual parity. While the first is defined by equal outcomes for members of different groups,

the second is about maximizing the accuracy of the outcome for each individual. The varying definitions of fairness in the ML literature point to a wider issue that is not merely technical in nature. What is considered equal treatment may vary, not just depending on the context but also depending on the (political, societal, philosophical, etc.) views about what entails an equal distribution. It is incorrect to think that choices as to fairness-unfairness or equal-unequal treatment can be translated into a mathematical, formal definition which is independent from social, legal and political contexts. There are a number of pitfalls:

- *Portability Trap*: Failure to understand how repurposed algorithmic solutions designed for one context may not be suited for another context;
- *Formalism Trap*: Attempt to capture social concepts (such as fairness) in mathematical models, while those concepts are procedural, contextual and contestable;
- *Ripple Effect Trap*: Failure to understand how the incorporation of technology into an existing social system changes the behaviour and embedded values of that system.

Technical challenges

There is a number of additional technical difficulties associated with the prevention or reduction of discriminatory effects of AI:

- "Bias in, bias out" phenomenon: The availability of data on some groups is simply greater than on others. There may be many data on who was arrested, but little data on who committed the offenses; there may be many data on who does or does not repay the loan, but no data on who would or would not have repaid it if they had been eligible to apply for the loan.
- Distorted model development: The data and phenomena that models reflect can be quite complex and heterogeneous. Traditional ML routines are aimed at minimizing the average error in the majority populations. This leads to a different distribution of errors in the sub-populations; the average error will be higher for the minority population than for the majority population.
- Formalizing fairness: The formalization of fairness for technical systems is problematic – on the one hand because a uniform definition of fairness is lacking, and on the other because contextual interpretations of fairness cannot be applied to an AI system, because a form of abstraction or formalization is typically adopted to achieve a general model.

Societal challenges

Finally, a number of broader societal difficulties have been identified in policy reports and scientific literature in regard to the risks of discrimination and inequality associated with AI:

- Effectiveness and reliability: The lack of effectiveness and the occurrence of errors can be partly explained by technical limitations and unavoidable bias. However, another explanation is the lack of standards for data analysis.
- Filter bubble and Matthew effect: Old patterns can significantly affect the possibilities and opportunities that groups have in the future. If no action is taken, existing social inequalities in society are replicated and deepened by AI.
- Legitimacy and trust: Internal audits and external inspections of AI are lacking. This undermines the trust of citizens in AI-driven decision-making processes. While contextualizing data and decisions is important for citizens' intuitive acceptance of AI, AI often leads to decontextualization of decision-making processes.

Chapter 2 – Existing Standards

Discrimination

Legal regulation of non-discrimination and equal treatment incorporates international, CoE, EU and national law. The legal framework consists of roughly four elements: (1) the assessment of whether a specific situation falls within the scope of the legislation, (2) the qualification of discriminatory treatment, (3) the assessment of whether exceptions or justifications apply, and (4) a proportionality test. This can summarized as follows.

1. Awareness	Do my aim, design or outcomes produce potentially "suspicious" distinctions? Suspicious grounds are, inter alia: • marital status • disability/chronic illness • sex/gender • religion • age • religious identity • nationality • political opinion • race/ethnicity • sexual orientation	Example 1: The algorithm I use to assess acceptance conditions gives lower scores to individuals who are or have been unfit for work Example 2: My recruitment algorithm is trained on successful résumés. All employees at my organization are male and 18 or older Example 3: I want to build an AI system that filters out people with two nationalities and selects them for additional screening
2. Discrimination?	Does this cause disadvantage?	Example 1: Persons with a disability/chronic illness are potentially excluded from the service I offer Example 2: The recruiter is potentially not presented with the résumés of women or persons under 18 Example 3: The group identified is subjected to a higher level of scrutiny
3. Can I justify my decision?	Do I have a good reason for the distinction that is made? (1) There is a legal justification for the distinction (e.g. positive discrimination may be allowed/required by law); or (2) There is an objective justification for the distinction. This means that the distinction is (a) Relevant: Appropriate for the selected aim (does it contribute to achieving it?), consistent (free from inherent contrarieties?) and coherent (does it take account of the context in which it is applied?) (b) Necessary: Necessary to achieve the goals. Also, there are no other, less drastic means available to achieve the same goal which are equally effectively. (c) Proportional: the means are proportionate to the goals pursued	Example 2: recruitment algorithm I am looking for candidates for a specific, high-risk job. Persons under 18 are legally unauthorized to practice this work. However, I am not permitted to exclude women from the recruitment process, and the algorithm may cause the systematic underscoring of female candidates. I must correct this.

Guidelines, CoC and practical tools

In recent years, national and international companies, governments and organizations have developed a large number of ethical guidelines for AI. Six different approaches and strategies can be differentiated to counter bias.

1. **Static approaches and software toolkits:** These include methods and techniques to detect or prevent bias in data sets, AI models or the outcomes models produce. A further differentiation can be made between:
   - ***pre-processing* aimed at data**: methods that focus primarily on creating balanced data sets and minimizing bias in the data set;
   - ***in-processing* aimed at algorithms**: methods that focus on adjusting the algorithm in such a way that, when training the model, it is explicitly instructed to minimize discriminatory effects;
   - ***post-processing* aimed at ML models**: methods that focus on the minimization of bias after training the classification model. This may be white box methods that adjust the model, or black box methods that adjust the model's predictions.

2. **Discursive frameworks, self-evaluation tools and learning material**: These methods are generally less technical in nature and serve to help developers – but also users, policy makers and other parties involved – recognize, prevent and mitigate bias. Impact assessments, questionnaires, evaluation cards and instructions for use are examples of methods that fall into this category.

3. **Documentation standards:** These methods are aimed at the standardization of descriptions of data sets and models. AI algorithms often use many different types of data or models. By documenting in a standardized manner how and why data sets are developed and which decisions are made in training the models, developers gain better insight in the data and models they work with, which makes them better equipped to detect and mitigate bias. Examples of these methods are data sheets, model cards and declarations of conformity. The underlying idea is that, for each data set, a detailed description is provided of how it was produced and what the strong and weak points of the data set are. Model cards are brief documents accompanying trained ML models, which provide a benchmarked evaluation under different circumstances, such as across different cultural, demographic or phenotypic groups (for example, race, geographic location, sex, and Fitzpatrick skin type) and across intersectional groups (for example, age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also provide a description of the context in which models should be used, details of the performance evaluation procedures and other relevant information. Declarations of conformity are (often not legally required) documents provided by suppliers to clarify how a product was produced, how it was tested, what the expected performance is, etcetera.

4. **Auditing**: These methods compare the outcomes of systems across different groups, based on different data sets and interactions, to examine whether the use of the selected algorithms leads to discrimination. Examples are surveys, A/B testing, non-invasive data scraping, and crowdsourced auditing in which users collect data by interacting with the system. These methods are generally applied after the system is developed, and provides little information about how bias could arise in the system. Auditing methods that can be differentiated include (1) institutional, (2) software-related and (3) hardware-related mechanisms. Relevant mechanisms for the development of the handbook are:
   - The use of audits by independent parties (1)
   - Bias and security bounties (1)
   - Explainability and documentation (2)
   - Compute support for university researchers (to be able to evaluate claims on largescale AI systems) (3)

5. **The development of technological standards and certification:** Several national and international institutions are currently working on the development of a broad spectrum of standards for AI, which generally also address discrimination, bias and fairness. Regarding certification, various initiatives are developing programs to clarify whether systems have been tested for bias and that measures have been taken to prevent bias.

6. **Socio-technical methods:** The methods described above generally put emphasize on the technical aspects of the issues that produce bias and discrimination. They pay little attention to the cultural, organizational and political context in which AI algorithms are developed and used. However, other strategies and methods have been developed that do focus on those aspects. These also consider the ways in which the development and use of AI algorithms is embedded in the broader context. For example, several initiatives stress the importance of having diversity in AI development teams, engaging and involving stakeholders in decisions, and paying attention to power relations and structures within the socio-technical ecosystem.

The six categories outlined above may overlap, and the different methods, approaches and strategies can complement each other. A documentation standard, for example, may also pose questions that stimulate reflection on the diversity of the team that built the data set.

Lastly, in addition to formulating principles and developing practical tools, anchoring them within the organization is crucial. After all, such initiatives are voluntary and can easily be dropped when circumstances change. Therefore, putting organizational mechanisms in place to safeguard these initiatives is of crucial importance to avoid the risk of “ethics washing.”

In short, although it is promising that so many strategies are being developed, the current multitude of strategies and the lack of effective enforcement mechanisms also make it difficult to choose the most fitting strategy. While some factors may indicate the success of these initiatives (such as involvement in law, specificity, reach, enforceability, iteration and follow-up), no extensive comparative research has been conducted to examine the effectiveness of one of these approaches. For the development of AI system principles, it is important to take these limitations and challenges into account.

Statistical principles

The collection of data and the use of insights and analyses for policy objectives and decision-making must adhere to legal requirements that are determined, among others, by privacy and data protection laws, the prohibition of discrimination, and the right to equal treatment. This applies to a lesser extent to the analysis of data, since no concrete decisions that affect citizens are made when data are analysed and such does not necessarily involve the processing of personal data (for example, if data are processed at aggregate level and general correlations are found). Still, here exist commonly accepted principles for statistics and statistical analysis.

Privacy and data protection

The GDPR provides a number of standards for AI systems that make use of personal data:

Legitimate	Fair	Purpose limitation
Governmental organizations must have the required legal basis and serve a public interest. Private organizations need consent or must serve an interest that is more important than the individual interest of the data subject.	The entire data process, from collecting and storing data to analysing data and using profiles for automated decision-making, must be lawful and fair.	Data collected for a specific purpose may, in principle, only be processed for that purpose. An exception to the purpose limitation principle is the processing of data for statistical purposes.
Data minimization	**Data quality**	**Transparency**
Only data that are necessary to achieve the specific objective of the AI may be collected; they must be deleted as soon as the objective has been met.	Data must be correct and up to date; citizens have the right to present additional data.	A citizen has the right to information on who processes which data and why and on what logics underlie automated decision-making.
Sensitive data	**Automated decision-making**	**Accountability**
Data on race, ethnicity, religious or philosophical beliefs, sexual orientation, and medical and criminal history may not be processed unless there is, inter alia, a significant public interest, or explicit consent.	Fully automated decision-making and profiling practices that have legal consequences or significantly affect individuals are prohibited, unless there is a legal basis or consent has been obtained.	Organizations have the obligation to keep a register of data processing activities, to conduct data protection impact assessments for high-risk projects, and to adopt data protection by design standards.

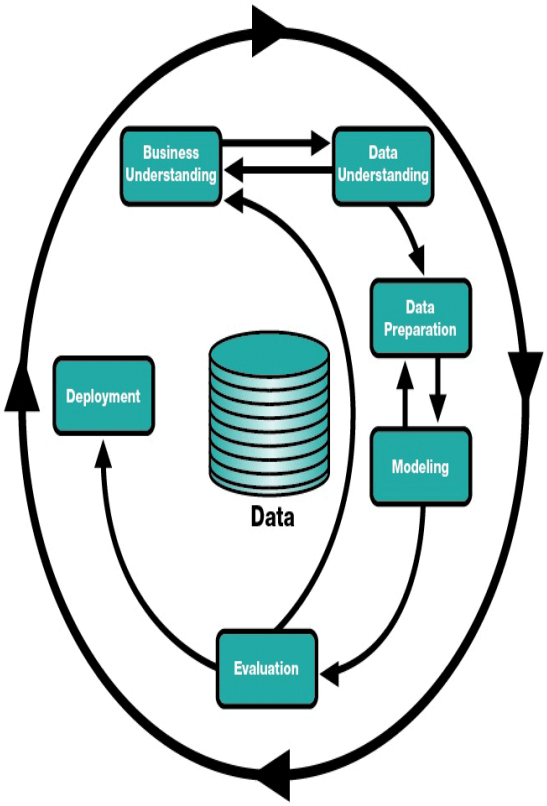
Procedural standards should be in place when governmental organisations make decisions that affect citizens' private sphere:

- Citizens must be informed when a decision is made that affects them;
- They must have access to all relevant information;
- The decision-making process must be neutral;
- Citizens have the right to be involved in the processes and to voice their views;
- Citizens have the right to contest decisions;
- Decisions must be made timely;
- Citizens have the right to legal support and representation;
- The decisions must be understandable and fair;
- The decision-making process must be fair and adequate.

Chapter 3 – Evaluation of existing standards and principles

Based on step 1 and step 2, a first overview of principles and standards is created. Three choices are of importance here.

Firstly, the problem is that anti-discrimination laws and equal treatment jurisprudence do not provide clear guidelines for process design and the ways in which a decision should ultimately be made. Once a decision is made, it can be examined *ex post* whether the decision was based directly or indirectly on one of the prohibited grounds or whether the decision disproportionally affects certain groups in society, and if so, whether there is an objective justification for it. The handbook largely covers the *ex ante* phase of a decision, during which it is ensured that an AI-driven decision-making process is as neutral and non-discriminatory as possible. Although some of the legal standards can be extrapolated to earlier phases in the decision-making process, this does not provide sufficient material to develop an exhaustive and complete handbook. Therefore, the handbook also considers principles from disciplines that traditionally focus on the process design of decision-making systems, such as AI literature, statistical principles and data protection law.



Secondly, the handbook differentiates between legal principles, technical principles and organizational principles. It must be noted, of course, that these different types of standards cannot always be clearly distinguished. For example, the General Data Protection Regulation contains many principles that are primarily organizational in nature; in the technical literature, a number of legal principles are included and elaborated; etc. In addition, there is significant overlap between the various principles. Nevertheless, it was decided to use this three-category division, in which organizational principles mainly deal with the process design (who is part of the team, how are decisions documented, etc.), technical principles deal with the design of the AI system as such (which fairness definition is selected, what bias is considered acceptable, etc.) and legal principles deal with the evaluation of the system (why are certain data collected, are the data truly necessary to achieve the purpose, etc.).

Thirdly, it was decided to structure the handbook in a variation on the visualization provided in the CRoss-Industry Standard Process for Data Mining (CRISP-DM). This model is well-known among AI experts and already structures the different phases of the process. The legal, organizational and technical principles will be clustered for each phase, in order to provide a step-by-step guide for building, developing and using an AI system. It must be kept in mind, however, that the development of an AI system is not a linear process– rather, it resembles the hopping procession of Echternach, in which three steps are taken forward, followed by two steps back. Specifically, adjustments to the final two stages of this model have been made in the handbook to create more space for permanent evaluation.

	Legal	Organizational	Technical
Phase 1 – Problem definition: identification of the problem/the objective and translating this into AI/ML task	**Necessity:** 1. Is it necessary to initiate an AI project? 2. Is it necessary and proportional to collect and process additional data for the project? 3. What is the expected effectiveness of the AI project? **DPIA:** 1. What is the impact on human rights? 2. How can perceived risks be mitigated? 3. If high risks continue to exist after taking additional measures, permission must be sought from the Data Protection Authority **Transparency:** 1. Document all choices that are made, and motivate them 2. Proactively provide citizens with information as much as possible 3. Respond to requests for information without delay	**Team:** 1. The project team is selected based on competences 2. The project team is diverse in terms of gender/cultural/religious backgrounds 3. The project team is diverse in terms of expertise/professional background **Mandate and resources:** 1. The team has access to the required resources 2. The team has the necessary authorizations **Involvement:** 1. Stakeholders or representatives of stakeholders are involved and heard throughout the process 2. Gain insight into the context of the problem and the stakeholders 3. Involve stakeholders in defining the problem and drawing up requirements	**System choices:** 1. Motivate the choice for an AI/ML system in relation to the problem/purpose 2. Motivate the choice for statistics in relation to the problem/purpose 3. Motivate the choice for the methodology, such as CRISP DM, in relation to the problem/purpose 4. Formulate the logics and the whys behind the AI system **Benchmarks:** 1. Formulate the target: when is the AI system successful? 2. Formulate an acceptable margin for false negatives and motivate that margin 3. Formulate an acceptable margin for false positives and motivate that margin **Ethics Canvas:** Conduct the Ethics Canvas technical evaluation, as developed by the Open Data Institute
Phase 2 – Initial data collection and storage	**Purpose limitation:** 1. Define a concrete objective 2. Only collect data that are necessary to achieve this purpose, and delete them as soon as the objective is met 3. Do not use the data for other purposes **Legitimate:** 1. Government organizations must have a legal foundation for an AI project; it must serve a public interest 2. Private organizations must have consent for the collection of data from all data subjects, unless the AI system serves an interest that is greater than those of the data subjects combined. **Secure and confidential:** 1. Ensure that unauthorized persons outside of the	**Data burden:** Data are collected equally among the different segments of the population; duplicates in data sets are avoided. **Objective compilation:** 1. Which data sources are selected, and are these representative? 2. Is there a historical bias in the data, and if so, which? 3. Which label categories are selected for the data, and why? **Quality inspections:** Organizations and teams continuously evaluate their work and procedures; if necessary, external experts are consulted	**Bias in, bias out:** Check for biases in the data set, especially if the data were obtained from third parties/public sources **Sampling method:** Choose a sampling method (e.g., random, stratified, oversampling) and motivate this choice **Documentation standards**: AI algorithms often use many different types of data or partly trained models. A standardized manner must be adopted to document data and choices that are made, for example by using data sheets and model cards.

	organization do not have access to the data 2. Ensure that unauthorized persons within the organization do not have access to the data 3. Ensure that, if an unauthorized person gains access to the data, the data are unusable/encrypted		
Phase 3 – Data-analysis and preparation	**Data quality:** 1. Check whether data are correct and up to date 2. Correct and update data if necessary 3. Inform citizens of their right to provide additional personal data **Sensitive data:** 1. Assess whether the data set includes data on race, ethnicity, religious conviction, sexual orientation, or medical and criminal history 2. Assess whether these data can be derived from the data set directly or indirectly 3. Determine whether it is possible to delete these data from the data set, without significant disadvantages **Sensitive data:** 1. If sensitive data are necessary, define their purpose 2. Ensure that there is a legitimate ground for processing those, such as explicit consent from data subjects or a significant public interest 3. If the data are stored solely for the purpose of preventing discriminatory effects, document this	**Objectivity:** Data sets, methods of analysis and decisions are selected with objectivity in mind; errors are documented and correct immediately **Quality:** Methods, procedures, definitions and classifications are used in a consistent manner; they are standardized across organizations as much as possible to allow for comparisons **Relevant expertise:** All necessary and relevant expertise is available to analyse the data and to prepare the data for modelling; employees continuously receive training to keep their knowledge up to date	**Check:** 1. Describe the composition of the data set 2. Examine the distributions in the data set 3. Check whether all relevant groups are represented **Pre-processing:** Data sets are compiled in a balanced way, for instance through: 1. Instance class modification 2. Instance selection 3. Instance weighting **Double check**: After the pre-processing methods, check whether the data set is balanced and representative; if not, repeat the various correction mechanisms
Phase 4 – Modelling	**Statistical principles:** 1. Trustworthiness: statistics must measure and represent reality as authentically, accurately and consistently as possible 2. Neutrality: statistics are developed, produced and distributed in a neutral manner	**Trustworthiness/comparability:** Methods used for data collection and analysis are defined, documented and made public. Accessibility and universal design are given priority so that everyone can use the products, including people	**In-processing:** The algorithm is adjusted to minimize biased outcomes, for instance through: 1. Classification model adaption; 2. Regularization/loss function and constraints; 3. Latent fair classes.

	3. Objectivity: statistics must be developed, produced and distributed in a systematic, reliable and unbiased manner; this implies the use of (context-dependent) professional and ethical standards 4. Comparability: the statistical concepts, measurement tools and procedures applied are compared, and harmonized to the extent possible, across geographical regions and societal domains 5. Consistency: the use of concepts, classifications an methods is consistent through time; deviations and adjustments are documented and explained	with a disability. Universal design principles should be applied to be able to serve as many users as possible **Accountability:** 1. Meta data are stored and documented; 2. Data are accessible to third parties for as far as possible; 3. The model must be explainable and understandable to the identified stakeholders **Participation of stakeholders:** Citizens, stakeholders and external experts are involved in the process of modelling	**Post-processing:** Bias is reduced after training the classification model. White box methods adjust the model; black box methods adjust the predictions. Examples of methods: 1. Confidence/probability score corrections; 2. Promoting demoting boundary decisions; 3. Wrapping a fair classifier on top of a black box base learner. **Causality:** If AI is based on causality, for instance because it is used for decision-making, deep learning is not an obvious choice. Motivate it.
Phase 5 – Evaluation (evaluate selected model based on success criteria formulated at step 1 and a "test set")	**Right to non-discrimination:** The organization that uses AI must demonstrate that the system does not directly or indirectly discriminate and that, if it does, this is legitimate and necessary. **Right to privacy:** If decisions affect citizens: 1. They must be informed about it; 2. They must be involved; 3. They must have the opportunity to contest the decision; 4. They must be offered the opportunity to receive legal counsel; 5. The decisions must be fair and understandable; 6. The decision-making process must be neutral **Data protection:** 1. Right to information, including information on the algorithm 2. Right to contest the decision 3. Right to contribute additional information 4. Right to not be subjected to automated decision-making or profiling	**Validation:** Statistical outcomes are validated by means of: 1. Prior testing 2. Reviewing 3. Monitoring 4. Editing 5. Designing **Improvements:** Errors in data or in models that have already been implemented in practice are adjusted as soon as possible and made public **Universal design:** Use an accessible and universal design so that everyone can use the products, including people with a disability	**Fairness:** Which definition of fairness is adopted (e.g., individual parity or group parity), and why? **Anti-classification:** A model is considered fair if it excludes protected characteristics when producing a classification or prediction. Some anti-classification approaches also attempt to identify and exclude proxies for protected characteristics. **Outcome / error parity:** Compare how members of the various protected groups are treated by the model. Following the fairness definition of outcome parity, a model is fair when the positive and negative outcomes that it produces are distributed equally across groups.

This overview was subsequently presented to three different test groups. Test group 1 consisted of lawyers, test groups 2 consisted of "techies," and test group 3 consisted of a mix of participants from different backgrounds. This way, expert feedback was obtained from legal and technical perspectives, while also obtaining an interdisciplinary perspective on these principles. The most important take away points from these test groups were:

1. **Limit ambitions:** It is virtually impossible to capture anti-discrimination jurisprudence in a clear model – not only because anti-discrimination law requires many different considerations and choices, but also because the legal factors that must be taken into account are complex and because the concretization of these factors and principles is context-dependent.

2. **Keep it soft/open:** A handbook cannot sum up what is and what is not permitted, because the law is not that straightforward. In addition, clearly defined principles have the disadvantage that they can people can work around them. The key is to give clear overview of the most important legal principles, so that they become instilled in people's minds: awareness.

3. **Organization:** Most problems and most solutions can be found in the organizational part of the overview.

4. **Diversity:** The team that builds and evaluates an AI system must be diverse, both in terms of their professional and personal backgrounds. It is important to also involve stakeholders here, preferably at all stages of the process.

5. **Intersectionality:** It is important to bring together as many disciplines as possible, also because, ultimately, everything is connected to everything.

6. **Domain knowledge:** AI system builders must always also have knowledge of the domain in which the system will be used.

7. **Iterative process:** A continuous iterative process must take place – in between one phase and the next, but also between the overarching principles and the practical application of them in concrete cases, and between the legal, organizational and technical principles, which are also interwoven.

8. **Documentation:** Documenting all questions and steps is pivotal, because it often concerns an iterative process.

9. **Continuous process:** AI systems continue to learn, which makes it important to continuously test and evaluate whether bias and discriminatory effects occur.

10. **Continuous updating:** A static handbook cannot work, because AI systems, anti-discrimination jurisprudence, as well as the interaction between those are constantly in flux.

Subsequently, a supplementary study was conducted of the jurisprudence of the European Court of Human Rights and the Court of Justice on anti-discrimination law to formulate additional principles for a handbook. In addition, a quick scan was conducted of jurisprudence in non-European jurisdictions on AI-systems, offering a variety of additional approaches.

Chapter 4 – Developing the handbook

Based on the overview of standards, the input of the three test groups and the additional research, a first concept of the handbook was developed. Although the three-category structure of legal, technical and organizational principles is maintained in this draft handbook, it was decided to begin each phase with a number of key questions. These key questions serve to guide the conversation within the organization, and to ensure that all relevant aspects are discussed. In addition, three fictional cases provide an illustration of what such a discussion may look like. The following questions are posed at the beginning of each phase:

Phase 1 – Problem definition	**Purpose and necessity** What is the problem and how will AI help solve it? What is the purpose of the project? Is the use of AI necessary, or could the problem also be addressed without using an AI system? Based on which assumptions about the various groups were the problem definition and the purpose of the system formulated? Are there different views about the purpose of the system, and have the various stakeholders been heard? What is the problem and according to whom and why must it be addressed? Which groups are differentiated in the problem definition(s) and why? What should the system change for whom, and why?	**Impact** Does this project require the collection of more data than currently available within the organization, and what consequences does this have for citizens? What impact does the system have on citizens and on society, both positive and negative? 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? What impact do false positives and false negatives have on citizens and society? What procedures have been taken up for stakeholders to contest a decision (false negatives/false positives)?	**Benchmarks** What are the financial, computational and organizational costs of this system, and what would the costs be of a non AI-driven alternative? When is the AI system considered a success (for example, at which effectiveness rate), and when must this benchmark be reached (for example, in 1 month or 2 years)? What percentage of false negatives and false positives is acceptable, and why? What do the various success criteria mean for different groups?
Phase 2 – Data collection	**Purpose and necessity** What data are needed for this project and why? To what extent are these data already available within the organization, and to what extent is externally collected data needed? Is it permitted to collect and process these data for this project?	**Data quality** What bias does the data set contain, and what are the consequences? Are the data representative and are all relevant groups represented equally? If multiple data sources are used, how is it ensured that these data are compatible and comparable?	**Data storage** How long will the data be stored and in which way? Will the data be treated safely and confidentially; what consequences does a data leak have for groups of categories of persons? 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?
Phase 3 – Data prep	**Inclusion and exclusion** Which of the collected data are relevant for the model and why?	**Integration and aggregation** How is it ensured that historical data and newly collected data fit together: are the data comparable,	**Labelling** How are data labelled and why?

aration	What happens with the data that are not used? Which criteria are used for data selection and how do they reflect distinctions made between groups? Does the selection of specific data or data operations influence the problem definition? Which aspects of the problem are not taken into consideration?	and what assumptions about groups and categories are inherent in the various data sources? How are the data aggregated, and what consequences does this have for the representativeness of the data? 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? Does combining different data lead to proxies, and if so, which?	Is this is line with the way other organizations label data and use datasets on which the algorithm has been trained? Is this in line with the way other stakeholders/citizens and domain experts would label data? Does the dataset contain sensitive labels, such as those referring to ethnicity, sexual orientation or gender, or labels that indirectly refer to these attributes. If so, why?
Phase 4 – Modelling	**Pre-modelling** Which algorithm is selected and why? What type of model will be built and why? How are criteria concerning explainability and fairness translated into a model selection strategy?	**Model (selection)** What parameters are chosen for the model and why? Does it suffice to build a single model, or would it be better to build multiple models and to compare them? Is the model based on existing models and why (not)?	**Test** How does the model perform on effectiveness? How does the model perform on the selected definition(s) of fairness? How does the model perform on the predetermined success criteria in terms of false positives and false negatives?
Phase 5 – Implementation	**Practical test** 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?	**Model adjustments** What adjustments are needed to improve functionality? What alterations are needed to increase the model's fairness? What alterations are needed to reduce the error rates?	**Application** What limitations arise from the previous steps with respect to the model's application possibilities and the implementation process? What are the key points of attention regarding application, and how can these be monitored in the implementation process? How will stakeholders and others be informed and involved?
Phase 6 – Evaluation	**Evaluation preparation** Will evaluation take place continuously, periodically or both? Will evaluations be conducted internally, externally or both? How will the evaluation be assessed, and based on which measurement points?	**Evaluation** How does the system perform with respect to the success criteria? Which improvements are needed with respect to the protected categories? How would the system perform if another model, fairness definition and/or algorithm would be adopted?	**Points of action** Should the system be (temporarily) put on hold? Can problems and obstacles be solved? How are the evaluation results perceived and interpreted by stakeholders and external experts?

Chapter 5 and 6 – Validation and finalizing standards

The first draft version of the handbook was discussed with a group of experts, representing various ministries, the Dutch Data Protection Authority, governmental organizations and semi-public organizations. In addition, the draft was presented in three workshops, in which the handbook was applied to existing use cases of AI. The most important points of attention that were identified during these steps were:

1. **Start:** In practice, a project is often initiated without having an elaborate plan in place, which makes it difficult to foresee how the project will develop. It is possible, however, to formulate milestones, benchmarks and objectives at the start of the project. This can be a broad vision of what the aim of the project is, whom and what purpose it serves, etc. It is important to also consider how non-discrimination will be embedded in the project's design at this stage.

2. **Ownership:** In principle, the individual citizen is the owner of her data. A key question, therefore, is whether the AI application benefits these citizens or the community at large. Ongoing involvement of citizens in the process is also key. In the health domain, dynamic consent is required.

3. **Necessity of storing sensitive data:** Sensitive data may be necessary to check the AI system for bias/discrimination. Therefore, these should not be deleted.

4. **Difference AI and human decision:** There is a difference in the error rate/bias that is accepted from human decision-making processes on the one hand and computer-driven decision-making processes on the other. Therefore, it can be advisable to conduct an analysis of the discriminatory effects of current (human) practices before the start of an AI project, and to examine the extent to which AI could actually improve/reduce existing bias.

5. **Need for brief summary:** Data analysts may not always choose to work through a long document before starting a project. Therefore, it might be beneficial to summarize the handbook and to provide an 1-page document listing the key points.

Based on the outcomes of the workshops and the input from the members of the expert group, an adjusted version of the handbook was developed. This new version was discussed with the expert group again. Fianlly, the authors presented the handbook to an internal auditor group, which included: Mark Bovens (Utrecht University; Scientific Council of Government Policy), Francien Dechesne (Leiden University), Ronald Leenes (Tilburg University), Egge van der Poel (TIAS), Johan Wolswinkel (Tilburg University) and The Institute for Human Rights. Based on this, the team developed the final version of the handbook.

Colophon

Text and research by: Bart van der Sloot (Tilburg University), Esther Keymolen (Tilburg University), Merel Noorman (Tilburg University), Mykola Pechenizkiy (Eindhoven University of Technology), Hilde Weerts (Eindhoven University of Technology), Yvette Wagensveld (Tilburg University), Bram Visser (Vrije Universiteit Brussel) and in collaboration with The Netherlands Institute for Human Rights.

Commissioned by: The Dutch Ministry of Internal Affairs