Ministry of the Interior and
Kingdom Relations

March 2022

Impact Assessment
Fundamental rights and algorithms

Introduction – applying the FRAIA

This fundamental rights and algorithm impact assessment ('FRAIA') is a discussion and decision-making tool for government organisations. The tool facilitates an interdisciplinary dialogue by those responsible for the development and/or use of an algorithmic system. The commissioning client is primarily responsible for the (delegated) implementation of the FRAIA.

The FRAIA comprises a large number of questions about the topics that need to be discussed and to which an answer must be formulated in any instance where a government organisation considers developing, delegating the development of, buying, adjusting and/or using an algorithm (hereinafter for the sake of brevity: the use of or using an algorithm). Even when an algorithm is already being used, the FRAIA may serve as a tool for reflection. The discussion about the various questions should take place in a multidisciplinary team consisting of people with a wide range of specialisations and backgrounds. Per question, the FRAIA indicates who should be involved in the discussion. This tool pays attention to all roles within a multidisciplinary team, which are included in the diagram below. However, the list is not exhaustive. Likewise, the role or function names may differ from one organisation to another.

Role	FRAIA Part 1	FRAIA Part 2	FRAIA Part 3	FRAIA Part 4
Interest group	●			
Management	●			
Citizen panel	●			
CISO or CIO	●	●		
Communications specialist		●	●	
Data scientist		●	●	
Data controller or data source owner		●		
Domain expert (employee who has domain knowledge regarding the algorithm's scope of application)	●	●	●	●
Data protection officer		●		
HR staff member			●	
Legal advisor	●	●	●	●
Algorithm developer		●		
Commissioning client	●	●	●	
Other project team members	●			
Project leader	●	●	●	●
Strategic ethics consultant		●	●	

Introduction – applying the FRAIA

The discussion on the basis of the FRAIA aims to ensure that all relevant focus areas regarding the use of algorithms are addressed at an early stage and in a structured manner. This prevents the premature use of an algorithm that has not been properly assessed in terms of the consequences, entailing risks such as inaccuracy, ineffectiveness, or violation of fundamental rights. To achieve this aim, it is important to follow all the relevant steps involved in using an algorithm, and to thoroughly think through the possible consequences, whether any mitigating measures may be taken, et cetera.

For each question, the answers and the key considerations and choices made are meant to be noted down. Thus, the completed FRAIA may serve as reference material and can be used to account for the decision-making process surrounding the development and implementation of an algorithm.

To facilitate this exercise, the FRAIA works on the assumption that the decision-making process regarding algorithms can be divided into three main stages:

- **Stage 1**: preparation. This stage is about deciding why an algorithm will be used and what its effects will be.

- **Stage 2:** input and throughput. This stage is about the development of an algorithmic system. In this stage, it is decided what the algorithm must look like, and which data is being used to feed the algorithm. Within this stage, the FRAIA further distinguishes between:
  - **Stage 2a:** data, or input. This involves asking questions that pivot on the use of specific data and data sources;
  - **Stage 2b:** algorithm, or throughput. This involves questions regarding the algorithm, and its operation and transparency.

- **Stage 3:** output, implementation and supervision. This stage is about how to use the algorithm, i.e., about the question which output the algorithm generates, how that may play a role in policy or decision-making, and how that can be supervised.

Introduction – Application of the FRAIA

In all stages, respect for fundamental rights must be ensured. Therefore, the FRAIA includes a special sub-section that pays attention to identifying risks of infringing fundamental rights and to the need to provide a justification for doing so.

Thus, the FRAIA is composed of four parts: three parts cover the three decision-making stages regarding the use of an algorithm, while the last part centres on the broader questions about fundamental rights. Per part, the FRAIA exists of a number of sub-themes, about which questions have been formulated that need to be discussed and answered within the team. Per sub-theme the questions are explained in smaller, dark blue print.

The FRAIA is closely related to a wide range of other rules, guidelines, reference or assessment frameworks and impact assessments, including the well-known data protection impact assessment (DPIA). Consequently, in the various parts, the FRAIA refers to a further elaboration or refinement, greater depth, or alternative terminology of classification provided by such tools. First, an introduction has been included for each part of the FRAIA listing the relevant tools. Second, in case of more specific questions – if relevant –, reference is made to a further elaboration or detailing in one or more of these tools as part of the explanation. References are printed in light blue print and contain links to the tools to facilitate clicking through to them during the discussion.

As a result of this approach, the FRAIA can function as an overarching tool in which the other tools have been logically embedded. Answers already entered into the FRAIA may be incorporated into other tools, such as the DPIA, and vice versa.

In some cases, the outcome of a discussion about a question in the FRAIA will be that an algorithm may be problematic or may entail serious risks. The FRAIA points out such cases in red print.

FRAIA flow chart

START

1 Why?

Intended effects (objective)

Reason
Objectives, values, ...

→ See Chapter 1

2 What?

A) Data (input)

Preconditions
Data quality, storage, archiving...

B) Algorithm (throughput)

Preconditions
Accuracy, transparency, explainability

→ See Chapter 2

3 How?

Implementation and use of algorithm (output)

Preconditions
Impact, evaluation, communication ...

→ See Chapter 3

4 Fundamental rights

Infringed fundamental rights
Specific legislation
Seriousness of interference
Justification

→ See Chapter 4

Apply if necessary
- Mitigating measures (Annex 2)
- Alternatives

Then go through process once more

FINISH

Once all questions have been answered satisfactorily, and if the fundamental rights assessment in Part 4 comes out positive, the FRAIA has been completed successfully.

Part 1: why?
Intended effects – objectives – preconditions

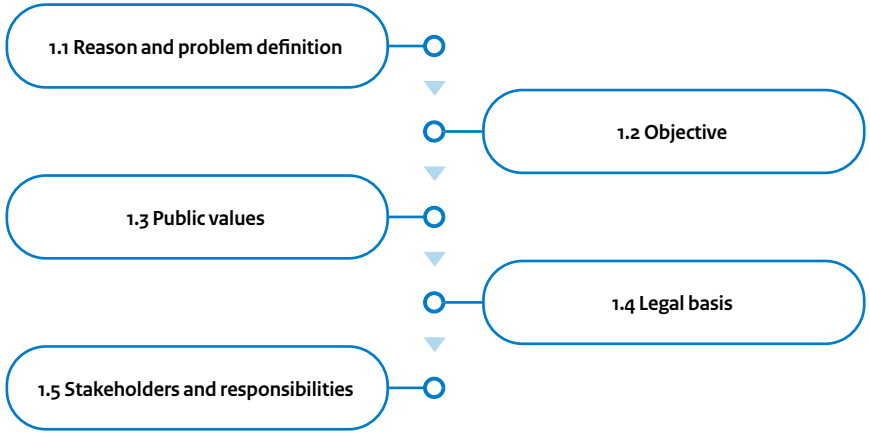
Introduction

Part 1 of the FRAIA is about the 'Why' of the intention to develop, purchase, adjust and/or use of an algorithm (henceforth for short: the use of an algorithm). What are the reasons, the underlying motives, and the intended effects of the use of the algorithm? What are the underlying values that steer the algorithm's deployment? These overarching questions must first be discussed in a decision-making process about the use of algorithms, before getting round to questions about, for instance, preconditions or possible impact on fundamental rights. The replies given to the questions in this part are relevant to answering the more specific questions in the next parts.

Instruction

Before discussing and answering the question, first read the explanation that follows after the question. For the various questions, it is always briefly specified which team members should be present at this discussion.

This section covers the following topics:



Enter here which persons complete this section. Also state their function.

Part 1: why?
Intended effects – objectives – preconditions

Part 1 builds upon/is related to:

- Integrated impact assessment framework for policy and legislation
  The integrated impact assessment framework for policy and legislation is a Dutch tool, a simplified version of which is also available in English. It includes a large number of relevant guidelines concerning, for example, the choice of policy tools, identifying policy objectives as well as other interests and values, and determining the legitimacy of policy and regulations. This first part of the FRAIA is optimally geared to this framework by providing a more specific translation of its principles into the context of using an algorithm for decision-making or policy. For most of the specific questions in this first part that require further clarification, the Integrated impact assessment framework may be consulted.

- Algorithm assessment framework of the Netherlands Court of Audit (2021)
  This assessment framework contains leads for audits of deployed algorithms. The risks described in the framework, for example with respect to formulating algorithm objectives, or its legal basis, may be helpful during discussions of the subsequent questions in this FRAIA. Among other things, they clarify the manner and specificity of formulating certain objectives, and which choices must be substantiated. This framework is only available in Dutch.

- Ethics Guidelines for Trustworthy Artificial Intelligence
  These guidelines, developed within the EU framework, set out to safeguard that algorithms be legal, ethical, and robust, outlining a framework to that effect. Principles such as autonomy, damage prevention, justice, and accountability play a pivotal role in the discussion about the questions formulated in this first part of the FRAIA. Thus, these guidelines can be incorporated in the discussion, also as regards answering questions about the objectives of using an algorithm and the other questions in this first part of the FRAIA.

- Non-discrimination by design guideline (Part 1: problem definition)
  This section of the guideline presents topics that, while similar to topics included in this part of the FRAIA, focus specifically on tackling unjustified discriminatory treatment in the data. Thus, the guideline may be used for further information regarding the questions formulated below, which are more generally focussed on the development and use of algorithms. This framework is only available in Dutch.

- Good Digital Governance Code

  This code pays attention to a wide range of rule-of-law and duty-of-care guarantees that are involved in algorithmic decision-making procedures. These safeguards must likewise be observed in decision-making processes regarding the use of an algorithm. Therefore, the below-mentioned questions must be answered against the background of this Code. The Code is only available in Dutch.

- Data Protection Impact Assessment

  In this first part of the FRAIA, questions are formulated about, among others, the objectives of using an algorithm. Such questions must also be answered when a Data Protection Impact Assessment (DPIA) needs to be performed. A DPIA will be required for the use of various algorithmic systems, particularly if this use implies processing of personal data. Among other things, the performance of a DPIA requires identifying the objectives of data processing. These objectives may also be relevant for the choices with regard to the use of the algorithm as such. For that reason, it may be useful to involve the analysis already made for a DPIA in the discussion of the FRAIA questions. Conversely, the answers provided in this part of the FRAIA may prove useful while performing a DPIA. The important thing is that a DPIA does not always need to be performed.

  Moreover, a DPIA relates to a narrower assessment than the FRAIA does. After all, a DPIA is primarily about processing personal data, while the decision-making about the use of an algorithm may involve many more elements. Therefore, a DPIA cannot replace the review of the FRAIA. A Dutch-language checklist for the question when a DPIA must be performed and a number of guidelines for performing a DPIA can be found on the website of the Dutch Data Protection Authority website of the Dutch Data Protection Authority. Information in English can be found in the EU's Guide to GDPR Compliance.

1.1
Reason and problem definition

1.1.1 Explain your proposal for the use/deployment of an algorithm. For which problem is the algorithm to provide a solution? What is the actual occasion or reason to use an algorithm? Why does it require an algorithm?

Briefly describe the proposal here, in such a way that an outsider could understand what you intend to do.

Please note that the purpose of the algorithm is addressed separately, under question 1.2.

1.1 Reason and problem definition

Expertise/role required for answering this question: commissioning client, project leader, domain expert, possibly citizen panel, possibly interest group representative

Directions and explanation

The first question of this theme is about reflecting on the actual **reason** to consider deploying an algorithm: what is the problem that the intended use of the algorithm should solve? Thus, it is about **problem definition and delineation**, rendering it essential to present the problem as explicitly and precisely as possible. Sometimes, the problem or reason may be an internal matter; for example, internal processes are not running efficiently or could be made more efficient by means of an algorithm. In other cases, an algorithm may be used to solve a social problem or a problem occurring among a specific population group.

The main goal of the second question within this theme is to decide **why it is desirable or requisite to use an algorithm**, knowing that other (non-digital) tools may be available to tackle a problem. From that perspective, a discussion must take place about the question why an algorithm may provide a better solution than a non-automated or non-digital process.

In order to achieve proper civic participation, gain an early insight into various points of view, and build support for the use of an algorithm, it may be useful to involve citizens in defining the objectives. This may be achieved, for example, by putting questions to a **citizen panel** about the objectives of the algorithm. Alternatively, it may be useful to involve a representative of an **interest group** or civil society organisation. These have vast substantive expertise and have been set up expressly to represent stakeholder groups.

1.2 Objective

1.2.1 What is the objective that the use of the algorithm needs to achieve?

What is the main objective here, and what are secondary objectives?

1.2 Objective

Expertise/role required for answering this question: commissioning client, project leader, possibly citizen panel, possibly interest group representative

Directions and explanation

It is important to make the objective as explicit as possible since you must be able to measure the performance of the algorithm against the objective at a later stage. Developing an algorithm on the basis of good intentions alone is not enough to safeguard absence of undue effects. Moreover, the intended objective must be legitimate.

Building on the problem definition at 1.1, it is crucial to define **the purpose of using an algorithm as concretely and specifically as possible**. Thus, not (only): 'protection of national security', but (also) 'using an automated process to chart and analyse indicators for terrorism risks at specific locations'. Efficiency or cost-cutting objectives can likewise be identified, for example, 'automated text analysis to significantly decrease employee workload and to facilitate reducing staff by 4 FTE in the new year'. Where possible, it can be helpful to quantify the objectives.

Moreover, it is important to **rank objectives if there are more than one** (as is nearly always the case): what are the main objectives and why? Which objectives are 'secondary objectives' for which it matters less if they cannot be achieved in full?

If a DPIA is performed as well – for example, because you expect personal data to be used during the development or use of the algorithm – the objectives of data processing need to be defined. These objectives may (but need not) overlap with the objectives of algorithm use. Therefore, it may be useful to store the objectives you have defined for the FRAIA and re-use them for a DPIA, or vice versa.

To achieve sound civic participation and create support for the use of an algorithm, it may be useful to involve citizens in defining the objectives. One way of doing so is putting questions about the objectives of the algorithm to a **citizen panel**. An alternative option is involving a representative of an **interest group** or civil society organisation. These have vast substantive expertise and have been set up to represent stakeholder groups.

1.3 Public values

1.3.1 What are the public values that prompt the use of an algorithm? If there are several public values prompting the use of an algorithm, can they be ranked?

1.3.2 What are the public values that may suffer as a result of using an algorithm?

1.3 Public values

Expertise/role required for answering this question: commissioning client, management, legal advisor, possibly citizen panel, possibly interest group representative

Directions and explanation

In short, the values that prompt the development and use of the algorithm in question, **are to be made explicit**. Rendering explicit the values that must be expressed in the algorithm may help facilitate the assessment of the algorithm's effects at a later stage of the process.

It is not always easy to know what the public values are, but they invariably concern manifestations of the public interest. The specific **public values** prompting government action may differ from one situation to another.[1] Public values tend to be distilled from essential **rule of law and democratic principles**, from **conditions that allow a society to function properly**, and from individual and collective fundamental rights and freedoms.

Examples of public values are equality, respect for personal autonomy, solidarity, liberty, security, responsibility, sustainability, legal security, distributive justice, respect for vulnerable groups, participation, and efficient use of resources. On a more hands-on level, protection of fundamental rights (such as personal data protection, freedom of expression, the right of access to information, and the right to due process) may rank among the public values.

Algorithms may serve to translate certain public values into tangible decision-making. In the process, algorithms may reinforce certain values, but they may also harm public values – such as fundamental rights. For that reason, it is important to chart which public values may be involved in the use of the algorithm.

When answering this question, you may consider the **reason** for using the algorithm (see question 1.1) and the **objectives** for doing so (see question 1.2), but also the core values of the organisation or organisational unit that will use the algorithm. Moreover, attention must be paid to the **fundamental rights roadmap** (see Part 4 of this FRAIA) and to the overview of fundamental rights clusters in Annex 1.

After all, if fundamental rights may be affected by the algorithm, or if, conversely, fundamental rights may be fostered by it, this means that fundamental rights are involved as public values.
When **multiple public values** prompt the use of the algorithm, it is useful to reflect on the **relative weight** of these values. Which values are most important? What are the harms if the values cannot be (fully) realised by the algorithm?

1.3 Public values

If **public values conflict** as a result of using the algorithm, a **comparative assessment** must be made between them and a **balancing exercise** may need to be conducted. As it may be difficult to do so at this stage of the FRAIA, it makes sense at this moment to do no more than point out the tension. The objective of identifying the various values concerned at this point in time is mainly focussed on consciousness-raising and being aware. While going through the questions in the follow-up sections of this FRAIA, the interests and public values involved, and any tension between them, may be kept in mind. At least for steps 5, 6 and 7 of the fundamental rights roadmap (see Part 4 of this FRAIA), the desirability of deploying the algorithm will eventually need to be considered in light of the values and fundamental rights that are affected by it. Also, a choice may need to be made if fundamental rights clashes cannot be adequately avoided, remedied or mitigated.

To achieve good civic participation and create support for the use of an algorithm it may be useful to involve citizens in the discussion about the public values concerned. You may do so, for example, by presenting a citizen panel with questions about the expected positive and negative impact of the algorithm and the comparative assessments to be made between them. An alternative option is involving a representative of an interest group. Interest groups often have vast substantive expertise and have been set up to represent stakeholder groups.

1.4 Legal basis

1.4.1 What is the legal basis of the use of the algorithm and of the targeted decisions that will be made on the basis of the algorithm?

Answering this question involves finding out whether there is a legal basis that explicitly and clearly allows for the use of an algorithm and renders this use sufficiently foreseeable. If an algorithm is expected to affect people's lives or freedom, and particularly if fundamental rights are expected to be affected, there must be a legal basis for its use.

If a legal basis that meets the quality requirements is lacking in such a case, the algorithm cannot be used.

Please note that requirements regarding good administration, transparency, and legal protection – which are related to the legality requirement – will be discussed more fully in later sections of the FRAIA.

1.4 Legal basis

Expertise/role required for answering this question: legal advisor, commissioning client

Directions and explanation

In the case of **fundamental rights infringements** (see Part 4 of this FRAIA) a number of specific requirements apply on the basis of the European Convention on Human Rights, two of which are particularly important for the use of algorithms: [2]

- The infringements on fundamental rights must be **sufficiently foreseeable**. This means that the legal basis must be defined sufficiently clearly for citizens and legal entities to know where they stand and, where necessary, can gear their actions to the expected outcomes.

- The legal basis must provide appropriate **guarantees against arbitrariness**. This means, among other things, that there must be sufficient transparency about the decisions to be made, and that sufficient opportunities for legal protection must be in place.

Further specification of the requirements for the legal use of algorithms may be found in the EU Ethics Guidelines for Trustworthy Artificial Intelligence.[3] Likewise, the Integrated impact assessment framework for policy and legislation provides leads for determining the legality of intended regulations and policy, which also apply where the use of algorithms is concerned. The discussion about the legal basis must take place in the light of the explanations provided in these tools.

1.5 Stakeholders and responsibilities

1.5.1 Which parties and persons are involved in the development/use/maintenance of the algorithm?

In answering this question, you should also take note of the following:

- If multiple parties are involved in the development/use/maintenance of the algorithm: can the parties and their roles be rendered explicit?
- If it should become clear in future that the algorithm is no longer desired/feasible/relevant, is there an exit strategy? What might such an exit strategy look like?

If it proves impossible to sufficiently safeguard the responsibilities of the stakeholders, the algorithm must not be used.

1.5.2 Have the responsibilities with respect to the development and the use of the algorithm been transparently allocated? How has it been safeguarded that these responsibilities continue to be transparently allocated once the development of the algorithm has been completed and it is being put to use?

1.5.3 Who is ultimately responsible for the algorithm?

1.5 Stakeholders and responsibilities

Expertise/role required for answering this question: commissioning client, project leader, other project team members, CISO or CIO

Directions and explanation

It may be tempting to look at the use of the algorithm primarily from a policy or administrative perspective. However, there may be a discrepancy between administrative decision-making on algorithms and the political responsibility for and control over it. Also, there is a risk that a citizen's perspective disappears from view where decision-making on the use of algorithms is concerned, partly because participation by citizens in the discussion about algorithm use is often limited.[4] That is why it is important to imagine during the discussion whether, apart from **administratively responsible parties, politically responsible parties and citizens** could or should play a role in decision-making about the use of the algorithm and, if so, how that role should be given shape. In particular, where possible and relevant, enabling citizens to think along about the development of an algorithm in the follow-up phase (i.e., stages 2 and 3 and the fundamental rights check of Part 4) is desirable.[5] Even in the early stages of developing an algorithm, the questions whether or not to apply civic participation and how to shape this participation may be considered.

In any case – and even if no civic participation is provided – thought should be given to the way the algorithm that is to be used may affect citizens and to **sound ways to protect the position of citizens**. In doing so, attention must be paid to the importance of a tailor-made approach and to citizens' needs on the one hand, while equal treatment and consistency must be guaranteed on the other and arbitrariness must be ruled out. Decision-making must be organised in such a way that it is mindful of these risks and tensions. How opportunities can be created for flexibility, deviation from the norm, et cetera should also be considered.[6]

1.5 Stakeholders and responsibilities

A **clear division and safeguarding of (various levels of) responsibilities** is crucial to running the process surrounding the use of an algorithm effectively and responsibly. Especially if undesired effects (may) occur during the implementation of an algorithm, short lines of communication and a clear division of tasks are important to enable timely adjustments or mitigating measures (see Annex 2 for an overview of measures). This must be considered beforehand, meaning that **sound decision-making structures and organisational embedding** of the algorithm need to be put in place.

This clear division of responsibilities must also continue to be **safeguarded in the future**. For example, it is not inconceivable that certain colleagues change jobs, or fail to continue their position for other reasons. Likewise, (units within) the organisation may be reorganised. Therefore, the discussion about these topics pivots on the question of how the organisation will continue to ensure proper allocation of responsibilities and relevant contextual knowledge.

If the development or implementation of the algorithm proves to be no longer desirable, it must be possible to adopt an **exit strategy**. This is important, as sometimes an organisation may already be 'too far into' the project and it may seem as if there is no longer any way back. An exit strategy will prevent any harmful consequences, such as wasting public funds or using an inaccurate algorithm. Such an exit strategy does not need to be fully developed beforehand; the broad outlines may suffice.

In this respect, it is also important to think about the pitfalls that may occur if people make decisions with the help of algorithms; it is important to consider early on how such **pitfalls** may be avoided. This is true, in particular, for pitfalls regarding cognitive prejudice, such as *automation bias*: people are inclined to see decisions generated by a computer as neutral and tend to believe them more easily. Another example of cognitive prejudice is *anchoring*: people are inclined to use the bit of information provided first as an anchor, or benchmark, and are less inclined to deviate from it if contradictory or additional information is provided at a later stage.

Part 2A: What?
Data – input

Introduction

Once it has been established why an algorithm will be used and how the organisation will safeguard public values and interests, it is important to think about the shape of the algorithm that is to be used. This is what Part 2 of this FRAIA is about: the 'What?' of the project.

This part is divided into two sub-parts: Part 2A concerns the *input* of the algorithm: the data (or digitally recorded details) to be used and the corresponding preconditions. Part 2B concerns the algorithm itself, i.e., the throughput of the project.

Like Part 1, Part 2 comprises a number of questions and points of interest that need to be included in any decision-making process about the use of an algorithm. For the work method, see the introduction to the FRAIA.

This section covers the following topics:

- 2A.1 Assessment: Algorithm type
- 2A.2 Data sources and quality
- 2A.3 Bias/assumptions in the data
- 2A.4 Security and archiving

Enter here which persons complete this section. Also state their function.

Part 2A: What?
Data – input

Parts 2A and 2B build on/are related to:

- Non-discrimination by design guideline (parts 2: data collection; 3: data preparation; 4: modelling)
  This guideline presents topics that, while similar to topics presented in this part of the FRAIA, focus specifically on tackling bias or discriminatory treatment in the data. Thus, the guideline may be used for further study and information regarding the questions defined below, in particular where checking data for bias is concerned. This guideline is only available in Dutch.

- Guidelines for Algorithm Application by Governments and public education on data analyses (Dutch Ministry of Justice and Security, 2021)
  These guidelines pivot on the themes of transparency, explainability, validation, responsibility, and verifiability. These themes also come up in the questions in this part of the FRAIA. Thus, these guidelines are important to the discussion about questions relating to these topics, as they provide a further explanation about the topics in question. These guidelines are only available in Dutch.

- Ethics Guidelines for Trustworthy Artificial Intelligence
  Point of departure of the ethics guidelines formulated by the EU is that 'trustworthy AI' must be provided by means of 1) human control and human supervision, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination, and fairness, 6) environmental and societal well-being, and 7) accountability. Furthermore, the guidelines further pay ample attention to commitment/ participation, controllability, and information provision. These various elements return in various places and ways in the FRAIA, including in the present part of it; element 5 is mainly discussed in Part 4 of the FRAIA (fundamental rights roadmap) while element 6 has received attention in Part 1 (identifying public values). The guidelines can offer further explanations regarding the various questions and may be helpful if doubts should arise. Furthermore, the 'checklist' included in the guidelines can help to finetune the questions included below.

- An audit framework for algorithms (in: Understanding algorithms) (Netherlands Court of Audit, 2020)
  This audit framework formulates five perspectives that need to be included in the decision-making about algorithms, one of which (ethics) encompasses the other four (governance and accountability, model and data, privacy, and IT general controls). The definitions components presented in this audit framework can be considered when answering the relevant questions in this part of the FRAIA, specifically where security is concerned (part 2A) and as regards accountability (Part 2B). The privacy element from the audit framework of the Netherlands Court of Audit is more specifically discussed in Part 4 (fundamental rights roadmap). It is only available in Dutch.

Part 2A: What?
Data – input

- Algorithm assessment framework Netherlands Court of Audit (2021)
  This assessment framework presents a wide range of standards for and potential risks inherent to the use of algorithms for public administration. While developing the FRAIA, this framework has been consulted to the fullest extent possible and will be referred to where relevant below. However, it is important to look up the exact requirements when accounting for the choices that have been made. This framework is only available in Dutch.

- Data Protection Impact Assessment (DPIA)
  In many cases, developing and using an algorithm will involve the collection and processing of personal data. In many of those cases, DPIA will need to be carried out. Since this involves fine-tuning the fundamental right to protection of personal data, we pay attention to it in part 4 of the FRAIA – the fundamental rights roadmap (step 2). Moreover, Part 1 of the FRAIA has already paid attention to the DPIA with respect to defining objectives for the processing of personal data. The information and answers generated during the performance of a DPIA may also be relevant when discussing the FRAIA questions below; conversely, the answers to the FRAIA questions may be useful to the performance of a DPIA.

- Government information security baseline (BIO)
  The BIO is a joint information security assessment framework that encompasses all layers of Dutch government. It is important to observe the BIO when answering the questions in Part 2A. The BIO is only available in Dutch.

- The FAIR and FACT-principles
  The FAIR and FACT principles uphold that the use of Big Data will only contribute to a better society for all if it has been developed on the basis of important public values. The FAIR principles relate to findability, accessibility, interoperability, and reusability. The FACT principles relate to fairness, accurateness, confidentiality, and transparency.

2A.1:
Assessment: Algorithm type

2A.1.1 Is it already (roughly) known what type of algorithm will be used?

Expertise/role required for answering this question: data scientist, algorithm developer, project leader

The next part of the FRAIA will feature an extensive discussion about the choice of a specific type of algorithm (Part 2B: What? | 'Algorithm – throughput'). Nevertheless, this question is presented here as well, for if there is no rough indication of what type of algorithm is to be used yet, it is hard to answer the questions in this part of the FRAIA. When answering these questions, it may be useful to have a preview at the variants mentioned in relation to question 2B.1 below.

Depending on the answer to the question whether there is a rough idea of the algorithm that is to be used, there are various options available for discussion:

- If there is a rough (or, alternatively, a clearer) idea regarding the algorithm that is to be used, the questions below can be discussed with regard to that algorithm.
- If various algorithm types can be considered, the questions may apply to the various types.
- If, as yet, you have no idea what algorithm type will be used, you may cease to use the FRAIA at this point. Once you have a rough indication of the algorithm type, or if there are several options, you may resume using the FRAIA.

2A.2
Data sources and quality

2A.2.1 What type of data is going to be used as input for the algorithm and from which sources has the data been taken? If no input data is used, proceed to topic 2A.4.

2A.2.2 Is the quality and reliability of the data sufficient for the intended data application? Please explain.

2A.2
Data sources and quality

Expertise/role required for answering this question: data scientist, algorithm developer, data controller/data source owner

Directions and explanation

The '**garbage in = garbage out**' principle is widely known. This expression reflects that the question which data is used as input for the algorithm, and its quality, is decisive for the algorithmic outcomes. Therefore, it is important that the data is complete and accurate. Data quality can be monitored in a variety of ways, for example using samples. The central question is: does the data describe the phenomenon that is to be investigated? The data that is collected has to be the right proxy for what needs to be identified.

If the algorithm in question uses **training data**, its origins and quality need to be investigated. In that case, it is likewise important to check whether scientists have criticised or even revoked the training data set.

The FAIR and FACT principles elaborate on these elements.

2A.3
Bias/assumptions in the data

2A.3.1 What assumptions and biases are embedded in the data? How are their influences on the algorithm's output corrected or otherwise overcome or mitigated (see also annex 2)?

2A.3.2 If training data is used: is the data representative for the context in which the algorithm will be used?

2A.3
Bias/assumptions in the data

Expertise/role required for answering this question: data scientist, algorithm developer, data controller

Directions and explanation

Bias (or 'prejudice') is a broad concept; **bias in data can take various shapes**. During the discussion, it is important to consider these various shapes and to assess how you need to deal with them.

One example of bias is representation bias. It occurs when, for example, certain population groups are under- or over-represented in training data. A well-known instance is facial recognition software that has been trained by means of photos of predominantly white and male persons. As a result, this training data reflects a strong bias, with the result that the algorithm is poorly trained in recognising dark and female persons and made many mistakes that had far-reaching consequences.[7] Therefore, it is imperative to prevent this type of bias in data.

When training data are being used, other factors play a role in preventing bias. For example, while training data may be of good quality, it may still **not be suited to the specific algorithm and the intended objective of the algorithm**. In order to be able to assess whether the training data is suitable, the objective of the algorithm in question and the nature of the training data need to be considered. The question is whether the training data is actually able to train the algorithm in such a way that it can fulfil the objective. In this respect, it may help to briefly investigate prior use of the training data set by scientists.

An example that shows why it is important to align training data and the objective of the algorithm concerns an algorithm that had learned to differentiate between huskies and wolves. The algorithm had been trained using balanced data, half of which consisted of pictures of huskies, while the other half consisted of pictures of wolves. However, the neural network had learnt to recognise the difference by observing that husky pictures tended to have a green environment (for example because their pictures are taken on the grass) while wolves tended to be photographed in a white environment (snow). Thus, there was still a bias, but it had to do with the background of the pictures. As a result, the algorithm did learn to recognise differences in the pictures, but they were not differences that were relevant from the perspective of the objective (differentiating between huskies and wolves). That means that, particularly with machine learning, it is crucial to shape the data set very precisely in order to be sufficiently specific in defining what the machine learns.

Moreover, it is important that the **dataset is complete and accurate**. If an algorithmic rule is formulated to grant people benefits, while not all persons relevant to the rule's objective have been included in a data set, the application of the algorithm may still result in bias.

2A.4 Security

2A.4.1 Is the data sufficiently secure? Differentiate between input data and output data.

When answering the questions, take the following into account:

- Is the data sufficiently secure from any attacks?
- Have adequate measures been taken to organise access for the (groups of) people who are authorised to have access?
- Is logging taking place to monitor access to and use of data? In order to prevent and/or trace any threats from within the organisation.
- Have sufficient measures been taken to protect the identity of the data, such anonymising or pseudonymising personal data?

2A.4 Security

2A.4.2 Is access to the data supervised?

2A.4.3 Is there compliance with relevant rules on archiving?

2A.4 Security

Expertise/role required for answering this question:
Data protection officer, CISO, data controller

Directions and explanation

When discussing these questions, it is useful to observe the Algorithm Audit Framework, as it presents extensive control measures taken by an organisation to ensure that the IT systems are reliable and incorruptible (IT General Controls). The main standards frameworks for the IT General Controls are the international ISO/IEC 27002 norm and the BIO.

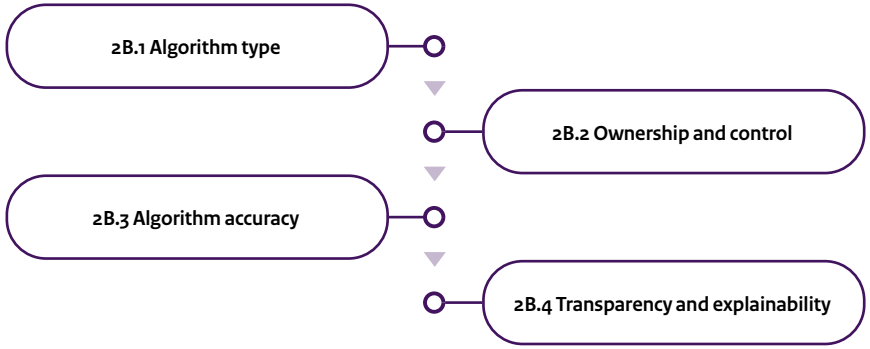
Where relevant, archiving rules must be observed. This prevents, among other things, that data is kept longer than permitted by law.

Part 2B: What?
Algorithm – throughput

Introduction

Just like Part 2A of the FRAIA, Part 2B is about the 'What?' of the project (see the introduction to the FRAIA), but this set of questions pivots on the algorithm itself. This concerns, for example, the question what type of algorithm is used and what the preconditions are for responsible use of the algorithm. For further information, please also see the introduction to Part 2 of the FRAIA.

This section covers the following topics:



2B.1
Algorithm type

2B.1.1 What type of algorithm will be used? Differentiate between:

a. A non-self-learning algorithm in which humans specify the regulations the computer must observe;

b. A self-learning algorithm, in which the machine itself is finding patterns in the data.

2B.1.2 Why is this type of algorithm chosen?

2B.1 Algorithm type

2B.1.3 Why is this type of algorithm best suited to achieve the objectives formulated in question 1.2?

2B.1.4 What alternatives are there and why are they less appropriate or useful?

2B.1 Algorithm type

Expertise/role required for answering this question: data scientist, algorithm developer, project leader

Directions and explanation

One example of a **non-self-learning algorithm** is an algorithm that checks whether a parking ticket has been paid in full.[8] In this instance, a human being might specify that 'if the payment reference for the transaction matches the outstanding ticket, and if the amount of money that has been paid in matches the fine, the parking ticket has been paid in full'. On the basis of this algorithm, a payment confirmation can be sent to the person the ticket was linked to. Another example is an algorithm for fraud alerts, for which an expert specifies that 'if the supplied administrative information contains more than three errors', the system should give off an alert that this dossier is to be investigated further.

One example of a **self-learning algorithm** is an algorithm that learns whether a picture is showing a dog or a cat on the basis of labelled sample pictures. Another example is an algorithm that learns how specific customer characteristics relate to the sales to these customers on the basis of sales figures and customer data. In the above example about fraud alerts, a self-learning algorithm may be used, for instance because the system is set up in such a way, based on examples of fraud and non-fraud, that it can learn which characteristics and behaviours turn out to be related to fraud.

In the case of **non-self-learning algorithms**, it is essential that **one specifies how the computer must function**, while in the case of **self-learning algorithms, one needs to specify what the computer must learn and how it is allowed to learn** (type of algorithm). Please note that we acknowledge that the definition of **self-learning** and **non-self-learning** algorithms is under discussion, but in view of the intended objective of the FRAIA, we are opting for the definition mentioned above. In addition, it is important to note that there are many variants of self-learning algorithms. One may think, for example, of more traditional algorithms that work on the basis of linear or logistic regression, but also of a decision tree or a Support Vector Machine (SVM), or of deep-learning algorithms such as convolutional neural networks.

The reason why the **distinction between these main types is relevant** is that they evoke different questions about the use of an algorithm. When a human specifies what the machine must do, it is relevant to reflect on the people who specify these rules, on the (learning) process they have followed to acquire this knowledge, and on the quality and legitimacy of the choices they make. Relevant to this type of algorithm are questions such as 'does the human expertise provide a sufficient foundation to shape the system this way?', 'is there sufficient consensus among experts that these "rules" are substantiated and apply to the context in which the algorithm is used?', and 'how to safeguard that human expertise will remain available to assess the quality and suitability of these rules in future?' By contrast, with regard to the self-learning algorithms, the reflection must chiefly concern the 'machine', that is, how it can learn, based on what, and whether that process is suited to the context in which the algorithm should be used.

2B.2
Ownership and control

2B.2.1 If the algorithm has been developed by an external party: have clear agreements been made on the ownership and management of the algorithm? What are those agreements?

Expertise/role required for answering this question: commissioning client, project leader, legal advisor

Directions and explanation

In some cases, a government organisation will choose to have an algorithm developed by a third party. In such cases, clear arrangements will need to be made about ownership and control of the algorithm. Also consider, in this instance, the impact of any algorithm updates to be carried out by the third party. Moreover, ownership is relevant to the explainability of the algorithm: even if the algorithm has been developed by a third party, the organisation actually using the algorithm must be able to explain how it operates. In this respect, you may also look back to Part 1, where the responsibilities regarding the development and use of the algorithm were discussed.

2B.3
Algorithm accuracy

2B.3.1 What is the accuracy of the algorithm? On the basis of which evaluation criteria is this accuracy determined?

2B.3.2 Is the level of accuracy (question 2B.3.1) acceptable for the way the algorithm will be will be used?

2B.3
Algorithm accuracy

2B.3.3 How is the algorithm tested?

2B.3.4 What measures can be taken to counteract the risks of reproduction or even amplification of biases (e.g. different sampling strategy, feature modification, ...)?

2B.3
Algorithm accuracy

2B.3.5 What assumptions underlie the selection and weighting of the indicators? Are those assumptions valid? Why or why not?

2B.3.6 How often is the algorithm wrong? (e.g. in terms of number of false positives, false negatives, R-squared, ...)

2B.3 Algorithm type

Expertise/role required for answering this question: data scientist, algorithm developer, domain expert (staff member with domain knowledge in the area where the algorithm is to be applied)

Directions and explanation

An algorithm achieves certain results on the basis of the input data and rules it follows. **It is desirable that these results are correct as often as possible**. As described in 2B.1, non-self-learning algorithms can be distinguished from self-learning algorithms. '**Accuracy**' is important to both these algorithm types. Moreover, in regard to both types, it is necessary to assess how often the algorithm will be right and how often it is wrong, but the methods to assess accuracy differ from each other. This has to do with the fact that for non-self-learning algorithms, the accuracy of the rules specified by people has to be assessed, while for self-learning algorithms, the accuracy of the machine-learned rules has to be considered.

As mentioned, the assessment of **non-self-learning algorithms** should focus on the **accuracy of the rules specified by people**. The example of the algorithm that checks whether a parking ticket has been paid in full may be used to illustrate this (see explanation in question 2B.1). A person might specify, for example, that 'if the payment reference for the transaction matches the outstanding ticket, and if the amount of money that has been paid in matches the fine, the parking ticket has been paid in full'.

The algorithm can present two solutions: [1] The ticket has been paid in full or [2] the ticket has not been paid in full. In order to assess the accuracy of this algorithm, four potentially occurring situations should be considered:

a. Cases in which the ticket has been paid in full and in which the algorithm states it has been paid in full (so-called true positives).
b. Cases in which the ticket has not been paid in full, but in which the algorithm states it has been paid in full (so-called false positives).
c. Cases in which the ticket has not been paid in full and in which the algorithm states it has not been paid in full (so-called true negatives).
d. Cases in which the ticket has been paid in full, but in which the algorithm states it has not been paid in full (so-called false negatives).

The algorithm can be characterised as 100% accurate if it always returns true positives and true negatives. In other words, if the ticket has been paid in full, the algorithm always arrives at that conclusion, and if the ticket has not been paid in full, the algorithm always arrives at that conclusion, too. In practice, however, the programmes rules may not always yield the desired results. Imagine, for example, that the algorithm only looks at the payment reference in the designated payment reference box rather than in the comments box. In practice, all transactions by persons who enter the payment reference in the comments box would then wrongly be regarded to be 'not paid in full' ('false negative'). In such a case, it is necessary to investigate and discuss whether [1] adjustments to the algorithm are needed and/or [2] this degree of accuracy would be acceptable in the context in which the algorithm will be used.

2B.3 Algorithm type

With **non-self-learning algorithms**, it is also important to **chart all possible situations as well as possible**, in order to **assess whether the algorithm yields the desired results in all those cases**. To assess the accuracy, it is important for stakeholders to have the opportunity to chart all possible situations and the interactions with the algorithm; this process requires sufficient human reflection.

With **self-learning algorithms**, there are alternative opportunities to assess accuracy. An example would be an algorithm that is shown pictures of dogs and cats and has to learn which pictures show a dog and which pictures show a cat. One widely used approach to developing this algorithm is that the developers are allowed to 'train' the algorithm using part of the pictures and are allowed to 'test' using a set of unseen pictures. On the basis of this as yet unseen test set, the developers may assess in how many cases the algorithm has produced the correct result.

More specifically, they can assess here, too, how many true positives, true negatives, false positives, and false negatives there are. The simplest measure for accuracy, again, is the number of true positives and true negatives relative to all assessed cases. However, this measure for accuracy is not always the most desirable one. For example, assume that the algorithm has to gauge whether a person has a fatal disease. In that case, it is crucial that the algorithm does not state there is nothing wrong with a person while they actually have a fatal disease. In this situation, a so-called false negative is a major concern. Thus, a seemingly high accuracy of 99% says little if it actually means that many human lives are lost. For such an algorithm it may therefore be necessary to adopt another accuracy/performance measure as the starting point for the discussion. Another point to note is that for assessing the accuracy of self-learning algorithm, it is vital that the training set and the test set are representative for the context in which the algorithm will be used.

The above examples relate to algorithms with a binary outcome ('yes' or 'no', '1' or '0'). Obviously, in practice, there are also algorithms that have a continuous scale as potential outcomes, or more than two categories. For these algorithms, too, there are measures of accuracy, such as the relatively well-known R2 for linear regressions, but they tend to be more technically complex and thus have not been specified in this explanation. Yet, the fundamental question remains the same: to what extent is the algorithm accurate and is this degree of accuracy acceptable in the context in which the algorithm will be used?

2B.4
Transparency and explainability

2B.4.1 Is it clear what the algorithm does, how it does this, and on what basis (what data) it does this? Please explain.

2B.4.2 For which people and groups (internal and external) will the operation of the algorithm be made transparent and how is this done?

2B.4
Transparency and explainability

2B.4.3 For which target groups must the algorithm be explainable?

2B.4.4 Can the operation of the algorithm be explained in a sufficiently understandable manner for the target groups identified in question B.4.3?

2B.4
Transparency and explainability

Expertise/role required for answering this question: project leader, data scientist, algorithm developer, communications consultant, legal advisor, strategic ethics consultant

Directions and explanation

A much-used term in connection with transparency is 'black box'. This term is used to describe an algorithm that is completely non-transparent; the only thing that is clear is which data goes in and which result comes out, but it is not clear what is happening within the algorithm.
Obviously, a black box algorithm is undesirable.

Where the transparency and explainability of algorithms is concerned, a fair number of directions are available in other tools, instruments and guidelines. Therefore, the discussion about these topics may primarily be conducted on the basis of those tools.

According to the Guidelines for Algorithm Application by Governments (Dutch Ministry of Justice and Security, 2021), transparency and explainability can be differentiated as follows. ('Technical') **transparency** is about insight into the algorithmic method that is being applied (decision tree, neural network), the source code, how the algorithm has been trained, as well as the data, input variables, parameters, and thresholds used, et cetera. **Explainability** is about being able to expound on the outcomes of data analyses and how they have come about incomprehensible language.

The above-mentioned Guidelines also differentiate between '**internal transparency**' (i.e., transparency within the organisation and for the benefit of the internal and external controllers, supervisors, judges and stakeholders (namely identified or identifiable persons whose personal data are processed by algorithms)) and '**external transparency**' (i.e., outward transparency, aimed at the general public).
With respect to internal explainability, the Guidelines and the Standards framework of the Netherlands Government Auditing Services (ADR. 2021) likewise stress the importance of **collegial explainability**, i.e., the need for teams to have complete access to and insight into each others' documentation, decisions, and code. It must be ensured that it is safeguarded according to the set standards. In the Algorithm assessment framework of the Netherlands Court of Audit, external transparency is also described as **public explainability**.

Please note, in this FRAIA, 'external transparency' or 'public explainability' has been embedded in the questions about 'communications' in Part 3.

According to the Guidelines for Algorithmic Application by Governments, **the extent to which explainability and transparency are needed** depends on (1) the impact of the algorithm on the decision, the outcome, and the citizen; (2) the degree of autonomy in decision-making (i.e., the degree in which human commitment is guaranteed; and (3) the type and the complexity of the algorithm.
A further explanation of these elements and the way they affect the discussion about transparency and explainability can be found in the above-mentioned Guidelines, which in turn refer to the EU Ethics Guidelines for Trustworthy Artificial Intelligence.

2B.4
Transparency and explainability

With respect to transparency, the term findability is also important. Findability means it is clear how an algorithm has arrived at a specific result. See the second question of this section: 'Is it transparent what the algorithm does, how it does this, and on the basis of what (which data) it does this?'.

Findability plays an important part in the FAIR principles. The FAIR principles are about findability, accessibility, interoperability, and reusability.

Please note, in the FRAIA, the three above-mentioned aspects correspond with the aspects of the influence the algorithm has on the intended measures/decisions and the context (see Part 3), also in relation to the impact on fundamental rights (see Part 4), and with the questions about the type of algorithm (See Part 2B). These aspects are discussed below. The above-mentioned factors can also have impact on the manner and nature of communications about the algorithm, as presented in the Guidelines. This are discussed in Part 3.

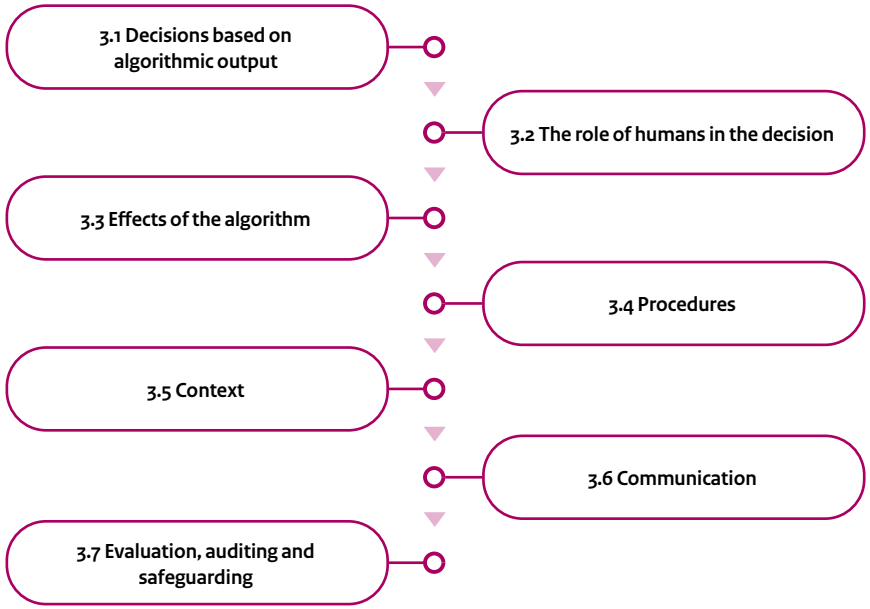
Part 3: How?
Implementation, use and supervision – output

Introduction

An algorithm as such does not have undesired effects. They are always caused by the implementation, use, or application of the algorithm, by the context in which the algorithm is used, or by the decisions and measures that are linked to the output of the algorithm.[9]

Therefore, Part 3 is about the implementation and the use of the algorithm concerned, i.e., about (handling) the algorithm's output. The work method has been presented in the introduction to the FRAIA.

This section covers the following topics:



Enter here which persons complete this section. Also state their function.

Part 3: How?
Implementation, use and supervision – output

Part 3 builds upon/is related to:

- Non-discrimination by design guideline (Parts 5: implementation and 6: evaluation)
  These parts of the Guideline present points of interest similar to the ones in this section of the FRAIA. However, they are more specifically focussed on tackling bias or discriminatory treatment in the data. Thus, the Guideline may be used for further study and information regarding the questions formulated below. This guideline is only available in Dutch.

- Guidelines for Algorithm Application by Governments and public education on data analyses (Dutch Ministry of Justice and Security, 2021)
  These Guidelines pivot on the topics of transparency, explainability, validation and accountability. They start from the assumption that the degree of transparency, explainability and accountability is connected with the degree of impact, the autonomy, and the complexity of the algorithm.
  The elaboration provided in the guidelines may be involved in the discussion about the questions below. These guidelines are only available in Dutch.

- Guideline for public education on data analyses (part of the Guidelines for Algorithm Application by Governments and public education on data analyses (Dutch Ministry of Justice and Security, 2021). This Guideline provides specific leads with respect to communicating about the operation and use of algorithms to a broader public (of non-experts). These leads are relevant to the questions about communication in this part of the FRAIA. These guidelines are only available in Dutch.

- Ethics Guidelines for Trustworthy Artificial Intelligence
  Among other things, these guidelines, developed within the EU framework, pay attention to information provision and communication about the operation of algorithms, as well as to the (human) control and supervision of the use of these AI systems, specifically considering the specific context in which these systems are to be used. These guidelines are relevant to answering questions about the implementation of an algorithm, so it can be useful to consult the Guidelines in that respect.

3.1 Decisions based on algorithmic output

3.1.1 What happens with the results or outcome of the algorithm? What decisions are based on them?

Directions and explanation

This question is particularly about discussing **which (types of) decisions** may be based on the algorithmic output. They may be various types of decisions, such as administrative decisions in individual cases, policy measures, or decisions about the use of public resources[10].

The second question focusses on getting a clear picture of whether– given the output the algorithm will generate – **the algorithm can meet the expectations**. To answer this question, you need to consider the answers to the questions about Reason (1.1), Objectives (1.2) and Values (1.3) from Part 1 of the FRAIA. When the algorithm is used in a different context from the one it has been developed or trained for (for example, in a neighbourhood where the population structure or other relevant data clearly deviates from the training data), this may result in incorrect or biased output. When the context changes, is it possible that the assumptions underpinning the algorithm are no longer applicable. Conversely, if the assumptions underpinning the algorithm change, the algorithm might no longer be usable for the chosen context.

Moreover, it is important to determine **the extent and nature of the impact made by the algorithmic output**, partly in the light of the public values and individual interests identified in Part 1. The impact may be limited (for example where an algorithm of a purely descriptive nature is concerned), but the algorithmic output may also play an important role (for example, where a prescriptive algorithm immediately determines which decision will be made and thus immediately generates legal effects).

3.1 Decisions based on algorithmic output

Impact can also be determined by other factors than the algorithm itself. When, for example, an algorithm only affects the efficiency of the internal work processes of a court, the impact is different from when an algorithm is used for a highly automated decision-making process that affects millions of citizens.[11] The degree of algorithmic impact is also connected to the **human role** in making decisions on the basis of an algorithm. Given the importance of the human role, a separate question about this has been included in this FRAIA (question 3.2).

It is important to **identify as many tangible factors as possible** that may affect the extent and the nature of the impact a decision has and to discuss the consequences of this impact, for example regarding decision-making procedures or civic engagement in the creation of policy measures. In connection to this, this section focusses on safeguarding **good governance and good administration** regarding decisions based on algorithmic output.[12] They also include civic participation and legal protection.

3.2
The role of humans in the decision

3.2.1 What role do humans play in decision-making on the basis of the algorithmic output? How are they empowered to make decisions responsibly on the basis of the algorithmic output?

3.2.2 Is there sufficient qualified staff in place to manage, review and adjust the algorithm, if needed, and will there be in future?

3.2 The role of humans in the decision

Expertise/role required for answering this question: project leader, data scientist, HR staff member, strategic ethics consultant

Directions and explanation

It is generally deemed important that humans have sufficient control over the output generated by an algorithm. That control can be exerted in various stages. For example, this FRAIA is already making clear that people are closely involved in the choice of an algorithm as a decision-making or policy development tool and in the design of that algorithm.

Even once an algorithm is implemented, human intervention is vital. In some cases, decision-making can **be fully automated**, so that the output of an algorithm directly generates a (legally binding) decision. In those cases, too, human intervention is usually deemed necessary, even if only in the form of sufficient control and supervision.

The necessity of human involvement particularly becomes evident from the Ethics Guidelines for Trustworthy Artificial Intelligence that have been drawn up within the framework of the EU. These Guidelines take a detailed look at human intervention in case of decisions that are based on the algorithmic output. The guiding principle there is that people must always be able to make informed, autonomous decisions with respect to algorithmic output. They need to have the knowledge and resources to be able to understand algorithms and deal with them satisfactorily, and, where necessary, they must be enabled, within reason, to control or challenge the system.

Moreover, according to the Ethics Guidelines for Trustworthy Artificial Intelligence, it is crucial that humans are involved in making decisions or taking measures on the basis of an algorithm. According to the Guidelines it is actually vital that humans are involved in all stages of an algorithm's lifecycle (i.e., from design to implementation and supervision) and are able to intervene. This is called 'human in the loop'. Other terms that can cover human intervention are 'human on the loop' and 'human in command'; these terms are further explained in the Guidelines.

If the algorithmic output does not directly inform a decision and it is **a human being who makes the ultimate decision**, it must at least be ensured that these decisions are made responsibly and carefully. Among other things, the risk of prejudice in making decisions that are based on algorithmic output should be considered (see the explanation to question 2A.3).

Developing and managing an algorithm can be a complex task, requiring specific knowledge and **expertise**. If the expertise available within the team or within the organisation is limited, this may be a reason to be extra careful with using an algorithm. In that respect, staff developments need to be considered. It should be ensured that the required expertise will also be safeguarded in future, and the measures needed to do so must be inventoried. Moreover, practical matters must be considered, such as dividing work between staff members. When cooperating with third parties it may be important to make specific agreements, for example that on completion of the project the algorithm will be turned over to the government institution in such a way that it is possible for the institution or other third parties to maintain it (i.e., including all the documentation, et cetera).

3.3 Effects of the algorithm

3.3.1 What will be the effects of using the algorithm for citizens and how will the 'human measure' be considered when making decisions based on the algorithm?

3.3.2 What are the risks of stigmatisation, discrimination or otherwise harmful or adverse effects on citizens? How will these be combated or mitigated?

3.3 Effects of the algorithm

3.3.3 How will the expected effects contribute to solving the problem that prompted the development/deployment of the algorithm (see question 1.1) and to achieving the proposed objectives (see question 1.2)?

3.3.4 How do the expected effects relate to the values being served (see question 1.3)? How are risks of undermining certain values handled?

3.3 Effects of the algorithm

Required expertise/role for answering this question: project leader, data scientist, HR employee, strategic advisor on ethics, client, possibly citizen panel or interest group.

Directions and explanations

This part is aimed at getting a clear idea of whether - given the output that the algorithm will generate - the algorithm can meet expectations. In answering this question, use should be made of the answers to the questions about Reason (1.1), Objectives (1.2) and Values (1.3) from Part 1 of the FRAIA.

In addition, it is important to determine the size and nature of the impact of the output of the algorithm, also in the light of the public values and individual interests identified in Part 1. The impact may be limited (for instance in the case of a purely descriptive algorithm), but the output of the algorithm may also play a major role (for instance in the case of a prescriptive algorithm that directly determines which decision is taken and thus generates legal effects almost directly).

3.4 Procedures

3.4.1 What procedures are in place for taking decisions on the basis of the algorithm?

3.4.2 How are the various relevant actors (administratively and politically accountable persons, citizens) engaged in decision-making?

3.4 Procedures

3.4.3 How is it being safeguarded that these procedures meet the requirements of good governance, good administration and – where necessary – legal protection?

Required expertise/role for answering this question: project leader, data scientist, HR employee, strategic advisor on ethics, legal advisor

Directions and explanations

It is important to **identify as many concrete factors as possible** that could influence the degree and nature of the impact of a decision and to discuss the consequences of this impact, for example on decision-making procedures or the involvement of citizens in the development of policy measures. In connection with that, this section is about ensuring **good and proper governance** and good administration in decisions that are based on the output of algorithms. This includes citizen participation and legal protection.

3.5 context

3.5.1 Time/period: when will the algorithm be used? For how long will it be used?

3.5.2 Area: where will the algorithm be used? Is it in a specific geographical area; does the algorithm concern a certain group of persons or cases?

3.5 Context

3.5.3 Can the algorithm still be used if context factors change or if the algorithm is used in another context than the one for which it was developed?

Required expertise/role for answering this question: project leader, data scientist, HR employee, strategic advisor on ethics

Directions and explanation

If the algorithm is used in a context other than the one for which it is intended or trained (e.g., in a neighbourhood where the population composition or other relevant data clearly differs from the training data), this may lead to incorrect or biased output. **When the context changes, the assumptions on which the algorithm is based may no longer be applicable**. Vice versa, when the assumptions on which the algorithm changes, the algorithm may no longer be applicable to the chosen context.

3.6 Communication

3.6.1 How open can you be about the operation of the algorithm in the light of the objectives and context of its deployment?

3.6.2 How do you intend to communicate about the use of the algorithm?

3.6 Communication

3.6.3 Will the algorithmic output be visualised, for example, in a table, graph, or dashboard? If so: is the shape of the visualisation or depiction a correct representation of the algorithmic output? Is the visualisation easy to read for various user groups?

3.6 Communication

Expertise/role required for answering this question: project leader, commissioning client, communications consultant

Directions and explanation

In question 2B.4, attention was paid to **transparency and explainability**. It stated that explainability is closely related to internal and external communication about the way the algorithm is used, how it operates, what effects it aims at generating, et cetera. Obviously, in connection to dealing with the algorithm's output this must be considered, too. The questions formulated here aim to inform the discussion about the nature and the shape of that communication.

In this context, **openness** is to be considered a spectrum: organisations may opt to be entirely open or entirely close about the operation and significance of the algorithm, while many forms are conceivable in between. To what extent openness may of should be provided is likely to vary per algorithm.

Thus, the **way in which** the operation of the algorithm is communicated about may vary. Some organisations are passive in providing information and do so when driven by demand, while other organisations are more active, for instance organising information gatherings, creating a 'dashboard' to provide information about the operation of an algorithm, or producing informative videos.

The degree and form of communication further depend on **the public**. This may, for example, be a professional public (judges, lawyers, advocacy bodies, auditors), but it may just as well be a more general public (citizens who will be subjected to automated forms of decision-making, children).
For example, auditors are eager to gain insight into the technical operation of an algorithm, and affected citizens need to know where they can ask questions.

For extra handholds regarding data visualisations, the online Kennedy tool may be consulted: http://seeingdata.org/developing-visualisation-literacy/top-5-things-to-look-for-in-a-visualisation/.

3.7 Evaluation, auditing and safeguarding

3.7.1 Have adequate tools been provided for evaluation, auditing and safeguarding of the algorithm?

3.7.2 Are there sufficient opportunities to render proper account about the algorithm?

3.7 Evaluation, auditing and safeguarding

3.7.3 What possibilities are there for auditors and regulators to attach (formal) consequences to the to the use of an algorithm by the government (e.g., feedback of findings, making recommendations, budgetary consequences, ...)

Expertise/role required for answering this question: project leader, commissioning, data scientist

Directions and explanation

These questions are primarily about **safeguarding accountability**, i.e., accounting for the operation and effects of the algorithm. This concerns the opportunity to ask questions, discuss the correctness of proceedings and being able to connect any consequences (informally or formally) to proceedings. The difference with transparency and explainability is that they concern fairly passive matters (for example, can an explanation of the system be found anywhere, or is the source code online?), while accounting is active; it concerns the opportunity to question and answer, to judge proceedings with possible consequences.

Further guidelines for and explanations of the concepts of accountability and responsibility and the corresponding criteria and standards can be found in the EU Ethics Guidelines for Trustworthy Artificial Intelligence.

Once an algorithm has been developed and is ready for use, it is **important to keep verifying the operation of the algorithm**. As a result of contextual factors or changes in data, algorithms may operate differently than expected, while their operation may change through time. Often, proper arrangements are made at the start of the development or use of the algorithm but this attention wanes during the later stages of the process. Therefore, it is vital to consider a number of steps to monitor the algorithm's operation and make any necessary adjustments even before the algorithm is used.

3.7 Evaluation, auditing and safeguarding

These steps may concern both i**nternal evaluation, auditing, and safeguarding processes** (i.e., processes that are set up in the government body applying the algorithm) and **external processes** (for example, supervision by an external supervisor).

With respect to internal processes, thought should at least be given to the following questions:

- How often and at which moments in time should the use of the algorithm be evaluated? Does your organisation have the right staff in place to do so?
- Are there processes that may be designed to ensure that the use of the algorithm continues to be future-proof?
- How can a check be built in to ensure that the decisions or measures continue to contribute to the reason and objectives for using the algorithm, even if a change takes place in the context in which the algorithm is applied (validation tools)?
- What safeguards ensure that the above-mentioned matters are maintained during the following stages of the development and use of the algorithm?
- Have the requirements mentioned in Part 2B regarding human capital at the organisation where the algorithm is to be used been complied with (i.e., required ICT and data infrastructure, staff with the required capacities, knowledge, and experience)?
- Specifically, with regard to self-learning algorithms: have processes and systems been set up to monitor models (for example, with respect to data drift, concept drift and accuracy)?
- Are sufficient opportunities available to adjust the algorithm or change the use of the algorithm if it turns out not to comply with the reason and objectives (any longer)?

With respect to external forms of supervision and auditing, attention should be paid to the following questions:

- Is there a mechanism for external auditing and supervision in place?
- Is sufficient information about the algorithm made available for the supervisor to be able to exercise supervision?
- Is the practice and frequency of performing audits communicated?

Further points of departure and preconditions for both types can be found in the Ethics Guidelines for Trustworthy Artificial Intelligence (EU).

Part 4: Fundamental rights
Roadmap

Introduction

The first three parts of the FRAIA contain questions and points of interest that are relevant to all algorithms. Discussing them may contribute to algorithms being used in a careful, thoughtful, and well-embedded manner.

Given the importance of protecting fundamental rights, and the particular risks that algorithms may pose for violation of those fundamental rights, it is essential to pay special attention to this subject. That is why Part 4 of the FRAIA includes a 'fundamental rights roadmap' with a twofold objective:

1. **It serves as a tool to identify whether the algorithm to be used will affect fundamental rights;**
2. **If so, it facilitates a structured discussion about the question whether there are opportunities to prevent or mitigate this interference with the exercise of fundamental rights, and whether there are reasons why the (mitigated or unmitigated) fundamental rights interference should nevertheless be considered acceptable.**

As stated in Part 1 of this FRAIA, various steps in this model (particularly step 1, and step 7 up to a point) are also relevant to answering the question which public values may be served and affected by an algorithm (question 1.1).

This section covers the following topics:

- 4.1. Fundamental right
- 4.2. Specific legislation
- 4.3. Defining seriousness
- 4.4. Objectives
- 4.5. Suitability
- 4.6. Necessity and subsidiarity
- 4.7. Balancing and proportionality

Enter here which persons complete this section. Also state their function.

Part 4: Fundamental rights
Roadmap

The fundamental rights roadmap broadly comprises the following seven steps

These steps can be succinctly explained as follows:

1. **Fundamental right:** does the algorithm affect (or threaten to affect) a fundamental right?
2. **Specific legislation:** does specific legislation apply with respect to the fundamental right that needs to be considered?
3. **Defining seriousness:** how seriously is this fundamental right infringed?
4. **Objectives:** what social, political, or administrative objectives are aimed at by using the algorithm?
5. **Suitability:** is using this specific algorithm a suitable tool to achieve these objectives?
6. **Necessity and subsidiarity:** is using this specific algorithm necessary to achieve this objective, and are there no other or mitigating measures available to do so?
7. **Balancing and proportionality:** at the end of the day, are the objectives sufficiently weighty to justify affecting fundamental rights?

Lawyers are bound to notice that this roadmap lacks the step of finding an adequate legal basis for a fundamental rights infringement. This may be explained by the fact that this step has already come up explicitly in Part 1. Therefore, including this step again into the roadmap would result in needless repetition. The roadmap refers back to earlier parts of the FRAIA with respect to a number of other points, so that there is no need to duplicate efforts.

Part 4: Fundamental rights
Roadmap

Expertise/role required for going through this roadmap: project leader, domain expert, legal advisor

This fundamental rights roadmap is connected to/builds on:

- Ethics Guidelines for Trustworthy Artificial Intelligence (High Level Expert Group on AI, 2019)
  Under the heading of 'principles of trustworthy AI', these guidelines define a number of ethical principles that are needed to be able to use AI and algorithms properly. They define accordance with 'fundamental rights as moral and legal rights' as one of the guiding principles, connecting them to core values, such as individual freedom, respect for democracy, justice and constitutional state, and equality, non-discrimination and solidarity. These core values also underpin this fundamental rights roadmap and the list of fundamental rights clusters in Annex 1. The guidelines provide a helpful further explanation of a number of these fundamental rights, specifically with regard to AI.

- Non-discrimination by design guideline
  This Guideline addresses principles that are similar to those contained in the present fundamental rights roadmap but focusses specifically on tackling bias or discriminatory treatment in the data. The guideline is only available in Dutch.

- Guidelines for performing a Data Protection Impact Assessment (DPIA) (Dutch Data Protection Authority)
  In many cases, developing and using an algorithm will involve collection and processing of personal data. In many of those cases, a Data Protection Impact Assessment will need to be performed. Since this involves the fine-tuning of the fundamental right to data protection, we will pay attention to this in step 2 of the fundamental rights roadmap. The document is only available in Dutch.

- Algorithm assessment framework Netherlands Court of Audit (2021)
  This framework includes a number of standards that are relevant to privacy rights, more specifically the right to protection of personal data. Therefore, the audit framework can be included in the discussion if personal data are collected and processed for the benefit of the development or use of the algorithm. This topic is addressed in greater detail in step 2 of this fundamental rights roadmap. The framework is only available in Dutch.

4.1 Fundamental right

4.1.1 Is any fundamental right affected by the algorithm that is to be used?

For the benefit of this study, four main clusters of fundamental rights have been defined in Annex 1 to this FRAIA:

1. Fundamental rights relating to the person (including a number of social and economic fundamental rights)
2. Freedom-related fundamental rights
3. Equality rights
4. Procedural fundamental rights

These four fundamental rights clusters can be split further into sub-clusters, which are also defined in Annex 1. For each algorithm it is necessary to identify which sub-clusters an algorithm affects or may affect. To facilitate this identification, each cluster as defined in Annex 1 is preceded by a brief explanation of its interdependence with the other clusters.[13] The idea is to go through the explanations and the clusters and note down which fundamental rights may be affected by the use of the algorithm.

4.1 Fundamental right

Directions and explanation

Some fundamental rights are bound to be affected more often and more easily than others. Most algorithms result in some form of **unequal treatment**. This is inherent to the categorisation and profiling for with many algorithms are being used, and to the fact that many data sets are biased, in the sense that they reflect stereotypes existing in society or historical patterns of structural discrimination.
Because many algorithms are trained or fed with personal data, the **right to protection of personal data** is likely to be affected.

Furthermore, it is conceivable that use of algorithms has consequences for **procedural fundamental rights and the right to good administration**. Since algorithms are often difficult to know about and difficult to explain, it may be hard for people to detect that an algorithm results in discrimination. This makes it less easy for them to start court proceedings to contest a decision, while it is also harder to prove that the algorithm has caused an unjustified difference in treatment. Differences in knowledge about the operation of the algorithm between the parties involved in litigation may affect the right to equality of arms. Automated decision-making with the aid of algorithms may imperil the principle of due diligence, the obligation to state reasons and the rights of defence that are part of the right to good administration.

Specific aspects of these three groups of fundamental rights have already been addressed in Parts 1-3 of the FRAIA. For example, guaranteeing findability, explainability, transparency, accountability and communication is important in the light of the right to good administration and to obtaining the right to an effective remedy. Careful handling of data (see Part 2A) is essential to countering certain issues of discriminatory treatment and data protection. When answering the questions whether the above-mentioned three types of rights are affected and whether an impermissible restriction or infringement applies (on the basis of the following steps of the roadmap), it is therefore useful to make use of the answers to the questions in Parts 1-3.

Obviously, **other fundamental rights** may also be relevant to an algorithm. Which fundamental rights can be affected depends on the policy area and on the nature and objectives of the algorithm. For example, an algorithm to detect hate speech is likely to affect the freedom of expression and of information, an algorithm to preventively detect health risks will affect various privacy rights and the right to health, while an algorithm that supports finding a job may affect individual autonomy.

It is important to at least investigate for every algorithm whether the three first-mentioned fundamental rights are affected by the algorithm (i.e., right to equal treatment, protection of personal data, procedural and good administration rights). It will then be necessary to consider whether other fundamental rights may be at risk. This can be achieved by looking into the tables in Annex 1 to this document and finding out which fundamental rights (clusters) may possibly be affected by an algorithm. Once this has been surveyed, it is necessary to go through the subsequent steps in this roadmap for all fundamental rights identified as relevant, in order to establish whether, in spite of these effects, it is permissible and acceptable to use the algorithm.

4.2 Specific legislation and standards

4.2.1 Do specific legal provisions or standards apply to the fundamental rights infringement?

Depending on the fundamental right that is affected (please note, there may be more than one) it needs to be identified whether specific legislation applies in which the exercise of that fundamental right is further elaborated on.

4.2 Specific legislation and standards

Directions and explanation

In many laws, **specific standards for the protection and restriction of fundamental rights have been laid down**.

Firstly, one may think of l**aws that are exclusively about fundamental rights**, such as the GDPR, equal treatment legislation or specific legislation pertaining to freedom of expression or the right to demonstrate. Moreover, **legislation may include specific provisions** that protect a specific aspect of a fundamental right, such as legislation about media regulation (which may be relevant to freedom of expression and of information) or legislation about medical treatment (which may encompass norms relevant to individual autonomy, privacy or bodily integrity). In many cases, such legislation include special requirements or criteria that a fundamental rights restriction need to comply with. Those requirements or criteria also apply if an algorithm is used to automate or support decision-making or implementation.

Sometimes, such **specific legislation aims to be exhaustive**, that is, it intends to detail a certain fundamental right (for specific cases) in such a way that it covers any type of restriction. For example, such is the case with the GDPR in relation to the right to data protection (see also below). In this case, assessing the consequences of the algorithm that is to be used in light of the relevant legal provisions will suffice and proceeding with the roadmap is no longer required, since the steps in the roadmap are already incorporated into legislation.

In other cases, **legislation is not exhaustive**, or only regulates a small part of the matter that the algorithm may relate to. In that case, a further, broader fundamental rights assessment is in order, on the basis of steps 3-7. Therefore, if it is evident that no (or limited) specific legislation applies to the algorithm, you have to continue applying the roadmap.

In connection with the discussion about specific legislation, special attention must be paid to the equal-treatment legislation and the legislation regarding protection of personal data.

If the **right to protection of personal data** is under discussion, then – depending on the context in which the algorithm is developed – the GDPR or the Law Enforcement Directive (2016/680) may apply. This secondary EU legislation includes detailed rules about the way data can be collected and processed. If this legislation applies, you will usually be required to perform a **data protection impact assessment (DPIA)** before deploying an algorithm.

4.2 Specific legislation and standards

Given the more general character of the FRAIA, we will not go into the special requirements imposed on processing personal data. However, it should be emphasised that it is essential to check whether a DPIA must be performed. As it happens, data processing must not take place before the DPIA has been completed. If it becomes clear from the DPIA there are large risks that cannot be sufficiently restricted, at least a consultation with the competent Data Protection Authority is needed about data processing (and about the algorithm that is based on or fed with it).

In addition, there are a number of EU directives about equal treatment, and there will be national legislation implementing and possibly complementing these directives. The relevant legislation prohibits unjustified unequal treatment on a number of grounds in employment and – in some cases – provision of goods and services and social protection. When an algorithm relates to one of these areas and, may cause a difference of treatment directly or indirectly based on one of the protected grounds (religion, belief, political opinion, race, gender, nationality, sexual orientation, age, disability, chronic disease, number of working hours, temporary or permanent employment), the starting point is that the difference in treatment is prohibited, unless there is a justification for it. For direct discrimination (i.e., unequal treatment that is directly related to a protected ground, such as race or gender) the exceptions are exhaustively listed in the legislation. For indirect discrimination (i.e., unequal treatment that is based on a non-protected ground – such as language – that does disproportionately burden a group characterised by the protected ground, such as people of non-national origin) a broader justification is required that roughly needs to comply with the requirements described below in steps 3-7.[14]

4.3 Defining seriousness

4.3.1 How seriously is a fundamental right affected by the algorithm?

If an algorithm affects or threatens to affect a fundamental right, this does not necessarily mean that it cannot be used. However, it does mean that additional requirements must be imposed on the justification of the interference and the amount of due diligence required before the decision to develop or use the algorithm can be made (see elements 3-7 in the roadmap). These justification and due diligence requirements do not always have to be equally severe. Thus, three grades of required justifications can be differentiated in decision-making:[15]

a. **Serious interference, thus compelling reasons required as justification (red)**

b. **Medium-serious interference, thus due diligence required (yellow)**

c. **Less serious interference, thus no special due diligence required (green)**

Steps 4-6 must be completed for all fundamental rights infringements. Depending on the 'colour', the requirements imposed for those steps may be severe (red), medium (yellow) or light (green). This will be further elaborated for each step below.

4.3 Defining seriousness

Directions and explanation

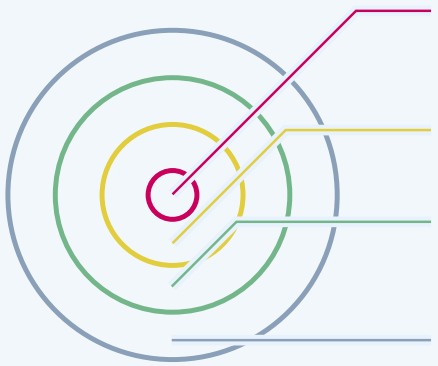
The attributable colour code (which determines the degree of due diligence needed in assessing whether steps 4-7 have been complied with) depends on two main factors:

i) **Which aspect of the fundamental rights identified in step 1 is affected?**
ii) **Is the expected interference of the fundamental rights serious, limited or something in-between?**

These two factors are closely interrelated, and the one should always be considered in conjunction with the other. To do so, you may examine to what extent a specific algorithm (or its application) will affect the **core of a fundamental right**. This **core is determined on the basis of the public value** the fundamental right expresses. The core of the privacy rights, for example, lies in the values of personal autonomy, human dignity, physical and mental integrity, and identity; the core of freedom rights lies in personal autonomy, pluralism, and democracy. A useful starting point may be that **as an algorithm (or its application) makes it harder to realise these values, greater demands should be placed on the justification of the fundamental rights restriction.**[16]

An example (which is not focussed on algorithms) may be found in the measures that can be taken in the context of countering the Covid-19 pandemic. It can be assumed that these measures interfere with a large number of fundamental rights, but looking at the freedom of movement, it can also be assumed that full liberty of movement (= autonomy to choose for yourself where, how, with whom, et cetera, you move outdoors) constitutes the core. The full realisation of that core is the benchmark. The harder it becomes to realise that core, the more serious the fundamental rights infringement will be, and the greater the due diligence and justification demands should be.

The further a fundamental rights infringement is away from the core of the fundamental right, the less severe that infringement is, and the more moderate due diligence and justification demands may be.



When discussing this step, such a graph should be created for each (intended) algorithm and each identified fundamental right, following which the expected fundamental rights infringement can be positioned on it. Selecting a colour code is likely to invite quite some discussion, since no objectively and unequivocally right answer can be given to the question of whether an algorithm affects the core of a fundamental right or rather the periphery. The main thing is to discuss this issue in a well-considered and open manner and to explain comprehensibly why you have opted for red, yellow, or green. This is vital, as the choice of colour determines how steps 4-7 will be applied.

4.4 Objectives

4.4.1 Which objectives are pursued by using the algorithm?

Directions and explanation

It is necessary to determine which objectives are intended in deploying the algorithm, in order to be able to assess, in steps 5 and 6, whether an algorithm is a suitable and necessary tool to achieve these objectives (or whether a different tool would suffice). If, for example, it proves impossible to achieve certain objectives by means of the algorithm, while its use does affect fundamental rights, in principle, the tool must not be deployed. In view of the ultimate balancing of interests or cost-benefit analysis in step 7 (given its objectives, is it desirable to use the algorithm in spite of the fact that fundamental rights are infringed?), it is likewise important to be clear about the objectives that are pursued.

In the context of the FRAIA, the objectives of developing or using the algorithm have been defined in question 1.2, to further fine-tune and substantiate the public values served by using the algorithm. Those objectives can be inserted into this part of the roadmap.

4.5 Efficacy

4.5.1 Is the algorithm that is to be used a suitable means to realise the set objectives?

Directions and explanation

The question about suitability (or efficacy, or effectiveness) is an **empirical question** that departs from '**evidence based**' policies and regulations. Answering it requires foresight: a realistic prognosis needs to be made of the specific yield of using an algorithm and of the various consequences that use will have. How likely is it that the algorithm will actually result in cutting costs or in gaining efficiency, if the costs of developing it and auditing it afterwards are considered? How great is the benefit of using an algorithm where predicting risk categories relative to the present situation is concerned (e.g., professionals appraise the risk that someone gets into debt compared to the situation in which an algorithm makes that appraisal)?

From a technical perspective, suitability may also have to do with the question whether an algorithm is **reliable and accurate** (see question 2B.3). If many false-positives or false-negatives are expected, this may be so problematic (certainly in the case of the red colour code (see step 3)) that the decision must be made not to develop or use the algorithm.

A measure (whether or not aided by the use of an algorithm) will seldom manage to be 100% effective, so such a high degree of suitability cannot be expected. The expected degree of efficacy and the extent to which proof for this appraisal may be expected can be determined by the **level of the justification and due diligence requirements** represented by the **colour codes** in step 3. The red colour code (serious interference with a fundamental right) means that a far more rigorous investigation may be expected to establish whether the algorithm will result in achieving the set objectives than is the case for the green code. The yellow colour code means that some additional requirements must be formulated, but they need not be as rigorous as for the red colour code.

If there is **more than one objective** (as will usually be the case), it may be unclear what is to happen if a number of those objectives cannot be properly achieved, while others can. This depends partly on the relative importance of these objectives (see step 3 and the objectives set out in Part 1 of the FRAIA) and on the justification and due diligence requirements that can be set, given the colour code that has been attributed in step 3.

A general guiding principle may be that **an algorithm is potentially problematic if it does not appear very efficacious and threatens to affect an important fundamental right (red or yellow colour code)**.

4.6 Necessity and subsidiarity

4.6.1 Is using this specific algorithm necessary to achieve this objective, and are there no other or mitigating measures available to do so?

Directions and explanation

In order to achieve policy objectives, a **wide range of tools and means** can be deployed, including algorithms. Even if one specific tool is chosen, it may often be used in various ways. Moreover, it may sometimes prove possible to soften the detrimental effects of a certain instrument by means of compensating or mitigating measures. The choice that can be made between various tools is central in the question regarding the necessity and subsidiarity of choosing a specific algorithm.

Specific points of interest that must be addressed in the discussion about this are the following:

- What are the various tools and means that can be deployed to achieve the set objectives, apart from the intended algorithm?
- Are non-algorithmic tools available?
- Can a choice be made between several types of algorithms or algorithm providers?

Some of these questions have been addressed in relation to question 1.1; the conclusion arrived at there can be used in the discussion about the present step.

4.6 Necessity and subsidiarity

Once the various options and tools have been charted, the question is **whether deploying the algorithm is the best option**, given its effects on fundamental rights. More generally, with respect to choosing between alternatives, the principle of the '**best available techniques**' applies.[17]

If, technically, **preventive, mitigating or compensating tools or measures** are available as a result of which the set objective still can be achieved, but fundamental rights are infringed to a lesser extent, it is desirable that those tools be deployed. Such preventive, mitigating or compensating measures can also be of a regulatory nature, for example, first providing scope for experimenting and only rolling out legislation following extensive testing; working with sunset clauses and evaluation deadlines; using reporting obligations and auditing resources, et cetera.

Annex 2 to the FRAIA contains a (non-exhaustive) overview of available preventive, mitigating and compensating measures.

If, on the basis of this requirement, the use of a mitigating or compensating measure is opted for, **the fundamental rights roadmap** must be followed once again for this measure. In particular, the question of whether the mitigating or compensating measures really result in a less serious fundamental rights infringement must be examined, especially in case of a red colour code (see question 3) but also in case of a yellow colour code. In case of a green code, there is more scope, and choosing between the various tools and measures is easier.

The question may come up **what the trade-off should be if an alternative appears to be slightly less effective but does appear to affect the fundamental right less**. No hard-and-fast rules are available for this instance – it will mainly require having an open discussion. However, if a red or yellow colour code applies and if doubts about the necessity cannot be removed by mitigating measures, there is very good reason to consider not to introduce or use the algorithm.

4.7 Balancing interests/proportionality

4.7.1 Does the use of the algorithm result in a reasonable balance between the objectives pursued and the fundamental rights that will be infringed?

Directions and explanation

Even if an algorithm appears to be a suitable and necessary tool to achieve the formulated objectives, it is still necessary to always take a final step. This step has to do with **the relative weight of the fundamental right at stake, compared to the relative weight of the social objectives and public values pursued**. Suppose, for example, that an algorithm is an eminently suitable as well as necessary tool to enhance the efficiency of decision-making, but there is a real risk that the tool reinforces historical discriminatory patterns. In that instance, is it considered reasonable to still deploy the tool?

It is not possible to formulate hard and objective criteria for determining the weight of and balancing the various rights, interests, objectives, and public values.[18] We can generally say, however, that **the more serious the expected fundamental rights infringement is, the heavier the social objectives need to weigh** in comparison. For example, in case of a red colour code (see step 3) choosing an algorithm (however suitable and necessary it is to achieve certain objectives) is only acceptable if those objectives themselves are weighty. Determining the relative importance of the objectives that have been identified in step 1 may help to appraise that.

Fundamental rights schedule

The form below can be used in addition to or instead of the answers given in part 4 of the FRAIA. The first row shows the questions that were also asked in part 4 of the FRAIA. Below the row of questions are several empty rows, which can be filled in for each fundamental right in order to arrive at a clear overview of the affected fundamental rights and their consequences. An example case is given in the explanation on the next page, which shows how this scheme can look when completed.

4.1. Fundamental right	4.2. Specific legislation	4.3. Defining seriousness	4.4. Objectives	4.5. Suitability	4.6. Necessity and subsidiarity	4.7. Balancing and proportionality
Which (aspect of) a fundamental right is affected? See explanation and appendix 1 (list of fundamental rights).	Does specific legislation apply to this fundamental right? If so, are they met? See explanation.	Which aspect of the fundamental right is affected and is the expected harm extensive, limited or something in between? Which color code belongs to this (red, orange, green)? See explanation.	What goals are being pursued? See answer to question 1.2. See explanation.	Is the algorithm to be used an effective/suitable/effective means of achieving the stated goals? See explanation.	Is deployment of this specific algorithm necessary, i.e. are there no other or mitigating measures available to do this? See explanation and appendix 2 (list of mitigating measures)	Are the goals – all things considered – weighty enough to justify the violation of fundamental rights? See explanation.

Fundamental rights schedule

Fundamental rights schedule

Completely fictitious fill-in example: a predictive algorithm that is used to combat social security fraud. Of course, this diagram can only be completed when all previous parts (parts 1-3) have already been fully covered and all questions asked must also be discussed. The examples below are therefore only intended to give an idea of what the scheme can look like for one specific algorithm to be used.

4.1. Fundamental right	4.2. Specific legislation	4.3. Defining seriousness	4.4. Objectives	4.5. Suitability	4.6. Necessity and subsidiarity	4.7. Balancing and proportionality
Which (aspect of) a fundamental right is affected? See explanation and appendix 1 (list of fundamental rights).	Does specific legislation apply to this fundamental right? If so, are they met? See explanation.	Which aspect of the fundamental right is affected and is the expected harm extensive, limited or something in between? Which colour code belongs to this (red, orange, green)? See explanation.	What goals are being pursued? See answer to question 1.2. See explanation.	Is the algorithm to be used an effective/suitable/effective means of achieving the stated goals? See explanation.	Is deployment of this specific algorithm necessary, i.e. are there no other or mitigating measures available to do this? See explanation and appendix 2 (list of mitigating measures)	Are the goals – all things considered – weighty enough to justify the violation of fundamental rights? See explanation.
Right to data protection	Yes, GDPR – DPIA required	Zie DPIA	See DPIA	See DPIA	See DPIA	See DPIA
Prohibition of indirect discrimination based on national origin	Yes, General Equal Treatment Act – so it is necessary to check whether the requirements of this law have been met. Outcome check: ...	Algorithm will lead to certain groups of people being checked more quickly for fraud; this can have a strongly stigmatizing effect and it is a 'suspicious' ground of distinction, so this is a far-reaching violation with code red.	Import from question 1.2: combating fraud; good use of government resources; efficiency.	Yes because ...	Yes, because... but only if the following mitigating measures are taken...	No: even if mitigating measures are taken, the disadvantage remains that a protected group can be indirectly very heavily and stigmatized by this measure (code red). These effects are so problematic that the benefits of using the algorithm do not outweigh them.

Fundamental rights schedule

Right to an effective remedy	Yes, General Administrative Law Act – so it is necessary to check whether the requirements of this law have been met. Outcome check: ...	The algorithmic decision-making procedure is set up in such a way that legal protection is open to citizens who are affected by a decision and that sufficient transparency, explainability, etc., is provided. The legal protection is therefore not severely limited. However, it appears that people often find it more difficult to know exactly what they have to fight against in algorithmic decision-making, so there may be some effect on legal protection. All things considered, the breach is relatively minor, so code green.	Import from question 1.2: combating fraud; good use of government resources; efficiency.	Yes, because the decision-making procedure is designed in such a way that all Awb requirements for legal protection are met, namely	Yes, extra attention is paid to explainability and communication to the following groups: ...	Yes, because it is important to ...

ANNEX 1- FUNDAMENTAL RIGHTS CLUSTERS

Introduction

For the benefit of step 1 of the fundamental rights roadmap, this annex includes a more detailed list of the four main fundamental rights clusters:

1. Fundamental rights relating to the person
2. Freedom-related fundamental rights
3. Equality rights
4. Procedural fundamental rights

Below, the relevant core values of each cluster will first be briefly summarised; that core may be important to determining how seriously a specific fundamental right is affected (step 2 of the fundamental rights roadmap). Moreover, we will give several examples of fundamental rights that are closely related to these core values.

This summary is followed by a schematic representation of the various fundamental rights that belong to a specific cluster. A cohesive cluster of fundamental rights is presented in the left-hand column, while examples of the types of fundamental rights that belong to this cluster can be found in the right-hand column. These lists of clusters can be checked systematically as part of the FRAIA, the objective being to identify the various fundamental rights that may be affected by the algorithm in step 1 of the fundamental rights roadmap.

A more detailed explanation of the majority of fundamental rights named in this overview, including further references to jurisprudence and literature, can be found in Gerards et al. 2020. Moreover, further information on this subject can be found in chapter 3 ('Monitoring trustworthy AI') of Ethics Guidelines for Trustworthy Artificial Intelligence' (High Level Expert Group on AI, 2019).

It is important that the before-mentioned fundamental rights are seldom 'absolute'. In many cases, reasonable restrictions can be imposed, or it is possible to regulate the exercise of fundamental rights. To assess whether this is the case, steps 2-7 of the fundamental rights roadmap (Part 4 of the FRAIA) must be followed.

1. RIGHTS RELATED TO THE PERSON

Introduction

The broad category of rights bound to the person is closely related to various core values, particularly to human dignity, personal autonomy, physical and mental integrity, personal identity, and social identity.[19] In short, these core values imply that we need to respect that people are who they are (and want to be), because they all have their own value as persons. Moreover, people must be allowed the opportunity (obviously within reason) to develop themselves and make their personal choices.

Firstly, these values concern 'internal' aspects of being human and of one's personality that must be protected from intervention and interference by others. Typical examples of rights that are closely related to these core values are the right to respect for one's personal environment (such as one's home), the right to protection of personal data, rights that are related to physical and mental integrity, the protection of honour and reputation, and freedom of conscience.[20]

At least as important are the external and social identity-focussed aspects of privacy and individual autonomy, which have to do the interaction with others. Often, people can only be themselves and develop if they can establish and maintain relationships. Specific instances of this fundamental right are the right to respect for family life, the right to get married and, more generally, the right to establish relationships with the outside world, both privately and work-related.

Third, the broad category of rights related to the person can be said to include a number of rights that can be seen as indispensable for exercising the afore-mentioned core values. The ultimate example is the right to life: if this right is not effectively protected, it is impossible to exercise any other fundamental rights or to be able to realise values such as human dignity or autonomy.

Lastly, important conditional fundamental rights are social or economic fundamental rights, such as the right to water and food, the right to decent working conditions, the right to a subsistence minimum, the right to accessible healthcare and the right to a healthy living environment. If such fundamental rights are insufficiently respected and protected, it is impossible (or at least a lot harder) to lead a dignified life in which autonomous choices can be made. [21]

Cluster	Examples
Personal identity/personality rights/personal autonomy	- Right to self-fulfilment - Freedom to decide one's own behaviour - Freedom to style one's outer appearance - Right to choose one's trade, one's education, one's training, etc. - Respect for one's own identity (gender identity/sexual identity, etc.) - Reproductive rights - Right to knowledge about parentage - Name rights - Freedom of contract
Social identity/relational privacy rights/relational autonomy	- Right to respect for family relations/family life - Right to get married - Right to family formation - Right to enter into sexual relationships - Right to start professional/business relationships - Right to access to work/profession - Right to access to a country/residence rights - Right to education
Physical and mental integrity	- Freedom of conscience/freedom of thought - Right to life - Prohibition of torture/inhumane or degrading treatment and punishment - Prohibition of refoulement - Prohibition of (body) searches - Informed consent requirement in case of medical treatment and research - Right to access to healthcare - Respect for legal capacity - Right to die - Right to abortion - Prohibition of (modern) slavery/servitude/forced labour/human trafficking/exploitation
Protection of data/informational privacy rights	- Protection from unauthorised/inaccurate data processing - Right to access to data - Right to correction of data - Right to be forgotten

Cluster	Examples
Communication rights	- Privacy of correspondence - Protection from eavesdropping/phone tapping/interception - Prohibition of unauthorised transmission of communication data - Confidentiality of communications with lawyer, doctor, etc.
Territorial privacy rights	- Freedom of movement - Habeas corpus (prohibition of custody, home arrest, etc.) - Freedom to choose place of residence - Free movement of persons (in EU context) - Right to leave the country - Prohibition of unauthorised tracking of persons (GPS tracker) - Prohibition of unauthorised camera surveillance
Proprietary privacy rights	- Home inviolability (protection from raids/house searches) - Right to property - Protection from expropriation - Protection from searching clothes/bags/laptop/computer, etc. - Intellectual property rights
Reputation rights	- Prohibition of punishable insult/defamation/slander - Protection of honour and good name
Healthy living environment	- Right to sustainable development - Right to environmental protection - Protection from emissions of harmful substances - Right to water - Right to sanitation - Right to access to energy
Social and economic rights	- Right to a minimum subsistence level - Right to social security and assistance - Right to access to education

2. FREEDOM RIGHTS

Introduction

Freedom rights are closely related to the above-mentioned core values of dignity, autonomy, identity, and integrity. However, they rather concern the external aspect: they ensure that individual may unrestrictedly express their own identity.[22] Thus, the freedom of expression ensures that people can share their personal opinions or feelings with others, while the freedom of religion provides scope to actively profess and spread a specific belief (or lack of one). The freedom of association and assembly allows people to gather with like-minded people, while the freedom of demonstration relates to communicating certain opinions to the outside world in a certain way, for example, by means of a protest march. The right to vote allows people to freely choose who will represent them in legislation and policy, et cetera.

The afore-mentioned core values – dignity, autonomy, integrity, and identity – are not the only values underpinning the freedom rights. The free exchange of opinions and ideas, the access to information, an ample freedom of association, the opportunity to vote for a certain candidate without hindrance during elections, the opportunity to submit petitions: all these are essential to allow a constitutional democracy to function well and to leave scope for debate and the confrontation of various opinions and ideas. Thus, freedom rights are also related to core values such as rule of law, pluralism, and democracy.[23]

Sub-Cluster	Examples
Freedom of expression	- Press freedom/journalistic freedom - Artistic freedom - Freedom of science/academic freedom - Freedom to choose manner of expression (oral/in writing, etc.) - Whistle blowing - Journalistic source protection/privilege
Freedom to receive information	- Passive information gathering (right to access existing information) - Active information gathering (right to be given access to government information/principle of open government) - Right to pluriform information - Right to unhindered access to the Internet
Freedom of religion	- Freedom to have or not to have a religion - Freedom of religious manifestation (symbols, rituals) - Freedom to gather with other religious people - Freedom of religious denominations/religious communities - Separation of state and religion (religious neutrality of the state) - Respect for religious/philosophical beliefs in education
Freedom of demonstration	- Freedom of assembly - Freedom of gatherings, protest marches, etc. - Free choice of subject, time, place, and resources - Right to be protected from hostile audiences
Freedom of association	- Freedom to be or not to be a member of an association - Internal freedom of association (one's own choice of members, activities) - Freedom of political parties - Freedom of trade unions - Right to bring collective actions - Right to strike
Political rights/freedoms	- Right to periodical organisation of free and confidential elections - Active and passive voting rights - Right to petition

3. EQUALITY RIGHTS

Introduction

The right to equality before the law serves important values pertaining to rule of law, namely those of legal equality, legal security, and protection from arbitrary government action. Everybody is supposed to be equal before the law, and if a law applies to a specific group, the implementation of that law must really be applied to all those. This prevents public bodies and civil servants from deviating from the law as they see fit and promotes predictable and consistent application of that law.

A second core value is the value of being treated as equal to others.[24] Everybody's needs, qualities, wishes, or talents are different, but those differences should not result in some people or groups being treated as 'inferior'. Everybody has the right to equally participate in society, the economy, and the labour market (the value of inclusion) and groups or persons should not be unduly excluded from access to important social services (the value of accessibility). This element is closely related to the afore-mentioned core values of human dignity, autonomy, identity, and integrity. After all, a lack of equal and inclusive access to important provisions makes it hard to comply with those core values.

The concepts of equality before the law, being treated as equals, inclusivity, and accessibility do not mean that people should always be treated completely the same.[25] There may be good reasons to differentiate between people or between the situations in which they find themselves. Unequal or differential treatment may even be fairer than equal treatment if it means that individual wishes, needs, or capacities are considered more. Moreover, not properly considering the differences between people in policy-making may result in those people being excluded after all (substantive inequality).

Thus, the right to 'equal treatment' and non-discrimination mainly implies a 'fair' treatment that is based on objective, rational grounds, where the values of equality, inclusivity and accessibility must be considered. Once the choice has been made to treat certain cases or persons in a specific way, the requirement of equality before the law applies.

Codifications of the principle of equal treatment usually assume that adverse treatment that is inspired by certain, explicitly stated personal characteristics ('protected personal characteristics') is fundamentally unlawful. That is because rather than being prompted by objective, neutral considerations, decisions based on certain personal characteristics usually relate to bias, or to incorrect or too broad stereotypes or prejudices. Also, such decisions may result from deep-rooted patterns of systematic subordination and discrimination of certain groups.

Thus, a decision that has been based (directly or indirectly) on a protected personal characteristic is suspicious or 'suspect', which means it is only admissible if weighty reasons can be given to justify the decision (cf. the red colour code mentioned in step 2 of the roadmap). Frequently stated protected personal characteristics are race, nationality, ethnicity, sex/gender, sexual orientation, religion, belief, political affinity, birth (legitimate/illegitimate/adopted...), disability, chronic illness, and age. This list is certainly not exhaustive; national or international codifications may contain a range of additional protected grounds. Moreover, the list may be added to in the course of time in light of developing insights that still other groups need to be protected against discrimination.

Sub-Cluster	Examples
Equality before the law	- Equality before the law (application of general legislation to all persons and cases coming within its scope) - Prohibition of arbitrariness - Consistency requirement - Principle of legal certainty
Prohibition of direct discrimination on certain grounds	- Decisions or rules must not (decisively) be based on protected personal characteristics
Prohibition of indirect discrimination on certain grounds	- Decisions or rules must not be disproportionally detrimental to persons belonging to groups with protected personal characteristics
Prohibition of discriminatorily motivated actions	- Prohibition of racist/xenophobically/homophobically etc. motivated actions (e.g., racist violence) - Prohibition of ordering discrimination
Right to substantive equality	- Duty to take differences between people and groups into account
Right to reasonable accommodation/positive action	- Right to periodical organisation of free and confidential elections - Active and passive voting rights - Right to petition
Prohibition of profiling	- Prohibition of creating categories or profiles on the basis of protected personal characteristics, which then underpin decision-making or policy
Prohibition of segregation	- Prohibition of territorial or other types of segregation of groups who are otherwise find themselves in a similar situation and are treated similarly

4. PROCEDURAL RIGHTS

Introduction

Procedural fundamental rights, such as the right to an effective remedy and the right to a fair trial before an independent and impartial judge, are essential to solving disputes effectively and objectively and to providing restorative justice when individual interests have been harmed. Furthermore, these rights are vital to monitoring the exercise of legislative and executive competences by the government. Particularly if fundamental rights have been infringed, it is important for an objective third party to establish this and – if need be – to annul a decision or declare a statutory regulation to be inapplicable.

Procedural fundamental rights are partly instrumental by nature, in the sense that they help realise the afore-mentioned, more material fundamental rights. At the same time, it may be assumed that procedural fundamental rights represent intrinsically established core values, particularly values such as honesty, openness, equilibrium, objectivity, and procedural justice.

In order to be able to realise these core values, an accessible and effective remedy needs to be provided; this may be an independent and impartial court established by law, but in some cases, it may also be another independent institution such as a human rights institution or an ombudsperson. If access to an independent and impartial court is granted, the procedure before this court needs to comply with a wide range of requirements, ranging from a trial within a reasonable period to equality of arms. A number of these requirements and guarantees specifically applies to criminal procedures, such as the presumption of innocence or the prohibition on retroactivity of criminal law. Elements of this may, when applicable, also play a role in other procedures, especially where (reparatory) enforcement mechanisms are concerned.

Procedural fundamental rights are traditionally assumed to particularly have significance for the so-called 'contentious phase', i.e., the phase in which a decision has been made that somebody wishes to complain about, or the phase in which a dispute has arisen between two civil parties. However, it is being increasingly accepted that procedural rights and guarantees should also be granted in the 'pre-contentious' or decision-making phase.[26] In particular, there is a tendency to consider specific general principles of good administration as fundamental rights; in the EU Charter of Fundamental Rights, a right to good administration has even been explicitly codified. This means that principles of transparency, reasoning and due diligence in decision-making can be regarded as procedural fundamental rights. This is particularly important where algorithms are concerned, as issues regarding explainability and findability of algorithmic outcomes frequently occur. In this FRAIA, we have allotted a special place to such principles in Part 1. This means that particularly the 'contentious' procedural rights are important to the contentious phase during the implementation of the roadmap of Part 4, even if it is always worthwhile to also investigate whether the right to good administration can be sufficiently guaranteed when using an algorithm.

Sub-Cluster	Examples
Right to good administration (pre-contentious phase)	- Right to transparency and information provision - Participation and defence rights - Right to due diligence in decision-making - Obligation to state reasons - Prohibition of arbitrariness - Prohibition of abus/détournement de pouvoir
Right to an effective remedy and access to justice (contentious phase)	- Right to an effective remedy - Right to access to a tribunal established by law - Full jurisdiction (entire case must be assessable) - Ius de non evocando (right to not be kept from the judge someone is assigned in a particular dispute) - Prohibition of disproportionately high thresholds (court registration fees, assistance by solicitor, immunities, terms) - Right to funded legal aid - Right to effective enforcement of judgments
Right to an independent and impartial court	- Personal independence (e.g., duration of office) - Functional independence (protection from external pressure) - Institutional independence - Subjective impartiality (no involvement with any of the parties) - Objective impartiality (no legitimate doubt about impartial assessment is possible)
Right to judgment with a reasonable time	- Entire procedure (including pre-procedure) must not take too long - Right to opportunities to speed up the procedure - Right to compensation if procedure takes too long
Right to a fair trial	- Adversarial procedure - Equality of arms (preparation time, access to file/documents) - Balanced and fair rules of evidence - Right to a fair and public hearing - Equal opportunities to hear witnesses/experts - Right to a public judgment - Legal certainty

Sub-Cluster	Examples
Right to a fair trial: criminal safeguards	- Presumption of innocence - Right to remain silent/right to refuse cooperation - Prohibition of entrapment - Ne bis in idem - Right to assistance by a solicitor - Right to assistance by an interpreter
Criminal requirement for legality (no punishment without legislation)	- Prohibition on retroactivity of criminal law - Lex mitior principle

ANNEX 2 - PREVENTATIVE AND MITIGATING MEASURES

Particularly in the fundamental rights roadmap, it has been stated (in step 6: necessity) that it may be necessary to look for measures to prevent the (potentially) detrimental effects on fundamental rights (preventative measures) or at least mitigate or compensate for them (mitigating measures).

On the basis of the existing tools and guidelines (presented in Parts 1-4 of the FRAIA) one may consider a wide range of options. Those options have been listed below (but please note that mitigating measures in the field of processing and protecting personal data have not been included, as they can be dealt with in a DPIA). More details can always be found in the tool referred to and/or in scientific literature.

Measures to prevent issues during development and use of the algorithm

- Lowering bias risks by instance class modification, instance selection or instance weighing.[27]
- Rendering the impact of protected personal characteristics on the operation of the algorithm visible by means of tools for e.g., 'gender tagging'.[28]
- Creating 'fairness-aware' algorithms.[29]
- Using in-processing or post-processing method of 'bias mitigation'.[30]
- Using mechanisms for 'ethics by design', 'equality by design', 'security by design' et cetera.[31]
- Testing and validating.[32]
- Using mechanisms to promote the transparency and explainability of algorithms, such as a 'crystal box' or multi-stage transparency.[33]

Regulatory and administrative tools that may mitigate the impact of algorithms on fundamental rights

- Working on the basis of an experiment or testing ground with a limited data set and limited accessibility; only to be broadened after extensive testing.[34]
- Including sunset clauses and evaluation deadlines into the regulation on the basis of which the algorithm can be used.[35]
- Including obligations to report and perform regular audits, in order to periodically monitor the effects of the algorithm and its impact on fundamental rights.
- Deeming impermissible or adopting a moratorium on a specific use of algorithms.[36]
- Formulating upper limits for the use of specific algorithms.[37]
- Deploying codes of conduct, professional standards or ethical codes for actors who are due to handle an algorithm. [38]
- Developing a variation on the Hippocratic oath for developers and adopters of algorithms (if accompanied by education or training on data ethics). [39]
- Education or training on data-ethics awareness.[40]
- Normalisation, accreditation and certification of algorithms on the basis of their compliance with the requirements for reliable algorithms.[41]
- Using checklists for decision-making on the basis of algorithmic output.[42]
- Providing scope for participation by affected or committed citizens.[43]
- Developing an exit strategy to stop the use of the algorithm in case it turns out to be no longer desirable.

LITERATURE

Alexy, R. (2002), A Theory of Constitutional Rights, Transl. J. Rivers, Oxford: OUP 2002

Altman, A. (2015), 'Discrimination', in: E.N. Zalta (ed.), The Stanford Encyclopedia of Philosophy, internet edition 2015, through http://plato.stanford.edu/archives/fall2015/entries/discrimination/

Netherlands Government Auditing Services (2021), Normenkader Algoritmen Auditdienst Rijk, 2021

Castelluccia, C. & Le Métayer, D. (2019), Understanding algorithmic decision-making: Opportunities and challenges, Report for the Panel for the Future of Science and Technology (STOA) of the European Parliament, 2019

Eubanks, V. (2018), Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, Macmillan 2018

Gerards, J., Xenidis, R. (2021). Algorithmic discrimination in Europe : challenges and opportunities for gender equality and non-discrimination law, Publications Office European Commission. https://data.europa.eu/doi/10.2838/77444

Gerards, J.H. (2005), Judicial Review in Equal Treatment Cases, Martinus Nijhoff: 2005

Gerards, J.H. (2018), 'Core rights and the interaction of normative and analytical elements in human rights scholarship', in: M. Schein-in (ed.), Methods of Human Rights Research (working title), forthcoming, working paper version available through https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333627

Gerards, J.H. (2019), General Principles of the European Convention on Human Rights, Cambridge, CUP: 2019

Janssen, H.L. (2020), 'An approach for a fundamental rights impact assessment to automated decision-making', 10 International Data Privacy Law (2020) (1) pp. 76-106

High Level Expert Group on AI (2019), Ethics Guidelines for Trustworthy Artificial Intelligence, Brussels: European Commission 2020

Khademi, A et al. (2019), 'Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality', through https://arxiv.org/pdf/1903.11719.pdf

Kleinberg, J et al. (2019), 'Discrimination in the Age of Algorithms', through https://arxiv.org/abs/1902.03731

Koops, B.J. et al. (2017), 'A Typology of Privacy', 38 University of Pennsylvania International Law Review 2017 (2), pp. 483-575

Meijer, A. et al. (2021), Code Goed Digitaal Openbaar Bestuur ('Good Digital Governance Code'). https://www.rijksoverheid.nl/documenten/rapporten/2021/04/30/code-goed-digitaal-openbaar-bestuur

Dutch Ministry of Justice and Security (2020), Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses (Dutch Ministry of Justice and Security, 2020)

Nickel, J.W. (2007), Making Sense of Human Rights, 2nd ed., Malden: Blackwell 2007

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

Soriano Arnanz, A. (2020), Posibilidades actuales y futuras para la regulación de la discriminación producida por algoritmos, Diss. Universidad de Valencia 2020 (available in English)

TILT 2020, Handreiking non-discriminatie by design (draft version, 2020)

Dutch Scientific Council for Government Policy (WRR) (2000), Het borgen van publiek belang, report no. 56, The Hague: WRR 2000 (An English summary can be found through https://english.wrr.nl/publications/reports/2000/04/22/safeguarding-the-public-interest)

Wieringa, M. (2020), 'What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability', in: Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain, ACM, New York, NY, USA

AUTHORS

Prof. Janneke Gerards, Utrecht University

Mirko Tobias Schäfer, Utrecht University

Arthur Vankan, Utrecht University

Iris Muis, Utrecht University

ENDNOTES

1 WRR 2000. See also the knowledge database of the Dutch centre on legislation and legal affairs (Kenniscentrum Wetgeving en Juridische zaken (KCWJ)): https://www.kcbr.nl/beleid-en-regelgeving-ontwikkelen/integraal-afwegingskader-voor-beleid-en-regelgeving/5-wat-rechtvaardigt-overheidsinterventie.

2 More details in Gerards 2019b.

3 High Level Expert Group on AI 2019, par. 22 ff.

4 Meijer & Grimmelikhuijsen 2020.

5 Nader Meijer & Grimmelikhuijsen 2020.

6 Meijer & Grimmelikhuijsen 2020.

7 See https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

8 For further information on the differences between self-learning and non-self-learning algorithms, see also Kulk, Van Deursen et al. 2020.

9 See for more detail, among others, Eubanks 2018; O'Neil 2016.

10 See for more detail about the definition of decisions Kulk, Van Deursen et al. 2020.

11 See also Kulk, Van Deursen et al. 2020.

12 See also Meijer & Grimmelikhuijsen 2020.

13 See also Vetzo, Gerards & Nehmelman, 2018; Gerards 2019a; Janssen, 2020.

14 See also Vetzo, Gerards & Nehmelman, 2018; for an assess model for equal treatment issues, see also Gerards 2002.

15 Loosely based on Gerards 2002; Gerards 2018; Janssen 2020.

16 Cf. Alexy 2002.

17 Soriano Arnanz 2020, p. 204 ff.

18 E.g., Den Houdijker 2012.

19 Gerards 2019a.

20 Vetzo, Gerards & Nehmelman 2018, p. 53 ff.; Koops et al. 2017.

21 Nickel 2007.

22 Vetzo, Gerards & Nehmelman 2018.

23 See also Vetzo, Gerards & Nehmelman 2018.

24 Gerards 2019a.

25 Cf. Altman 2015.

26 More details in Julicher 2019.

27 See TILT 2020, p. 39.

28 Kleinberg et al. 2019; Khademi et al. 2019.

29 Castelluccia & Métayer 2019.

30 TILT 2020, pp. 47-48.

31 High Level Expert Group on AI 2019, par. 98.

32 High Level Expert Group on AI 2019, par. 100-101; Dutch Ministry of Justice and Security 2020.

33 Dutch Ministry of Justice and Security 2020.

34 Janssen 2020; cf. Dutch Ministry of Justice and Security 2020.

35 Janssen 2020.

36 Janssen 2020.

37 Janssen 2020.

38 High Level Expert Group on AI 2019, par. 105; Janssen 2020.

39 Gerards & Xenidis 2021.

40 Gerards & Xenidis 2021.

41 High Level Expert Group on AI 2019, par. 106-107; Gerards & Xenidis 2021.

42 High Level Expert Group on AI, 2019, par. 117-118.

43 Cf. Meijer, Schäfer & Branderhorst 2013; Meijer & Grimmelikhuijsen 2020.