Artificial Intelligence Impact Assessment
ECP
Platform for the Information Society
www.ecp.nl

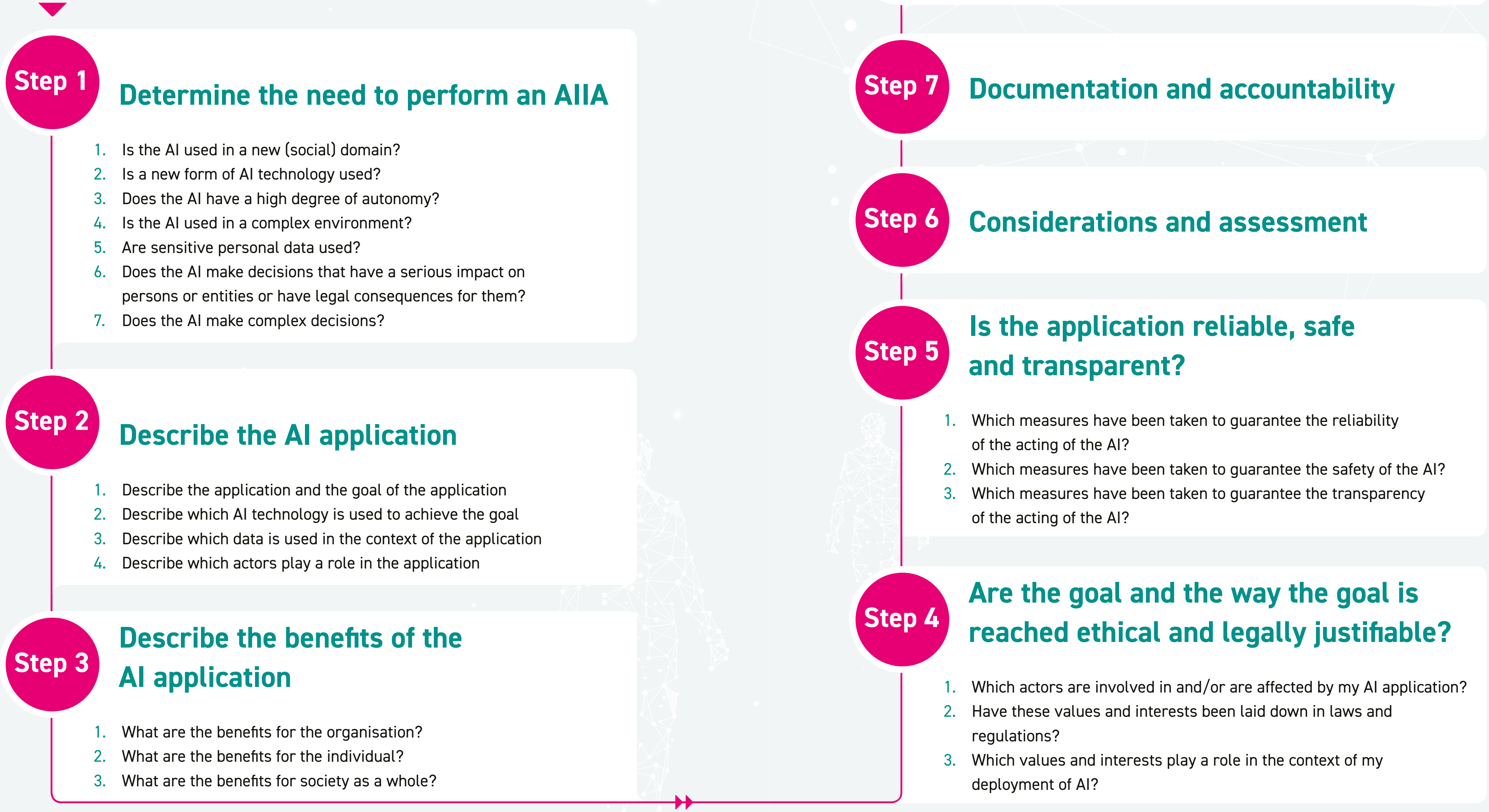
Artificial Intelligence Impact Assessment

ECP
Platform for the
Information Society

THINK BEFORE YOU PRINT

Roadmap for conducting the AIIA

Organisations who want to conduct the AIIA can follow the roadmap below. An explanation to this plan can be found in ‘Part 2: Conducting the AIIA’, page 35.



Step 1 Determine the need to perform an AIIA

1. Is the AI used in a new (social) domain?
2. Is a new form of AI technology used?
3. Does the AI have a high degree of autonomy?
4. Is the AI used in a complex environment?
5. Are sensitive personal data used?
6. Does the AI make decisions that have a serious impact on persons or entities or have legal consequences for them?
7. Does the AI make complex decisions?

Step 2 Describe the AI application

1. Describe the application and the goal of the application
2. Describe which AI technology is used to achieve the goal
3. Describe which data is used in the context of the application
4. Describe which actors play a role in the application

Step 3 Describe the benefits of the AI application

1. What are the benefits for the organisation?
2. What are the benefits for the individual?
3. What are the benefits for society as a whole?

Step 4 Are the goal and the way the goal is reached ethical and legally justifiable?

1. Which actors are involved in and/or are affected by my AI application?
2. Have these values and interests been laid down in laws and regulations?
3. Which values and interests play a role in the context of my deployment of AI?

Step 5 Is the application reliable, safe and transparent?

1. Which measures have been taken to guarantee the reliability of the acting of the AI?
2. Which measures have been taken to guarantee the safety of the AI?
3. Which measures have been taken to guarantee the transparency of the acting of the AI?

Step 6 Considerations and assessment

Step 7 Documentation and accountability

Step 8 Review periodically

Colofon

2018 © ECP | Platform for the Information Society

With thanks to Turnaround Communication

Contents

Foreword

The public debate around AI has developed rapidly. Apart from the potential benefits of AI, there is a fast growing focus on threats and risks (transparency, privacy, autonomy, cyber security et cetera) requiring a careful approach. Examples from the recent past (smart meters, ov-chipkaart (the smart card for public transport)) show that the introduction of IT applications is not insensitive to the debate about legality and ethics. This also applies to the deployment of AI. Mapping and addressing the impact of AI in advance helps to achieve a smooth and responsible introduction of AI in society.

> What are the relevant legal and ethical questions for our organisation if we decide to use AI?

The AIIA helps to answer this question and is your guide in finding the right framework of standards and deciding on the relevant trade-offs.

The "Artificial Intelligence Code of Conduct" is the starting point for this impact assessment and is an integral part of the AIIA. The code of conduct is attached to this document as annex 1. The code of conduct offers a set of rules and starting points that are generally relevant to the use of AI. As both the concept of "AI" and the field of use are very broad, the code of conduct is a starting point for the development of the right legal and ethical framework that can be used for assessment.

The nature of the AI application and the context in which it is used, define to a great extent which trade-offs must be made in a specific case. For instance, AI applications in the medical sector will partly lead to different questions and areas of concern than AI applications in logistics.

"Artificial Intelligence is not a revolution. It is a development that slowly enters our society and evolves into a building block for digital society. By consistently separating hype from reality, trying to read and connect parties and monitoring the balance between options, ethics and legal protection, we will benefit more and more from AI."

— Daniël Frijters,
MT member and project advisor at ECP | Platform for the Information Society

The AIIA offers concrete steps to help you to understand the relevant legal and ethical standards and considerations when making decisions on the use of AI applications. AIIA also offers a framework to engage in a dialogue with stakeholders in and outside your organisation. This way, the AIIA facilitates the debate about the deployment of AI.

"AI offers many opportunities, but also leads to serious challenges in the area of law and ethics. It is only possible to find solutions with sufficient support if there is agreement. The code of conduct developed by ECP and the associated AI Impact Assessment are important tools to engage in a dialogue about concrete uses. This helps to develop and implement AI in society in a responsible way."

— Prof. dr. Kees Stuurman, Chairman of the ECP AI Code of Conduct working group

AI Impact Assessment as a helping hand

The AIIA is not intended to measure an organisation's deployment of AI. Organisations remain responsible for the choices they make regarding the use of AI. Performing the AIIA is not compulsory and it is not another administrative burden. To the contrary; the AIIA is a support in the use of AI. Indeed, responsible deployment of AI reduces the risks and costs, and helps the user and the society to make progress (win-win).

The AIIA primarily focuses on organisations who want to deploy AI in their business operations, but it can also be used by developers of AI to test applications.

We hope that the AIIA will find its way to practice and that it will constitute an effective contribution to the socially responsible introduction of AI in the society.

Prof. dr. Kees Stuurman
Chairman ECP Working Group
AI Code of Conduct

Daniël Frijters
MT member and project
advisor ECP

Drs. Jelle Attema
Secretary

Mr. dr. Bart W. Schermer
Working group member
and CKO Considerati

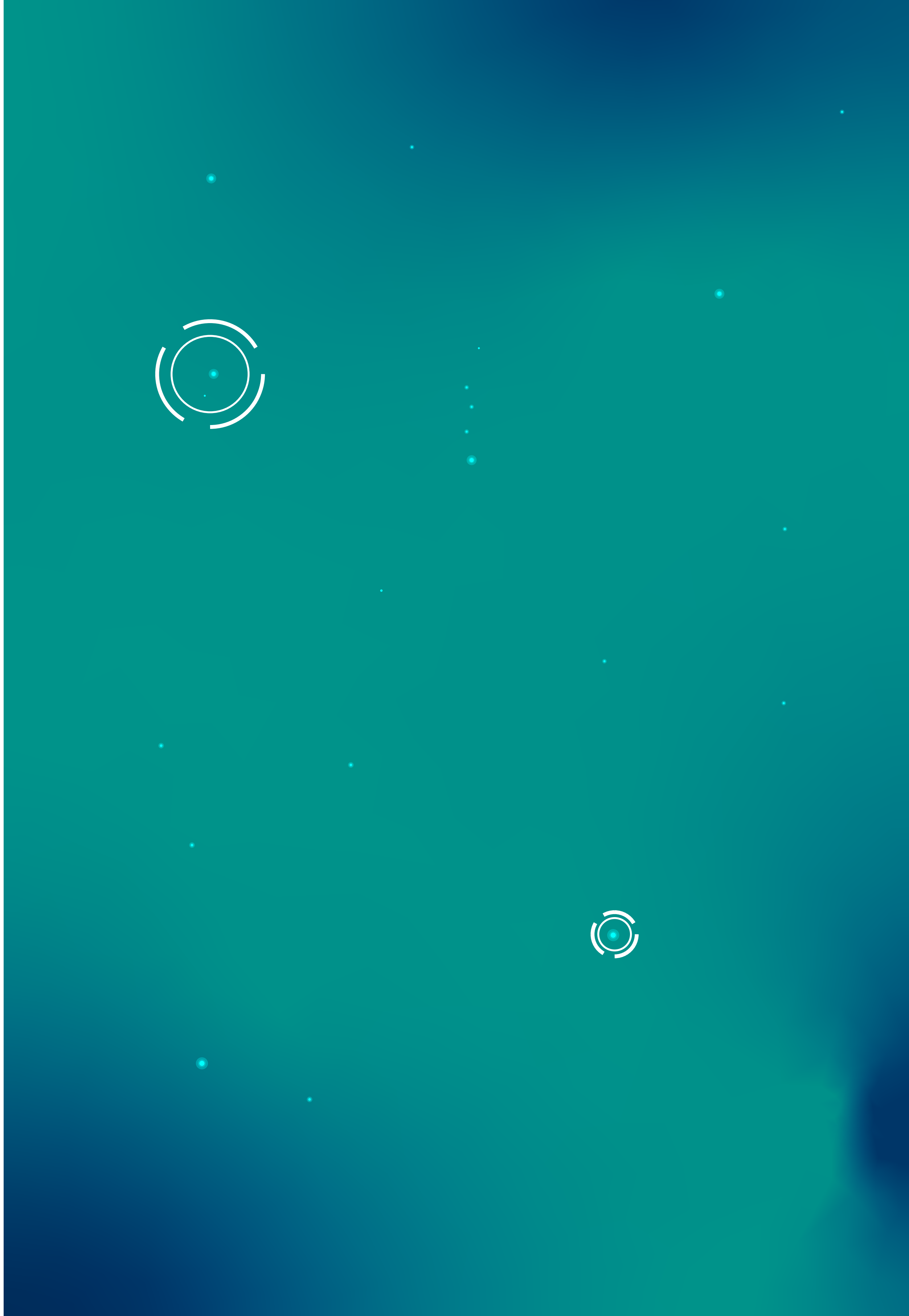
The working group “Artificial Intelligence Impact Assessment” consisted of (in a personal capacity):

Kees Stuurman (chairman) *Van Doorne advocaten, Tilburg University* • **Bart Schermer** *Considerati, Leiden University* • **Daniël Frijters** *ECP | Platform for the Information Society* • **Frances Brazier** *Technical University Delft* • **Jaap van den Herik** *Leiden University* • **Joost Heurkens** *IBM* • **Leon Kester** *TNO* • **Maarten de Schipper** *Xomnia* • **Sandra van der Weide** *Ministry of Economic Affairs and Climate Policy* • **Jelle Attema (secretary)** *ECP | latform for the Information Society.*

The following persons and organisations have made useful comments on the draft version (in a personal capacity):

Femke Polman en Roxane Daniels *VNG, Data Science Hub* • **Staff of the Ministry of the Interior and Kingdom Relations,** *department Information society of the management board of Information Society and Government* • **Marc Welters** *NOREA, EY* • **Marijn Markus, Reinoud Kaasschieter en Martijn van de Ridder** *CAP GEMINI* • **Rob Nijman** *IBM* • **Stefan Leijnen** *Asimov Institute.*

Considerati, on directions of ECP, made a considerable contribution to the preparation of the AI Impact Assessment. We thank in particular Joas van Ham, Bendert Zevenbergen and Bart Schermer for their efforts.



Introduction

The Artificial Intelligence Impact Assessment (AIIA) build on the Guidelines for rules of conduct of Autonomous Systems (“Handreiking voor gedragsregels Autonome Systemen” (ECP.NL, 2006)), which focused on the legal aspects of the deployment of autonomous systems: systems that perform acts with legal consequences. The guidelines were written by a group of various experts: lawyers, business scientists and technicians, from science, industry and government. The initiative for the guidelines comes from ECP. The guidelines at that time were created at the request of ECP participants, from industry and government, because of the seemingly rapid expansion of autonomous systems at the time, and for so-called “autonomous agents”.

In 2006, the Guidelines focused mainly on the legal aspects.The AIIA is broader and now also includes the ethical aspects: a broadly shared opinion in the working group (still consisting for the greater part of the same organisations and people as in 2006) is that AI must improve wellbeing and must not only respect, but also promote human values.

Need for AIIA

As the interest for AI is highly fluctuating, it is legitimate to wonder if and why an Artificial Intelligence Impact Assessment is necessary.

The most important reason is, that AI takes more and more tasks over from people or carries tasks out together with people, whereby the notice of ethics of people has a leading role: in education, care, in the context of work and income and in public bodies. In addition, thanks to AI, organisations can assume new roles, where ethics play a role. For instance in the prevention, control and detection of fraud.

Many of these examples of autonomy and intelligence are not very spectacular, but may nevertheless have a great impact on those who get to deal with these systems.

The AIIA is useful in AI applications that perform acts or make decisions, together with people or not, that used to be done by people and where ethical questions play a role. The Impact Assessment is also relevant if an organisation pursues new goals or performs activities that are made possible by AI and where questions of well-being, human values and legal frameworks are relevant.

The value of the AIIA is not dependent on the degree of autonomy or intelligence of ICT. Even if rapid developments in the area of AI make this question more concrete and more urgent.

Definition of Artificial Intelligence

There is little agreement on the definition of Artificial Intelligence (AI).[1] The AIIA follows the description and approach of the IEEE (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2017).

> "Autonomous and/or intelligent systems (AI/S) are systems that are able to reason, decide, form intentions and perform actions based on defined principles."

The IEEE has taken the initiative to ask more than two hundred experts and scientists around the world to think about

the ethical aspects of autonomous and intelligent systems. Working groups have been created for the various aspects, to define standards and rules of conduct. The document reflects the consensus among a broad group of experts from many parts of the world and cultures.

Core elements from the approach of the IEEE, and also the AIIA, is that applying AI in an ethical way means that AI must contribute to the well-being of people and the planet. The IEEE follows the operationalisation of the OECD of well-being (OECD, 2018). This covers many topics such as human rights, economic objectives, education and health, as well as subjective aspects of well-being. What "contributing to well-being" means for a specific project, requires the analysis and balancing of often many (sometimes contradictory) requirements with a view to the specific cultural context. The AIIA offers the "Artificial Intelligence Code of Conduct" (Annex 1) as a starting point for that analysis. The third aspect that the IEEE emphasizes is that the user of AI is responsible for the impact of AI and must set up processes to realise the positive effects and prevent and control the negative effects.

For whom is the Impact Assessment?

The Impact Assessment is for organisations who want to use AI in their (service) processes and want to analyse the legal and ethical consequences. At the design stage (where expensive errors can be prevented), but also during the use: organisations will often want to see the consequences of their service. Carrying out the Impact Assessment is a lot of work, however, a part can be reused because an important part of the ethical and legal starting points will be generic for a particular technology, for a specific sector or a certain profession.

The organisation that wants to apply AI, conducts the Impact Assessment. Technology should function within the legal and ethical frameworks of the organisation deploying AI, within the frameworks of the professionals who work with AI or transfer parts of their work to technology, end users and society.

The outcomes of the Impact Assessment sometimes lead to certain demands on the technology (specific features), organisational measures (for example a fall-back when end users want human contact, or new task distributions to prevent and deal with incidents), further education and training (how does a doctor, accountant, lawyer or civil servant bear his professional responsibility when tasks are performed by AI; how does a professional interpret the advice of AI, what are the weaknesses and strengths of this advice and how do they come about) and the gathering of data on the exact results in practice.

The provider and producer of the AI solution must ensure that a number of technical preconditions are met (for example, integrity of data, safety and continuity), but must also offer facilities allowing the organisation deploying the AI to take responsibility and to be transparent about the consequences. The provider of the technology can use the Impact Assessment to help organisations ask the right questions and make trade-offs.

> The starting point of this Impact Assessment is that the organisation deploying AI takes responsibility for AI.

This is fundamental for the working group: the black scenarios surrounding AI are usually about technology in which the ethical frameworks are set by an external party (perhaps the manufacturer, a malicious person or the technology itself).

Based on general principles and starting points in hand, this assessment helps to examine what these principles mean for a specific application:

for the design of the technology, for the organisation or the organisation that applies technology, for the administrators who have to account for it, the professionals and specialists working with the technology or delegating tasks to it, for the end users who experience the consequences, and for society

How does the roadmap look like?

Whether it is useful to conduct the Impact Assessment often depends on the combination of service, organisation, end users and society.

Step 1 of the Impact Assessment consists of a number of screening questions to answer the question whether it is useful to carry out the assessment. These questions relate to:

1. the social and political context of the application (experience with technology in this domain, the technology touches on sensitive issues),
2. characteristics of the technology itself (autonomy, complexity, comprehensibility, predictability),
3. and the processes of which the technology is part (complexity of the environment and decision-making, transparency, comprehensibility and predictability of the outcomes, the impact for people).

> With one or more positive answers to the screening questions, it may be useful to carry out the Impact Assessment.

The Impact Assessment then starts with step 2 , the description of the project: the goals that are pursued by using AI, the data that are used, the actors such as the end users and other stakeholders. Think also of the professionals in an organisation who have to work with AI or who transfer work to AI.

The goals of the project are formulated in step 3 , not only at the level of the end user, who experiences the consequences of the service, but also at the level of the organisation offering the service and of the society. This broad approach to goals is important, because ethical and legal aspects are at stake that relate to the relationship between an organisation and its environment.

Step 4 addresses the ethical and legal aspects of the application. In this step, the relevant ethical and legal frameworks are mapped and applied to the application. There are many relevant sources for ethical and legal frameworks for an application: some are formal (laws, decisions), others more informal: codes of conduct, covenants or professional codes.

In **step 5** organisations make strategic and operational choices with an ethical component: how they want to carry out their activities in relation to their customers, employees, suppliers, competitors and the society.

The different facets related to ethical and legal aspects, are weighted in **step 6**. In this step, decisions are made about the deployment of AI.

These steps are concluded by **step 7**: proper documentation of the previous steps and justification of decisions taken,

and by **step 8**: monitoring and evaluating the impact of AI. As the deployment of AI will often lead to changes in the way that ethical and legal aspects are looked at, this will often be the subject of that evaluation.

Interdisciplinary questions and starting points

The Impact Assessment and the Code of Conduct have been fleshed out by a broadly selected group of experts. An important challenge was bridging the different perspectives. A lawyer looks at ethics differently than a provider of these systems, an engineer, an official or an IT auditor. The Impact Assessment and the Code of Conduct have attempted to formulate common questions and starting points that address various disciplines from their own perspective and expertise. The guidelines do not make those discipline-specific analyses superfluous.

Updating AIIA

The Impact Assessment and the Code of Conduct have been adopted according to the insights of today. However, expectations, roles, norms and values change under the influence of the public debate and experiences with new technology. This changes the content of professions and the criteria on which professionals are assessed. The expectations of end users also change when certain technologies become commonplace. It is difficult if not impossible to foresee these changes; that is why planning new assessments and collecting data on the impact of technology are important elements in the Impact Assessment. And this is always done against the current state of affairs in the field of applicable (legal) rules and the public debate.

Social questions

The Impact Assessment examines the consequences of using AI in organisations. It does not give an answer to many issues surrounding new technology: for example, what automation and robotisation does with the content of work and employment, or what AI means for market relations. Issues such as interoperability of datasets and data control are not addressed. The public and political debate on these issues is very important for the requirements that AI must meet. Readers who want to get an idea of these aspects are recommended to read publications such as "Upgrading" (Rathenau Institute) or "Man and Technology" (SER).[2]

Ethical considerations

The Impact Assessment assumes that ethical questions do not only play a role in forms of AI that are not yet possible: the current (simple forms of) AI and much older ICT systems already raise ethical questions.

An important distinction between ethics and AI is the distinction between Artificial Narrow Intelligence (ANI) and Artificial General Intelligence (AGI): the aim of AGI is to have machines perform intellectual tasks just as well as people. To achieve this, these systems need information about what they can do, what their limitations are, what goals they have to strive after and which strategy fits. Often, this information is called "self-consciousness".

The difference between AGI and ANI is that ANI carries out intellectual tasks in a limited domain. Ethical principles also apply, but not exclusively, to systems with ANI or AGI. An important design principle is that people using ANI or AGI must be able to exercise control: to set the ethical frameworks within which the systems act. Organisations are not used to making ethical frameworks explicit and might leave this easily to the designers of systems.

> An objective of the Impact Assessment is that organisations define their ethical frameworks themselves.

A second important distinction between ethics and AI is that many systems classified as "AI" are no longer "pre-programmed" like the ICT systems we know, but are self-learning and adjust their actions and judgment. The classical systems had more computing power than their creators, but they could not be smarter than their inventors. The self-learning systems can ultimately make decisions or perform tasks better than "creators". "Self-learning" means that these systems must be able to make mistakes. And that they sometimes perform tasks in new ways, incomprehensible and unpredictable for people. Issues such as control, transparency and accountability are crucial themes in these self-learning systems: how can we control a system that is better than us, without understanding how. Sometimes this may mean that systems cannot be applied: for example in the domain of the government, where being able to explain a government decision in clear language is a right of citizens.

> The Impact Assessment assumes that control, accountability and transparency do not always have to be part of the system.

If a system is better than people, other measures are needed so that people can exercise control and are accountable. For example, by not allowing a system to "learn" when performing tasks. Or by having a system perform actions only within specific (ethical) boundaries, formulated by the organisation using the systems.

A third consideration is that most AI does not work independently: it is part of a service or a product. And AI often works together with or advises people. For example, a web shop based on AI can customize product offerings for a visitor, determine the price, test if the information the visitor provides about address and payment data is reliable and predict when the package is probably delivered at home. Each of these forms of AI has different ethical and legal aspects. But ethical questions can also be asked about the entire web shop, such as: does the shop help visitors to make sustainable choices or does it focus on temptation and impulse purchases (or does it combine both principles). In that case, the Impact Assessment concerns the entire service and the individual components.

Transparency

Transparency about how an AI application works gives individuals the opportunity to appreciate the effects of the application on the freedom of action and the room to make decisions.

> Transparency means that actors have knowledge of the fact that AI is applied, how decision-making takes place and what consequences this may have for them.

In practice, this can mean various things. It may mean that there is access to the source code of an AI application, that end-users are involved to a certain extent in the design process of the application, or that an explanation is provided in general terms about the operation and the context of the AI application. Transparency about the use of AI applications may enlarge the individual's autonomy, because it gives the individual the opportunity to relate to, for instance, an automatically made decision.

When it comes to transparency, it is important to remember that services (but also products such as the self-driving car) are often made up of countless components. Some of these components can be called AI. Many of these components are not under the direct management of the organisation that offers the service: public bodies use each other's data; self-driving cars rely on data from road managers, other cars on the road and providers of navigation systems. Often these services will use data from a variety of data sources that change continuously. In many cases it is no longer clear which data played a role at the time of a decision. The question then is which knowledge and organisational measures are necessary to be able to take responsibility and to prevent undesirable consequences and repetition: sometimes algorithmic transparency can be important.

The starting point of the Impact Assessment is that with every deployment of AI, we look at what is required for transparency and what that means for the design of the technique, the organisation or the people working with the technology.

Part 1 - Background Artificial Intelligence Impact Assessment (AIIA)

An Artificial Intelligence Impact Assessment (hereafter: AIIA or Impact Assessment) is a structured method to:

1. Map the (public) benefits of an AI application.
2. Analyse the reliability, safety and transparency of AI applications.
3. Identify values and interests that are concerned by the deployment of AI.[3]
4. Identify and limit risks of the deployment of AI.
5. Account for the choices that have been made in the weighting of values and interests.

Conducting an AIIA results in an ethical and legally justifiable deployment of AI. By thinking at an early stage about the opportunities and risks, problems are prevented. This not only ensures that the deployment of AI is justified; it also helps to protect the reputation and investments of the user.[4]

There is no statutory obligation to conduct an AIIA. The AIIA is a self-regulating instrument with which an organisation comes to a socially responsible deployment of AI.

Ethical and legal assessment

The performance of an AIIA must result in an ethical and legally justifiable deployment of AI. For an AI application to be ethical and legally justifiable, two conditions must be met:

Is the deployment of AI reliable, safe and transparent?

Reliability, safety and transparency are required prerequisites for a safe use of AI. If an AI does not work properly or is unsafe, it will not be easy to justify its use (regardless of the concrete goal). So these are generic conditions an AI application always has to comply with.

Reliable
Reliability refers to the systematically correct operation of the system: does it work efficiently and are the results technically and statistically correct. In other words, does the AI application do what it has to do and are the outcomes of the system correct and is it possible to reconstruct where necessary how the AI has come to a decision?

Safe
Safety of AI plays a role at various levels. Above all, the AI must not pose an (unacceptable) danger to the environment. This is particularly the case when it comes to AI systems that are situated in the physical world (think, for example, of self-driving cars). In addition, an AI application, being an information processing system, must be safe itself (digital security).This means that the integrity, confidentiality and availability of the system and the data it uses must be guaranteed. This is not only to protect the operation of the AI application, but also to protect the rights of (end) users, such as the right to privacy and data protection.

Transparant

A third aspect is transparency and by extension the possibility to explain the actions of AI and to account for its use (to the outside world). The individual and/or the society must be able to get an understanding of how decisions are made and what the consequences are for social actors. This applies first of all to decision-making that has a substantial influence on the individual or society. Transparency does not necessarily imply that algorithms and data usage must be understood.

Is the application ethical and legitimate?

Reliability, safety and transparency are necessary preconditions for the ethical use of AI. But even if these preconditions are properly met, the use of AI is not ethical by definition. For example, the purpose for which AI is used may be illegal itself (e.g. discrimination). Other values or interests could outweigh the goal, or the way the goal is achieved is not ethical.

Purpose

The purpose of the AIIA is not to tell what is and is not allowed when deploying AI. It is first of all up to the users of AI to decide what they consider ethical and which values they pursue with an AI application. Obviously, this consideration must be in line with the social views on what is ethical and comply with laws and regulations in force. The “Artificial Intelligence Code of Conduct” in annex 1 offers a roadmap to develop the ethical framework.

Values

In society, values translate into standards, laws and rules. That is why the legal framework is the first concrete assessment frameworkto use to determine whether an AI application is ethical. This concerns laws and regulations, codes of conduct and ethical codes.

Context

The context is also relevant, for instance within a sector. E.g. in the context of health care,

the Law on the professions in individual health care, the Law on the medical treatment agreement and the Law on medical devices are relevant. In addition, numerous guidelines and codes of conduct apply. These laws, rules and codes of conduct form the framework in which the AI application must operate in any case.

Moral compass

However, legal does not necessarily mean that an application is also ethical. In case of more advanced forms and applications of AI, the legal framework will often not be clear or concrete yet. It is then up to the organisation to make choices based on its own

moral compass. The "Artificial Intelligence Code of Conduct" can help to define this compass (see annex 1).

The design stage

An AIIA is conducted at the beginning of a project in which AI techniques are applied. In this way, the ethical considerations can be included in the design of the application (*value based design or value aligned design*). This also makes sense in terms of costs and feasibility, because if a product has already been built or a project has already been executed, it is often impossible or very expensive to make any changes.

Involving Stakeholders

In addition to the internal stakeholders (the business or the policies, legal, compliance, IT, etc.), the involvement of the outside world is also relevant. The discussion with stakeholders (politics, government, civil society, science) and in particular end users who are affected by the deployment of AI (citizens, patients, consumers, employees, etc.) and their representatives is essential to gain support for the results of the AIIA.

Relation Privacy Impact Assessment (PIA)

The AIIA and the *Privacy Impact Assessment* (PIA), also called *Data Protection Impact Assessment* (DPIA) are both risk assessment tools and partly use the same logic. Both instruments are complementary, but not interchangeable. A PIA only focuses on the risks that processing of personal data may bring to the data subject (the person whose data are being processed). The AIIA is a broader instrument, which focuses on all possible ethical and legal issues that can be associated with the deployment of AI. Furthermore, the AIIA not only looks at risks, but also offers a framework for making ethical choices for the use of artificial intelligence. If a PIA has already been carried out within the framework of the application, it is strongly recommended to include the results in the AIIA.

Practical application AIIA and ethics

Ethics is a philosophical discipline that addresses the question of what doing the right thing means. Ethics does not offer a checklist with what is right and wrong; it is rather the method to assess what is right and wrong.

> Ethics as a discipline helps to approach and fathom a conflict, a problem or a dilemma, to weigh different solutions and to analyse outcomes on the basis of human and social values.

Ethics does not guarantee a flawless implementation. An ethical analysis can lift a discussion about the design or implementation of an AI system to a higher level and help to make the right choices (ethical use of AI).

The deployment of AI must be in line with the objectives and ethical guidelines of the organisation itself. The values that the organisation strives for (the relationship with customers, sustainability, diversity, etc.) must be reflected in the deployment of AI. Furthermore, the deployment of AI cannot be separated from its broader embedding within an organisation and the interaction between the employees and the deployment. Within the organisation, it is also necessary to make choices about control measures in order to achieve a reliable, safe and transparent deployment of AI.

Ethical lenses

Ethics has different reasoning methods. It is, as it were, the lens you use to look at a problem. It is important to be aware that there are different lenses, which can lead to different conclusions. The most typical ‘ethical lenses’ are:[5]

1. **Consequence ethics** (or consequentialism) emphasises the consequences of an action. An action is morally good if the result is positive. When a person in an emergency situation has to choose to kill one person so that ten people can survive, then the right choice is to kill this person.[6]

2. **Deontology** literally means the science of duties. Instead of focusing on the consequences of an action, the starting point is compliance with obligations. Doing the right thing means doing your duty. So the effect of fulfilling the duty is not relevant in terms of ethics. When someone finds it morally unacceptable to kill, it is the right choice for him or her not to kill the person, even if the result is that ten other persons cannot be saved and die.

3. **Virtue ethics** looks at actions from the perspective whether they are inspired by or contribute to a certain virtue.[7] What is virtuous, varies per actor. Whether it is a good choice to kill one person to save 10 people depends on what a virtuous person would do. The right choice is the choice that a virtuous person would make.

4. **Care ethics** Care ethics is focused on care for each other and building good relations. The emphasis is not on general principles but on the individual. Abstract ethical questions, for example what is good, are overlooking the individual, according to care ethicists (with the result that there is no morality). So the choice of killing someone to save others depends on what relationship you have with the individuals.

The ethical lenses offer you a starting point for analysing whether your deployment of AI is ethical and form as it were your 'moral compass'. What values do you put first and what is your starting point when using AI? Are you going for the greatest happiness for the largest group, or are you paying more attention to vulnerable groups? These lenses represent the main currents in ethics and are therefore sufficient for a practical approach to ethics in an AIIA.

Making choices clear

Social actors can look at the same ethical dilemma through different ethical lenses and therefore draw a different conclusion about what is 'ethical' in a given situation. By clarifying your choices and considerations and the lens you are looking through, you can enter into a dialogue with other social actors.

When using these lenses, keep in mind that one lens does not necessarily exclude the other. For example, choices can primarily be inspired by the expected results (consequentialism), but the action can nevertheless be restricted or controlled by certain principles (deontology).

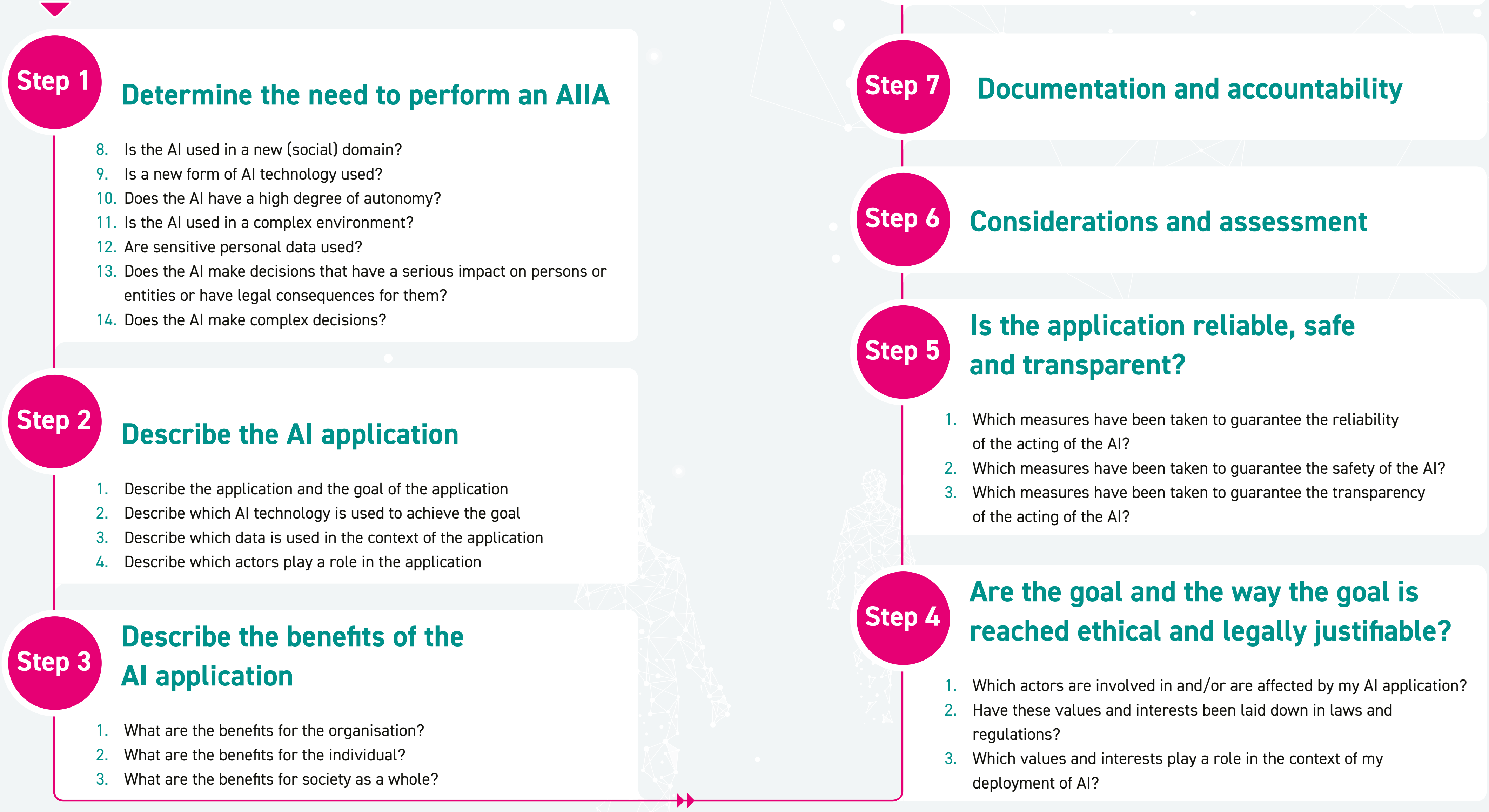
Part 2 - Conducting the AIIA

Organisations that want to implement an AIIA can follow the roadmap below:

1. Determine the need for conducting an AIIA.
2. Describe the application and context of the application.
3. Determine the benefits of the application.
4. Determine whether the purpose and the way in which AI is used are justified.
5. Determine whether the application is reliable, safe and transparent.
6. Document the results and considerations.
7. Evaluate periodically (create a feedback loop).

It is worthwhile to enter with each step into a dialogue with the outside world (representatives of end users, civil rights organisations, customer panels etc.), to test whether your assumptions and considerations are in line with the public views on ethics.

Roadmap for conducting the AIIA



Step 1 Determine the need to perform an AIIA

8. Is the AI used in a new (social) domain?
9. Is a new form of AI technology used?
10. Does the AI have a high degree of autonomy?
11. Is the AI used in a complex environment?
12. Are sensitive personal data used?
13. Does the AI make decisions that have a serious impact on persons or entities or have legal consequences for them?
14. Does the AI make complex decisions?

Step 2 Describe the AI application

1. Describe the application and the goal of the application
2. Describe which AI technology is used to achieve the goal
3. Describe which data is used in the context of the application
4. Describe which actors play a role in the application

Step 3 Describe the benefits of the AI application

1. What are the benefits for the organisation?
2. What are the benefits for the individual?
3. What are the benefits for society as a whole?

Step 4 Are the goal and the way the goal is reached ethical and legally justifiable?

1. Which actors are involved in and/or are affected by my AI application?
2. Have these values and interests been laid down in laws and regulations?
3. Which values and interests play a role in the context of my deployment of AI?

Step 5 Is the application reliable, safe and transparent?

1. Which measures have been taken to guarantee the reliability of the acting of the AI?
2. Which measures have been taken to guarantee the safety of the AI?
3. Which measures have been taken to guarantee the transparency of the acting of the AI?

Step 6 Considerations and assessment

Step 7 Documentation and accountability

Step 8 Review periodically

Figure 1. The roadmap for an AIIA

Overview

The figure below is an overview of the logic of an AIIA.

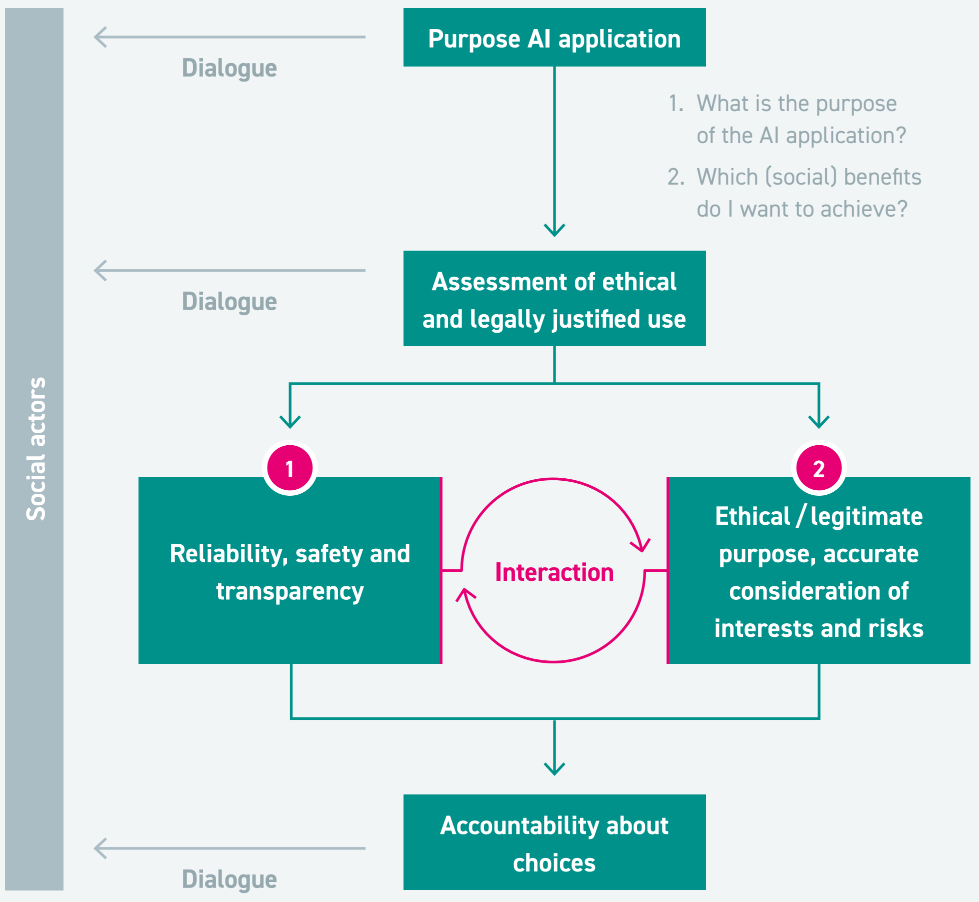

Social actors
Dialogue
Purpose AI application
1. What is the purpose of the AI application?
2. Which (social) benefits do I want to achieve?
Dialogue
Assessment of ethical and legally justified use
1
2
Reliability, safety and transparency
Interaction
Ethical / legitimate purpose, accurate consideration of interests and risks
Accountability about choices
Dialogue

Figure 2. The logic of an AIIA

Step 1 **Determine the need to perform an AIIA**

Not every deployment of AI justifies performing a complete AIIA. Only perform an AIIA if it is useful and necessary.The screening questions below are used to estimate whether an AIIA is necessary or desirable. If your answer to one of these questions is 'yes', then it would be a good idea to conduct an AIIA. If your answer to multiple questions is 'yes', then AIIA is highly recommended.[8]

The questions relate to the social and political context (questions 1 and 2), the characteristics of the technology (questions 3, 4 and 5) and the processes of which the technology is part (questions 6 to 9).

1. **Is the AI applied in a new (social) domain?**
Is the AI applied in a domain where it has not been used before? For example, an application that is used for the first time in healthcare while previously, it was only used for marketing purposes. Due to the change of domain, it is possible that the application will rise (new) ethical questions.

   When the application takes place in a sensitive social area, the risks and the ethical issues are potentially greater. Think of topics such as care, safety, the fight against terrorism or education. Think also of vulnerable groups such as children, minorities or the disabled.

   > Keep in mind that ethical dilemmas may also arise in seemingly innocent usage contexts.

   The “Artificial Intelligence Code of Conduct” (annex 1) and other sectoral or service-related and professional ethical codes can also help determine whether AI is applied in a sensitive area or topic.

2. **Is a new form of AI technology used?**
Risks of technology are usually greater when they are new and innovative than when they have been used and tested for a long time.

3. Does the AI have a high degree of autonomy?

The more an AI acts more independently and has more free room to make decisions, the more important it is to properly analyse the consequences of this autonomy. In addition to the room to make decisions, autonomy can also lie in the possibility of selecting data sources autonomously.

4. Is the AI used in a complex environment?

When the AI is situated in a complex environment, the risks are greater than when the AI is in a confined environment. The diversity of the input and the number of unexpected situations to which an AI must anticipate in an open environment is many times greater than in a confined environment, which can lead to unexpected or undesirable outcomes. For example, the use of an autonomous truck that drives in a closed container terminal has fewer risks than an autonomous truck driving on the public road.

5. Are sensitive personal data used?

If sensitive personal data are used in the deployment of AI, the risk is higher. Think for instance of medical data, data about ethnicity or sexual preferences.[9]

6. Does the AI make decisions that have a significant impact on persons or entities or that have legal consequences for them?

When the AI makes decisions automatically (without human intervention) and the decision can lead to someone experiencing legal consequences of that decision or being significantly affected otherwise, the risk is greater. Think of: not being able to get a mortgage, losing your job, a wrong medical diagnosis or reputational damage due to a certain categorisation.[10]

7. Does the AI make complex decisions?[11]

As the decision making by the AI is more complex (for example, more variables or probabilistic estimates based on profiles) the risks increase. Simple applications based on a limited number of choices and variables are less risky.

If the way in which an AI has come to its decisions can no longer be (fully) understood or traced back to people, then the risk of the acting or the decision is potentially greater. With complex neural networks, for example, it is not always possible to reason back how the AI came to the decision.

Step 2 **Describe the AI application**

The analysis starts by describing the goals that an organisation wants to achieve by applying AI. Which policy goal or commercial goal does the organisation pursue and how does the deployment of AI help to achieve this goal?

Without a clear description of the goal, it is impossible to assess whether the application is ethical.

1. Describe the application and the goal of the application

AI can be deployed in many forms, from relatively simple decision support systems to fully autonomous cars or even weapon systems. Therefore, describe the product, service, system or process in which the deployment of AI plays a role, the form in which AI will be deployed and the goal.

> In addition to the general description of the goal, it is also important to describe in more detail the 'room' the AI has and the values that are being pursued. To this end, the following questions must be answered:

1. Are the specific objectives of the deployment of the AI and the desired final state (goal state) sufficiently clearly defined?
2. How does the output of the AI contribute to achieving the goal?
3. Is the context in which the AI must achieve this goal sufficiently clear and delimited?
4. Is there a hierarchy of goals / interests?
5. What are the rules / constraints that the AI has to respect?
6. What is an acceptable tolerance / margin of error?

AI should have an understanding of ethical behaviour. This means that the AI 'understands' within the relevant context what is regarded as ethical behaviour by the user / society.[12]

What is ethical must therefore be made explicit and quantifiable as much as possible, so that the AI can seek an optimal solution to the problem based on the desired values and interests. This can be achieved by defining the desired goal state and possible rules and constraints to achieve this goal state. For more complex situations, this can also be the definition of 'target' or 'utility' functions. These purpose functions describe the utility of a particular state for an AI. The AI bases its choices on the consequences this has for the defined purpose functions, whereby it seeks maximum utility.

However, different purpose functions and the associated values and interests may conflict. It is therefore up to man, not only to make explicit what the purpose functions are, but also how they relate to each other.[13] The following (strongly simplified example) illustrates this:

The dilemma of the autonomous car

An autonomous car has been given the purpose function to get from point A to point B as quickly as possible. Given this function, the car will probably drive as fast as possible and will not take into account the safety of other road users, because this is not relevant to the assignment. If the same autonomous car has only been given the assignment to guarantee road safety, then the car will probably not leave, because the most secure option is not moving.

In the previous example, both purpose functions must therefore be combined to achieve an optimal result. To this end, it must be made explicit what road safety means in concrete terms and what the importance is in relation to achieving the other goal (going from A to B). If this is explicit (quantifiable), the AI can design an optimal strategy to achieve its goals.

Here too, ethical lenses play a role (zie page 27): is an AI designed to make choices that are consistent in nature or does the AI act deontologically? In other words, does the AI make decisions based on what yields the most for the defined value, or does the AI always act in accordance with specific ethical principles, even though the result may be less or even negative for the defined value? It is therefore again important to realise that one lens does not necessarily exclude the other.

2. Describe which AI technology is used to achieve the goal

Give a description of the AI technology or technologies used. This mainly concerns the features of the system, the input and output, the system's autonomy and how it effectively acts within the room that is given.

3. Describe which data are used in the context of the application

Describe the data sources that are used to have the AI make decisions (the input) and the origin of these sources. Think of the training data that are used to train an algorithm and the data that the system then uses to actually work

Include sensor data in the description of the data that the system uses as input. Also take into account the quality of the data and the nature of the data (e.g. synthetic data or real data).[14]

4. Describe which actors play a role in the application

Describe which actors play a role in or with the application, what their position is and what their expectations or wishes are (a stakeholder's analysis). This concerns in particular the actors in society with whom the application comes into contact. Think of citizens, other organisations and the government.

Step 3 **Describe the benefits of the AI application**

When AI is used to achieve a certain goal, it is with the idea of realizing benefits for the organisation, the individual and / or society as a whole. Benefits can be, for instance, freedom, well-being, prosperity, sustainability, inclusiveness and diversity, equality, efficiency and cost reduction.[15]

Describe in this step the benefits of using AI for the organisation, the individual and society as a whole. These benefits should be taken into account in the consideration of the ethical and legitimate deployment of the AI.

Benefits of the application are available at different levels and for different actors. For example, the organisation that applies the AI will first of all have to focus on realizing its own benefits (reducing costs, increasing profit, et cetera). In the case of the government, the benefits will often go hand in hand with social benefits (realizing policy objectives). In addition, there may be social benefits in addition to or complementary to the benefits for the government organisation. For example, the deployment of AI in the context of HR can ensure the selection of the best candidate (organisation benefits), but at the same time also prevent discrimination in the selection process (individual and social benefits).

1. **What are the benefits for the organisation?**
   How is the objective described in step 2 achieved and what advantages does this have compared to other methods (cost reduction, efficiency *et cetera*)? Also take into account at this point how the benefits to be realised relate to the standards and values of the organisation. To what extent does AI contribute to the goal and the way in which this is achieved and does this fit within the norms and values of the organisation? Does the application contribute to the organic objectives and is it in line with the ethical guidelines of the organisation?

2. **What are the benefits for the individual?**
   What benefits does the new application have for the individual? For example, is the deployment of AI safer, more objective or fairer than existing decision-making? Or does the use of AI enable a product or service for the individual that was not possible without AI?

3. **What are the benefits for society as a whole?**
   A deployment of AI may also have social benefits. Ask the following questions to map the social benefits:

   1. Which social interest is served with the deployment of AI?
   2. How does the project / system contribute to or increase well-being?
   3. How will the project / system contribute to human values?

Step 4 Are the goal and the way the goal is reached ethical and legally justifiable?

In this step, you determine whether the goal and, more specifically, the manner in which this goal is achieved is ethical and legally justifiable.

The starting point for your analysis is the existing legal framework. But this framework can be incomplete or inadequate for a good ethical assessment. That is why you identify the values and interests that are at stake in the deployment of AI. In particular, you look at the possible risks of your application. Identifying these risks is important, because it makes you see what you could improve in the design and the deployment of AI. The choices you make (are we going to exclude or limit risks, how much residual risk do we accept, do we accept that our application creates risks?) are the ethical trade-offs of the organisation. The ethical lens used to look at the application plays an important role here.

In order to assess whether the use of AI is ethical, you must determine which values and interests may be at stake in your deployment of AI. To this end, you can ask yourself the following questions:

1. Which actors are involved in and/or are affected by my AI application?

Values (honesty, equality, freedom) are ideals and motives that a society and the actors within it strive for. In Annex 1 you find the “Artificial Intelligence Code of Conduct” with the ethical principles of the European Group on Ethics in Science and New Technologies, which can provide guidelines for the analysis of relevant values. Because values are abstract, it is often difficult to assess whether the deployment of AI is in line with the values within a society. In general, acting in violation of values means that the interests of the actors are directly or indirectly harmed (see figure 3). For example, impairing the value 'equality' can mean that a person or group is discriminated against. That is why translating values into interests can give direction to the assessment whether an AI application is ethical or not.

Social actors have different interests. Through AI, existing power relations can change and the interests of actors can be harmed or strengthened. AI applications can therefore affect interests at different levels. For example, the deployment of AI can very specifically affect the interests of an individual (for example, a violation of his / her privacy), but the deployment of AI can also influence interests and relationships at the level of society. Think, for example, of changes in employment through the deployment of AI. In this AIIA, the emphasis is on the interests of the individual. These correspond to a large extent to the traditional fundamental rights (right to freedom of expression, privacy et cetera) and the social fundamental rights (right to education, employment, et cetera).

2. Have these values and interests been laid down in laws and regulations?

Standards, values and ethical principles within a society are (partly) crystallized in laws and codes of conduct. The goal of these rules is to promote well-being, to protect (human) rights and to organise society.

These laws and regulations form the concrete framework within which your application must remain. Insofar as the frameworks are unclear or incomplete, the design of your application must be in line with the values that apply in society. In so far as your application affects the interests of third parties, you must be able to substantiate why this is justified.

Ethically Aligned Design

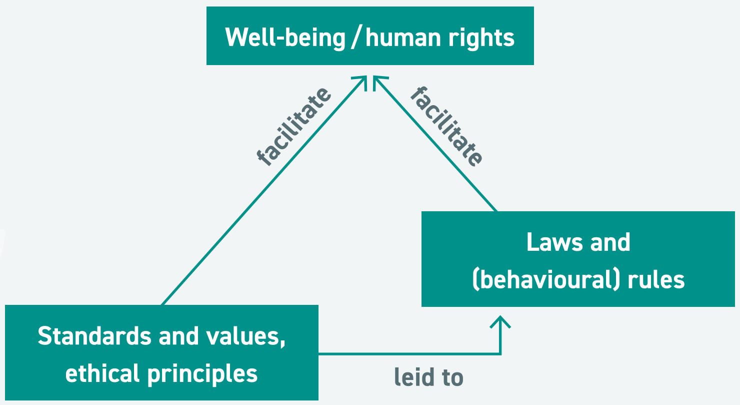

Well-being / human rights
facilitate
facilitate
Laws and (behavioural) rules
Standards and values, ethical principles
leid to

Figure 3. The relationship between norms and values, laws and regulations, well-being/human rights (IEEE, 2017)

3. Which values and interests play a role in the context of my deployment of AI?

There is a lively debate on values these days and many parties are developing codes of conduct and standards frameworks. The Impact Assessment is based on these frameworks, without choosing a single one: Annex I contains a code of conduct (from the European Group on Ethics in Science and New Technology) and rules of practice (an update of the rules of practice of the guide Autonomous Systems, published by ECP in 2006), which can help to describe values and interests that may be affected by the deployment of AI.[16] A number of questions (as examples) have been formulated for each of the values and interests that help to orientate thoughts about the ethical aspects and risks associated with the value in question.

Step 5 Is the application reliable, safe and transparent?

Reliability, safety and transparency are required prerequisites for a safe use of AI. Many of the risks of AI stem from inadequate reliability, safety or transparency of AI. The questions in this step help avoid common pitfalls in the deployment of AI and ensure a responsible application.

This step is not only about the reliability, safety and transparency of the AI application itself, but also about the broader embedding of the AI and these preconditions within the organisation. In other words, it concerns the entire system of organisational and technical control measures that ensure that an AI is deployed reliably, safely and transparently by an organisation.

1. Which measures have been taken to guarantee the reliability of the acting of the AI?

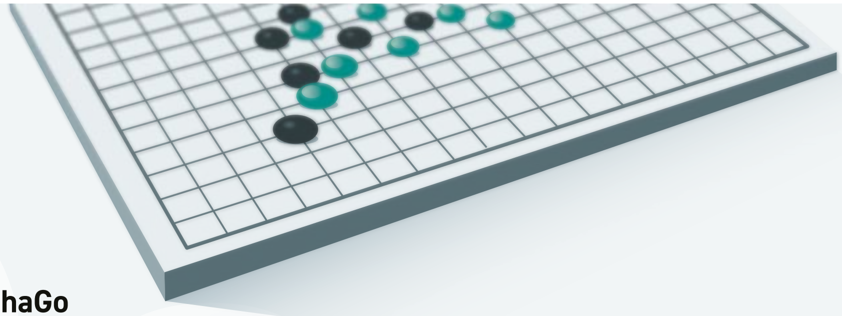
The first precondition is that an AI application is reliable. In short, this means that given the purpose function of the system, the system consistently makes the right decisions. To be able to determine the reliability, at least the following points must be taken into account.

Are there clear criteria / parameters for the correct functioning of the AI?
Based on the purpose target description from step 3, clear parameters must be defined for the correct functioning of the AI. What is the purpose of the AI? How should the goal be achieved? What are possible constraints for the actions? On the basis of these parameters, it must be tested whether the AI acts in line with the set parameters. When determining the parameters, the values and interests as described in step 5 must also be taken into account.

Does the AI act consistently?
The action of an AI must be consistent. This means that in comparable situations, the AI should not suddenly produce totally different outcomes. In order to determine whether an AI acts consistently, it must be tested

on the basis of the set parameters whether the AI acts consistently. With learning AI systems, it must be taken into account that the actions can change over time.

How is dealt with the unpredictability of the AI's actions?
Given the complexity of decision making by AI, there is an increasing number of situations in which it is not (completely) clear and / or reproducible why an AI has made a certain decision.



AlphaGo

A well-known example is the AI AlphaGo who made a completely unpredictable and inimitable move in round 37 of a game of Go against the human world champion Lee Sedol, which ultimately resulted in the win.[17] This is not a problem within the context of a defined and limited game with clear rules. But if an AI is situated in the physical world, where the complexity is endlessly larger so that 'correct' action is more difficult to define, unpredictable behaviour is potentially risky.

The fact that AI is now too complex to fully understand should not be an excuse for the uncontrolled introduction of an AI in a 'live' environment where it has to make important or risky decisions. With the deployment of AI, it must be determined how to deal with the unpredictability of the system and to what extent failure to reconstruct results is problematic for the deployment of the AI. It must also be indicated how reacted to unexpected outcomes and what mitigation measures there are to limit the negative consequences. It is important here whether it concerns decision making that people must be able to understand in real time, or whether it is sufficient that the process can be reconstructed if necessary (for example for a judicial review).

Are the correct data available and chosen for the deployment of the AI?
AI depends on the data used for correct operation. It is therefore important to assess whether the correct data are used and how these data are offered by the AI. This applies both to the phase in which the AI learns and is trained, and in the phase of actual application.

Insofar as the system is trained using training data, it must be assessed whether these data accurately reflect the actual environment and the problem area in which the AI will operate. In particular, account must be taken of teaching wrong behaviour through the selection and use of data (e.g. sample selection bias). This is to prevent things like bias and discrimination.

> When applied in a 'live' environment, measures must be taken to ensure the availability of the correct sources and the integrity of these sources.

Have the right algorithms been chosen that enable artificial intelligence to act effectively and achieve the goal?
In order to guarantee the correct operation of the system, the correct components must be used, in particular the algorithms used for decision making.

Which methods are used to verify whether the AI remains within the set parameters?
In order to be able to check whether the AI is acting correctly and is reliable, the functioning of an AI must be tested. This involves technical and organisational measures to check afterwards whether the choices of the AI have led to the correct result and that there is no adverse effect on the individual or society. The testing of an AI application plays an important role in this.

2. Which measures have been taken to guarantee the safety of the AI?

An AI must not pose a danger to its environment. In the deployment of AI, therefore, the safety aspects of the application must be taken into account. This is particularly important when the AI system is situated in the physical world and can also cause physical damage there.

> A large part of the risk-limiting measures aimed at ensuring reliability of the actions of the AI will also be relevant to guarantee safety.

Which safety measures must be taken (also) depends on the risks identified in step 5.

In addition to the fact that an AI must operate safely in a certain environment, the (digital) safety of the AI itself is also important. This concerns in particular the information security and the associated interests of confidentiality, integrity and availability.

Which measures have been taken to guarantee the safety of the AI?
In many cases, the data processed by an AI must remain confidential. Measures taken to ensure confidentiality are aimed at preventing unauthorized access to the data that an AI processes in order to be able to function.

Which measures have been taken to guarantee the integrity of the AI?
For the AI to work properly and to protect the rights and freedoms of third parties, the data processed by the AI must be protected against manipulation and damage. Measures taken to ensure integrity are aimed at preventing unauthorized access and adaptation of the data that an AI processes in order to work properly.

Which measures have been taken to guarantee the availability of the AI?
For the AI to work properly to protect the rights and freedoms of third parties, the data processed by the AI must be and remain available. Without data, the AI cannot function (properly). The same applies to models and algorithms used. Measures taken to ensure availability are aimed at removing or reducing threats that make data unavailable (e.g. taking measures to stop DDoS attacks).

3. Which measures have been taken to guarantee the transparency of the acting of the AI?

Transparency can constitute an important contribution to the legitimate and ethical deployment of AI. This concerns the public nature of the use and functioning of AI. Transparency can be offered at different levels (openness of use and operation, insight into the consequences and the possibility of accountability). The degree of transparency is partly dependent on the potential impact that the application has on the end user. As the size increases, a higher degree of transparency is appropriate.

To what extent is AI transparent?
The organisation must assess to what extent it is transparent about the deployment of AI. The purpose of transparency is to explain the use and operation of AI. For example, the operation of AI can be checked ex post (for example during an audit or after an incident). The most complete form of transparency is the publication of all algorithms, the datasets used and the results. This way, everyone can verify whether the AI application is correct. However, this form of transparency requires technical knowledge and does not necessarily provide insight into the use and the operation of the AI application.[18] In addition, however, there may be considerations for organisations to keep their algorithms and datasets secret. In addition to commercial considerations such as not wanting to disclose intellectual property, other issues such as protecting the privacy of individuals or national security may also play a role.

In addition to full transparency, more limited forms of transparency can also be envisaged, such as can be found in Article 13 or Article 22 of the General Data Protection Regulation. Article 13 states that when there is 'automated decision-making without human intervention', the logic of decision-making and the possible consequences thereof for the person concerned must be clearly communicated.

Is it possible to have insight into the consequences of AI?
In addition to the existence and operation of the AI, it is also relevant to provide insight into the consequences of the deployment of the AI. This may involve providing individuals with insight into how decision making by an AI influences their (legal) position, but also the broader social consequences of the deployment of AI, for example on issues such as employment.

Which control measures are applied?
For a careful deployment of AI, technical and organisational (management) measures must be taken. An organisation must be able to account externally and, where relevant, externally (*accountability*) about these control measures. This can be done, for example, through audits.

To test the measures taken and the deployment of AI in general, an 'algorithmic audit' is recommended. This audit, carried out by an independent third party, is to check the use of the algorithms and data.

Step 6 **Considerations and assessment**

In assessing whether the AI application is ethical, the benefits (step 3) and the identified risks (step 4) must be considered jointly. At least the following elements must be kept in mind.

Is the application proportional?
In order to be able to form an opinion on the legitimacy of the application, it must be assessed whether the deployment of AI is proportional. The question that has to be asked is: what is the goal that is being pursued with the AI application and how does this goal relate to the impact of the AI application on the individual and / or society as a whole? Although the end does not sanctify the means, in general a more legitimate purpose will be more permissible.

Can the same goal be achieved with less drastic means?
Subsidiarity must be assessed together with the proportionality of the application. This means that there should not be any less drastic way to achieve the same goal. In the context of AI, the need for data processing, the degree of autonomy and the complexity of the AI must be considered.

Is it about positive sum instead of zero sum?
It is important that the assessment of whether an application is legitimate / ethical is not a *zero sum game*. This means that it should not be purely a choice for one interest that outweighs the other interest. All values and interests must be optimally served by an application. Only where interests can no longer be united with each other or are at the expense of each other, it must be determined which interest outweighs the others.

Are there any residual risks?
When assessing whether an application is legitimate, even with a *positive sum* approach, it will be necessary to consider which risk is acceptable. Determine whether the risks that you have identified have been or are eliminated through risk-limiting measures or that there is still residual risk.

Where residual risk persists, it must be substantiated why the residual risk is accepted and what measures are taken to limit and repair damage if the risk manifests itself.

How is responsibility taken for further use?
A final point that must be considered is the responsibility for further use (*downstream responsibility*). AI applications do not work in isolation, but are linked to many other systems and processes. Account should be taken of how data, decisions and observations generated by AI can work in other systems and be used by other actors.

Step 7 **Documentation and accountability**

Record the results of steps 2 to 5. Pay particular attention to justifying the legitimacy of the application. A good record provides direction for the actual construction and layout of the application, but also enables you to enter into the social discussion and, where necessary, to be accountable for your choices.

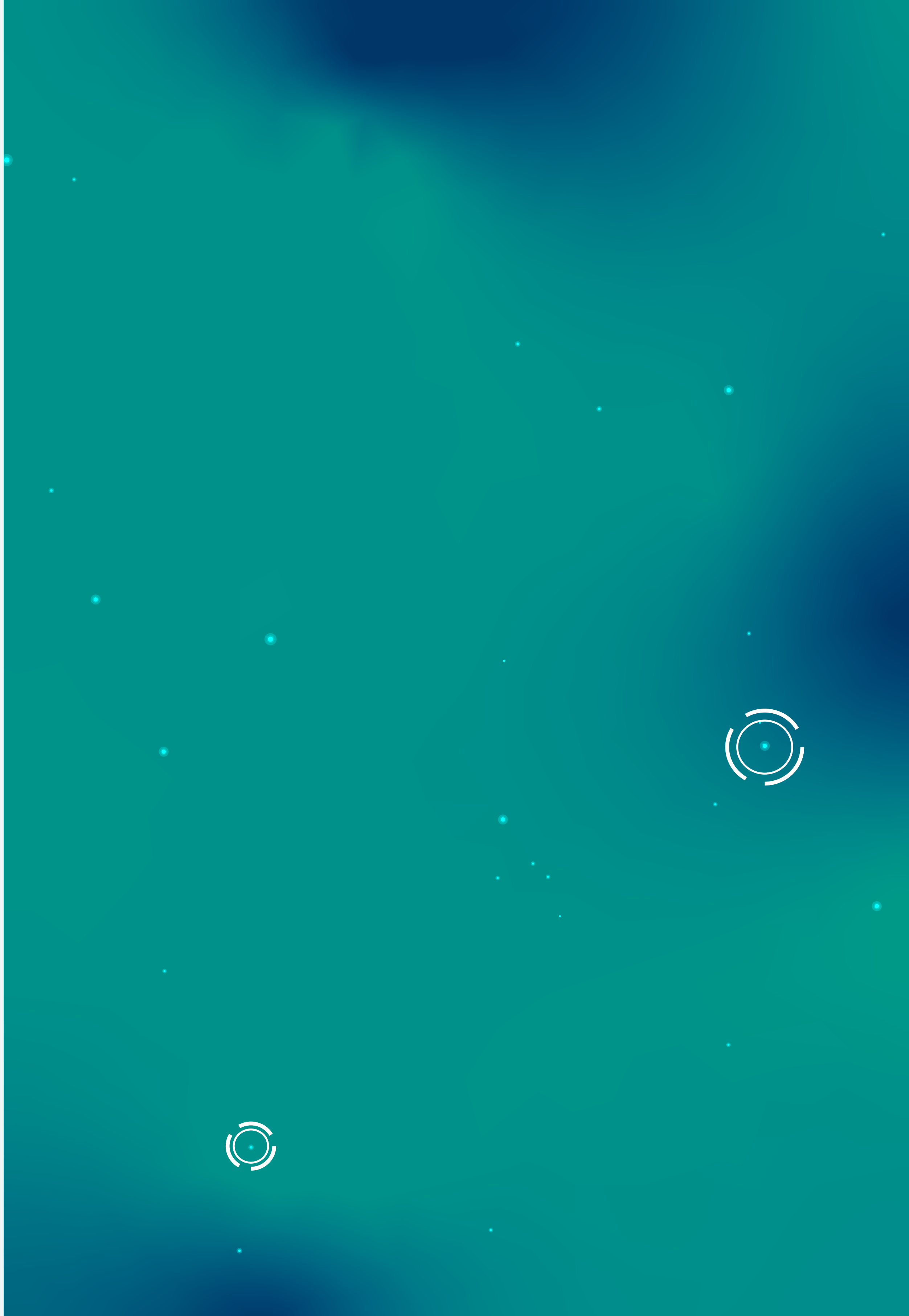
Whether an AI application is ethical and legally justifiable depends on the person or organisation that makes the judgment and the ethical lens used to look at the application. The opinion of the organisation applying AI may deviate from the opinion of that of other social actors and / or society as a whole. By performing an AIIA, you can substantiate and justify your choices and considerations and enter into a structured discussion.

Step 8 **Assess periodically**

Assessing whether an AI application is ethical is not a one-off process. The organisation and the outside world are changing. This may influence the ethical and legal frameworks of the AI application and thus the legitimacy. That is why it is important to periodically evaluate whether the application is still justified. Especially for learning AI it is important to follow how the AI develops and influences the environment. By periodically evaluating, new risks are discovered in time, but also a *feedback loop* can be created that makes the deployment of AI better and more effective.

An evaluation can take place with a certain time interval (for example every year), but it is wise to also define situations that require a reassessment. Think of:

1. The AI application is used for a purpose other than that for which it was originally intended;
2. The decision-making room of the AI is extended or modified in some other way;
3. New data sources are being used;
4. Existing data sources are modified or no longer used.



Bibliography

Advisory Council on International Affairs (2017), *The will of the people? Erosion of the democratic constitutional state in Europe*

Committee on Technology National Science and Technology Council. (2016). *Preparing for the Future of Artificial Intelligence. CreateSpace* Independent Publishing Platform. Retrieved 05 01, 2018, from https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF

Eck, M. v. (2018). *Geautomatiseerde ketenbesluiten & legal protection: Een onderzoek naar de praktijk.* Retrieved 5 1, 2018, from https://pure.uvt. nl/portal/files/20399771/Van_Eck_Geautomatiseerde_ketenbesluiten.pdf

ECP. (2018). *Artificial Intelligence. Gespreksstof en handvatten voor een evenwichtige inbedding in de samenleving.* Retrieved from http://www.ecp.nl/AI

ECP. (2018). *Het verhaal van digitaal. Samen vormgeven aan onze digitale samenleving.* Retrieved from http://ecp.nl/publicaties/het-verhaal-van-digitaal

ECP.NL. (2006, May 15). *Handreiking voor gedragsregels Autonome Systemen. Juridische aandachtspunten voor de bouw en het gebruik van autonome systemen.* Leidschendam: ECP.NL. Retrieved May 1, 2018, from ecp.nl: https://ecp.nl/wp-content/uploads/2017/04/Handreiking-voor-gedragsregels-autonomous-systems-2006.pdf

European group on ethics in science and new technologies. (2018). *Statement on artificial intelligence, robotics and autonomous systems.* Brussels: European Commission. Retrieved 05 01, 2018, from https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf

Future of Life. (2018, 05 01). *AI principles.* Retrieved from Future of Life: https://futureoflife.org/ai-principles/

ISACA. (2016). *Cisa Review Manual,* 26th edition.

Kiran, A., Oudshoorn, N., & Verbeek, P. (2015). *Beyond checklists: toward an ethical-constructive technology assessment.* Journal of Responsible Innovation, 2(1), 5-19.

KNMG. (2013). *Gedragsregels voor artsen*, version 3.1.

Kool, L., Timmer, J., Royakkers, L., & Est, R. v. (2017). *Opwaarderen - Borgen van publieke waarden in de digitale samenleving.* The Hague: Rathenau Instituut.

Motivaction. (n.d.). *Burgerschapsstijlen.* Retrieved 01-05-2018, from https://www.motivaction.nl/onderzoeksmethoden/burgerschapsstijlen

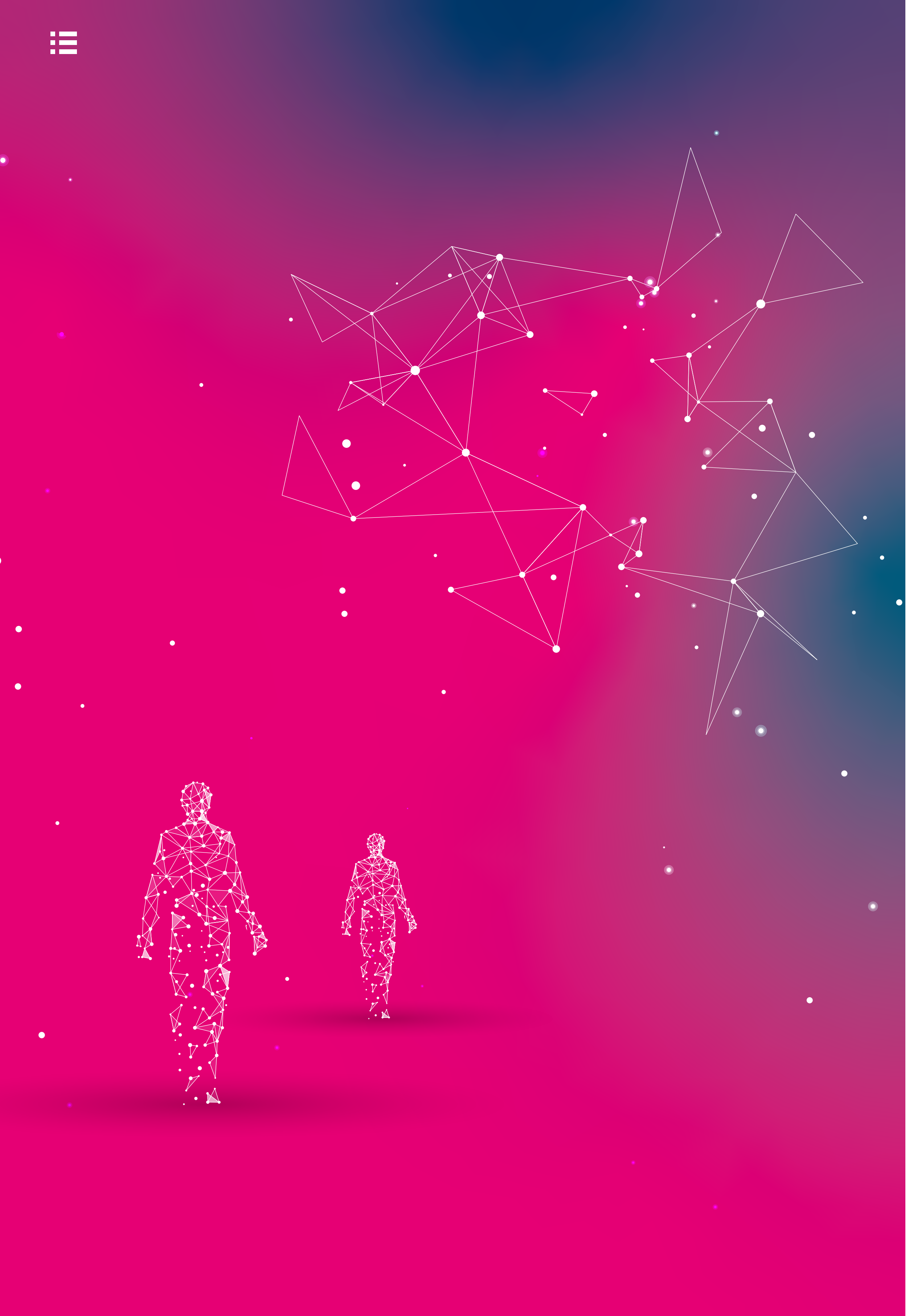
OECD. (2018, 05 09). *measuring-well-being-and-progress.htm.* Retrieved from oecd.org: http://www.oecd.org/statistics/measuring-well-being-and-progress.htm

Tewari, W. (2011). *A structured approach to IT auditing: model based development of audit terms of reference.* Amsterdam: VU University Press.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.* Version 2. IEEE. Retrieved from http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

Wildlak, A., & Peeters, R. (2018). *De digitale kooi. (On)behoorlijk bestuur door informatiearchitectuur of: hoe we de burger weer centraal zetten in een digitaliserende overheid.* Boom.

WRR. (2011). O*verheid. WRR-Rapport nr. 86.* Den Haag: Scientific Council for Government Policy.

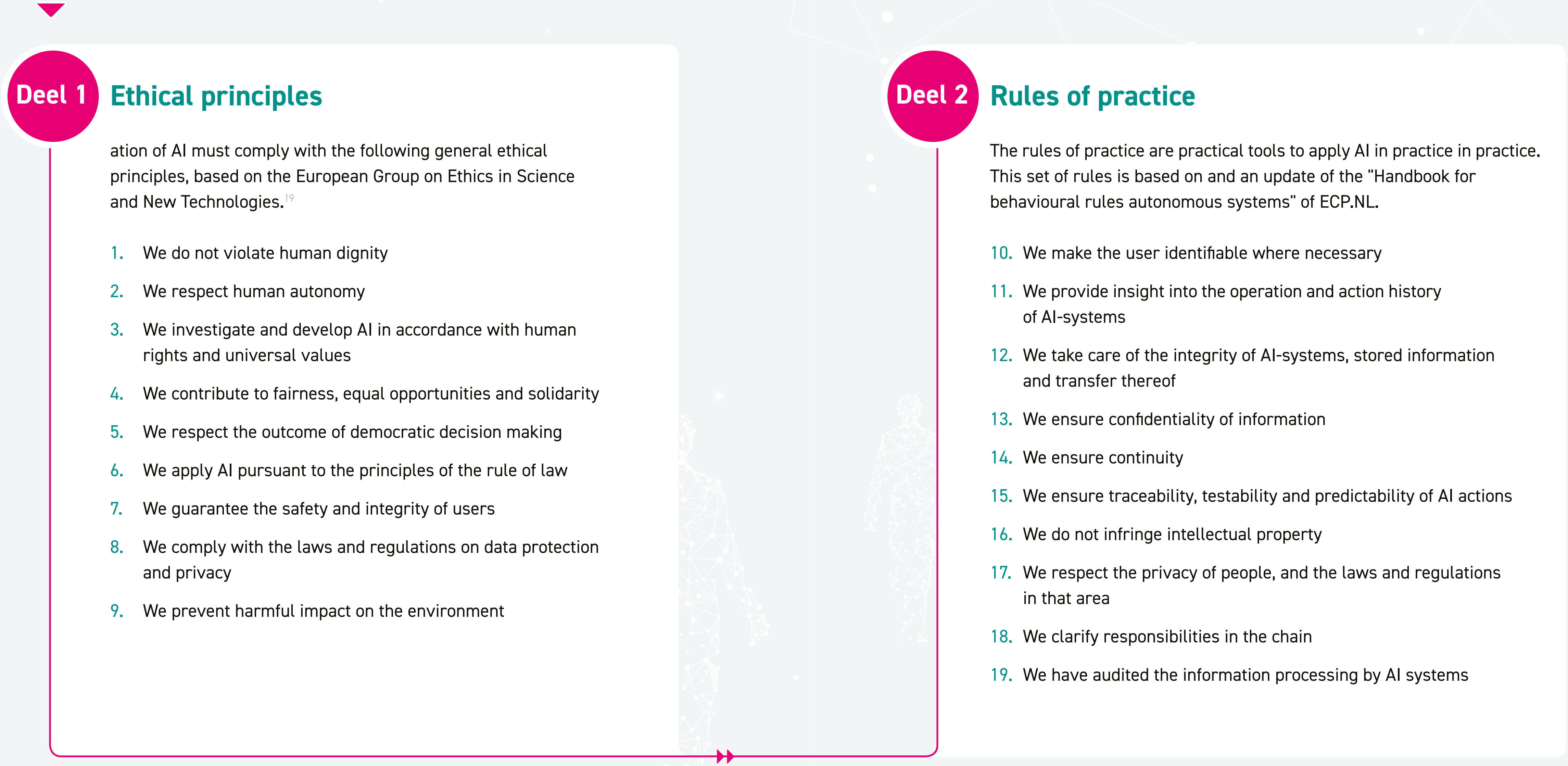


Annex 1 - Artificial Intelligence Code of Conduct

The Artificial Intelligence Code of Conduct offers a guideline for establishing the standards framework against which a concrete AI application is tested when conducting an Artificial Intelligence Impact Assessment (AIIA). This guide is generic in terms of the nature and context of the application. In a way, the Code of Conduct is also a snapshot. The debate about the frameworks within which AI is developed and applied is very dynamic and has a broad spectrum of opinions and visions. It is expected that further steps will be taken in the near future to come to European and, if possible, international frameworks for the development and deployment of AI. If further results are achieved in that process, it is obvious to adhere to this code of conduct.

Artificial Intelligence Code of Conduct

The Artificial Intelligence Code of Conduct is an integral part of the Artificial Intelligence Impact Assessment (AIIA). This set of rules is the fundament under the AIIA.



Deel 1 Ethical principles

ation of AI must comply with the following general ethical principles, based on the European Group on Ethics in Science and New Technologies.[19]

1. We do not violate human dignity
2. We respect human autonomy
3. We investigate and develop AI in accordance with human rights and universal values
4. We contribute to fairness, equal opportunities and solidarity
5. We respect the outcome of democratic decision making
6. We apply AI pursuant to the principles of the rule of law
7. We guarantee the safety and integrity of users
8. We comply with the laws and regulations on data protection and privacy
9. We prevent harmful impact on the environment

Deel 2 Rules of practice

The rules of practice are practical tools to apply AI in practice in practice. This set of rules is based on and an update of the "Handbook for behavioural rules autonomous systems" of ECP.NL.

10. We make the user identifiable where necessary
11. We provide insight into the operation and action history of AI-systems
12. We take care of the integrity of AI-systems, stored information and transfer thereof
13. We ensure confidentiality of information
14. We ensure continuity
15. We ensure traceability, testability and predictability of AI actions
16. We do not infringe intellectual property
17. We respect the privacy of people, and the laws and regulations in that area
18. We clarify responsibilities in the chain
19. We have audited the information processing by AI systems

Figure 4. Artificial Intelligence Code of Conduct

Terminology

When the AIIA refers to the 'user', we refer to the organisation that uses AI. This can also be the employee who works with AI in an organisation. When the assessment speaks about the 'individual' or the 'end user' we refer to the natural person who uses the AI of an organisation (for example the driver of an autonomous car) or is subject to the decision-making of the AI (for example an applicant assessed by an AI). By 'stakeholders' the assessment means all individuals and parties having an interest in AI application and experience direct or indirect consequences of AI and subsequent decision-making. 'Builders and providers' are the parties that develop AI systems. Many AI applications are offered via the cloud.

Parts

The code of conduct consists of two parts:

1. Ethical principles and democratic preconditions as formulated by the European Group on Ethics in Science and New Technologies;
2. Rules of practice for dealing with AI applications.

Part 1
Ethical principles

Application of AI must comply with the following general ethical principles, based on the European Group on Ethics in Science and New Technologies.[19]

These nine basic principles and democratic preconditions, published on EU initiative, are a first step towards establishing a global ethical framework. The principles are laid down in the "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems" of the European Group on Ethics in Science and New Technologies. These principles are based on the fundamental values laid down in the EU treaties and the Charter of Fundamental Rights of the European Union.

1. **Human dignity: AI must not infringe on human dignity20**[20]
   Every person has a self-contained and intrinsic value as a person that cannot be compromised. Humiliation, dehumanisation, instrumentalisation and objectification (using people as an instrument for a goal, without seeing them as an end in themselves) and other forms of inhumane treatment harm this dignity. In AI applications, consideration must be given to human dignity and the way in which a proposed application affects this dignity. Respect for human dignity means above all that the application must be in line with human rights. In addition, where necessary, it must be made clear to the individual that it interacts with an AI application. Respect for human dignity can also force the abandonment of the deployment of an AI application because a human intervention or interaction is more appropriate.

- To what extent is human deliberation replaced by automated systems?
- Can people take over the automated decision-making process?
- Is there a strong incentive for people to follow the automated decisions?
- Individuals who come into contact with the AI application are they aware of this?
- Are people objectified and possibly dehumanized by the deployment of the system?

2. Autonomy: AI must respect human autonomy[21]

Autonomy is the ability of an individual to act and decide independently. AI applications can restrict people's freedom of action and decision-making space. It also enables actors to influence people unconsciously (*nudging*) or even to manipulate them. Paternalism is a specific form of limiting the autonomy of the individual from the point of view of protection. The idea is that the organisation (or the algorithm) is better at decision-making, because it makes better choices than the individual. So an AI application can limit (or increase) autonomy for the individual, both consciously and unconsciously.

Transparency about the operation of an AI application gives individuals the opportunity to appreciate the effects of the application on the freedom of action and decision-making space. Transparency means that actors have knowledge of the fact that AI is applied, how decision-making is achieved and what consequences this may have for them. In practice, this can mean various things. It may mean that there is access to the source code of an AI application, that end users are involved to a certain extent in the design process of the application, or that explanations are given in broad terms about the operation and the context of the AI application. Transparency about the use of AI applications may enlarge the individual's autonomy, because it gives the individual the opportunity to relate to, for instance, an automatically made decision.

- To what extent is there access to the source code of the AI application (openness of algorithms) and is this knowledge usable for outsiders?
- To what extent can the operation of the application / the algorithm be explained to end users and those involved?
- Is clear to end users (and other relevant actors) what the consequences are of decision making by the AI?
- Can the used datasets be made public?
- Can the sources of used data be made public?
- Can the organisation be transparent in a different way for users and stakeholders?
- Does the domain in which the AI application is used demand a higher degree of transparency for users and those involved (e.g. care or justice)?
- To what extent does the organisation or AI application take decisions about or for the individual?
- Has a balance been found between the benefits of the goal and the freedom or the individual?
- Is there a time when the individual can influence decision making by the AI? Should this functionality be made available?
- To what extent does the AI direct the user in a direction desired by the organisation (*nudging*)?
- To what extent can an individual withdraw from (unconscious) influence?

3. Responsibility: the principle of responsibility must underlie every research and every deployment of AI[22]

Responsibility means that AI applications are only developed in accordance with human rights and other universal values. This means that during the entire process an AI application must have an ongoing view on (research) ethics and individual and the effects that the deployment of AI has on the individual and society. Because the negative effects of AI applications are potentially large, risk awareness and well-considered application are important.

- Which technical and organisational measures have been taken to prevent or limit any negative effects of AI (risk reduction)?
- How can any unforeseen effects be mitigated after deployment of the AI application?
- Is it clear who the legal controller rests for using the AI application?
- Can the organisation account for the application? (*accountability*)?

4. Fairness, equal access and solidarity: AI must contribute to fairness, equal opportunities and solidarity[23]

Fairness has various definitions. Fairness can mean that people get what they earn according to relevant criteria. Fairness can also mean that equal cases are treated equally (equality).Fairness can also refer to the concept of social equality, the idea that the weaker should be given priority over those who benefit from institutions that produce inequality. When using AI, the user must assess whether the deployment of AI and the decisions that are taken lead to just results. It should be kept in mind that information systems are never entirely value-neutral. In the design of the system, (implicit) choices for certain values are often decided (e.g. efficiency versus accuracy). Applications of AI can exhibit unwanted bias when the system design does not take conscious or unconscious bias into account (think, for example, of a bias in the selection of data with which an AI is trained). This can not only lead to incorrect or discriminating decisions, but also, for example, that groups, behaviour or information deviate from the prevailing norm (or the standard of the developers / users).

It is important to examine what effects the AI application has, in addition to the fairness of individual decisions, on more abstract norms such as legal certainty, equal opportunities and equal access.

- What values has the organisation decided to promote, and how?
- Are there specific groups that are favoured or disadvantaged in the context where the AI application is used?
- What is the possible harmful effect of uncertainty and error margins for different groups?
- Which choices are implicitly made in the architecture of the system? Have these choices been made by the organisation that will use the AI, or by the developer?
- Does the AI application take less biased decisions than the human decision-making process?
- To what extent is the AI application a continuation of human bias?
- Are prevailing images and stereotypes reinforced by the application or AI?
- Are values such as inclusiveness and diversity actively included as functional requirements for the AI application?

5. 5. Democracy: AI must respect the results of democratic decision-making[24]

A democratic constitutional state has an electoral dimension and a constitutional dimension. The electoral dimension includes aspects such as free and fair elections, a pluriform supply of parties and space for debate and consultation. The constitutional dimension includes aspects such as equality before the law, the right to redress, legal certainty, protection of civil liberties, a free and pluralist press.[25] As the scandal with Cambridge Analytica has made clear, the deployment of AI can influence the election process.[26]

In particular, governments in the deployment of AI should take into account the impact that this application has on the democratic constitutional state, especially where the constitutional dimension is concerned. The democratic dimension may also be relevant in applications that are further from the rule of law, because democratic values such as diversity, moral pluralism and equal access to information can be affected.

- To what extent does the AI-application undermine the principles of democracy, for example because the technology enforces policy without public deliberation?
- To what extent does the deployment of AI influence legal certainty and civil liberties? Is this influence clear to end users, stakeholders, and (popular) representatives?
- To what extent does the AI application affect free speech and the forum for public debate?
- To what extent does the AI application influence democratic values such as moral pluralism and diversity?
- To what extent does the AI application filter information from or for the user (curation)?
- To what extent does AI block access to information?
- What are the criteria on the basis of which information is filtered, blocked and curated?
- Does the AI have a bias regarding the information to be filtered?

6. Rule of law, accountability and liability: applications of AI must comply with and submit to the principles of the rule of law[27]

7. Safety, physical and mental integrity: AI systems must respect the safety and integrity of users[28]

Safety in the context of AI applications is about more than the physical safety for the user or the environment in which the AI application is used.[29] Internal safety and reliability (*cybersecurity*) and emotional safety in human-machine interaction must also be guaranteed. Special attention should be paid to vulnerable groups that may come into contact with the AI application.

- What is the effect on the physical safety of the users and environment of the AI application?
- To what extent is the cyber security of the application guaranteed?
- What effect does the AI application have on the emotional safety of users and stakeholders?
- Which vulnerable groups can come into contact with the AI application? How has it been ensured that these groups do not suffer any adverse effects from the application?

8. Protection of data and privacy: AI must comply with the laws and regulations regarding data protection and privacy[30]

The right to privacy is the right to the protection of privacy. What privacy means in practice strongly depends on the context. In the case of AI applications, the informational dimension of privacy in particular plays a role (the right to protection of personal data). Specifically, it can be linked to the principles and rules of the General Data Protection Regulation.

- Has the organisation determined how the privacy of those involved is protected?
- Does the application only collect and process the data necessary for the application?
- Are end-users capable of determining which data of / about them are collected and which conclusions are drawn from them?
- Can the user delete his data from the system?

9. Sustainability: AI must not have a harmful effect on the environment.[31]

AI applications, like other technologies, have an impact on the quality of life of our planet and the future prosperity of humanity and the living environment for future generations. AI has a direct influence on the living environment (think of increasing or reducing energy consumption and e-waste) and an indirect influence, for example by stimulating environmentally conscious behaviour (for example through decision support or nudging.

- What are the environmental effects of the AI application?
- Does the use of AI increase or decrease the use of raw materials and natural resources?
- What influence does the AI application have on the life of future generations?

Part 2
Rules of practice

This set of rules is an update of the "Handbook for behavioural rules autonomous systems" (2006) of ECP.NL. Both components of the code of conduct (ethical principles and practice rules) each have their own value and function. The ethical principles offer a broad framework for AI at a somewhat higher level of abstraction. The practical rules are generally a bit more concrete. However, they have not been designed as a (conclusive) elaboration of the aforementioned ethical principles, but do well with them and provide direction for the deployment of AI in practice.

1. Identification

Where necessary, the user of an AI system must be identifiable. It must be possible to link this identity to the AI system.

2. Transparency

The parties must check whether they have a corresponding picture of the possibilities and impossibilities of the AI system used.

If possible, builders and users of AI systems provide clear insight into the functioning of the AI systems they have built or offered.

Builders and users always give the end user insight into the history of the AI systems that they have built or offered. This principle is only an exception in those cases in which the generation of an action history is not legally required and is not reasonably possible

3. Integrity

Parties shall ensure the integrity of the AI system, the information stored therein and the transfer thereof.

Parties take appropriate measures to detect violations of the integrity of an AI system, and make agreements about the actions that need to be taken when a violation is detected.

4. Confidentiality

Parties ensure the confidentiality of the stored information in AI systems built or used by them.

Parties take appropriate measures to detect unauthorized disclosure of confidential information, and make agreements about the actions that should be taken when an unlawful disclosure is observed.

5. Continuity

Parties ensure the continuity of the AI-systems offered or used by them.

Parties take appropriate measures to prevent an error in an AI system or the platform on which it is running, leading to the complete loss of an AI system.

6. Testability, predictability and traceability

Parties ensure the traceability, testability and predictability of the actions performed by an AI-system.

Parties ensure the integrity of the logs generated by AI-systems.

Parties ensure the confidentiality of generated logs.

7. Intellectual property

Parties (builder, user and other stakeholders and / or end-user) will make prior clear agreements about the intellectual property rights and trade secrets relating to the system. This includes in any case: ownership / use of existing intellectual property rights and business secrets of one or more parties, and ownership / registration / enforcement of intellectual property rights and trade secrets arising from the development and / or use of the system.

Before use, it must be examined whether and if so which intellectual property rights of third parties play a role in the system. Subsequently, the parties must ensure that no such intellectual property right is infringed.

8. Privacy

The processing of personal data must be lawful, proper and transparent.[32]

The collection and further processing of personal data must be bound to specific goals.[33]

The data must be adequate, relevant and limited to what is necessary.[34] The data must be correct.[35] The data may not be stored longer than necessary.[36] The data must be properly secured. [37]

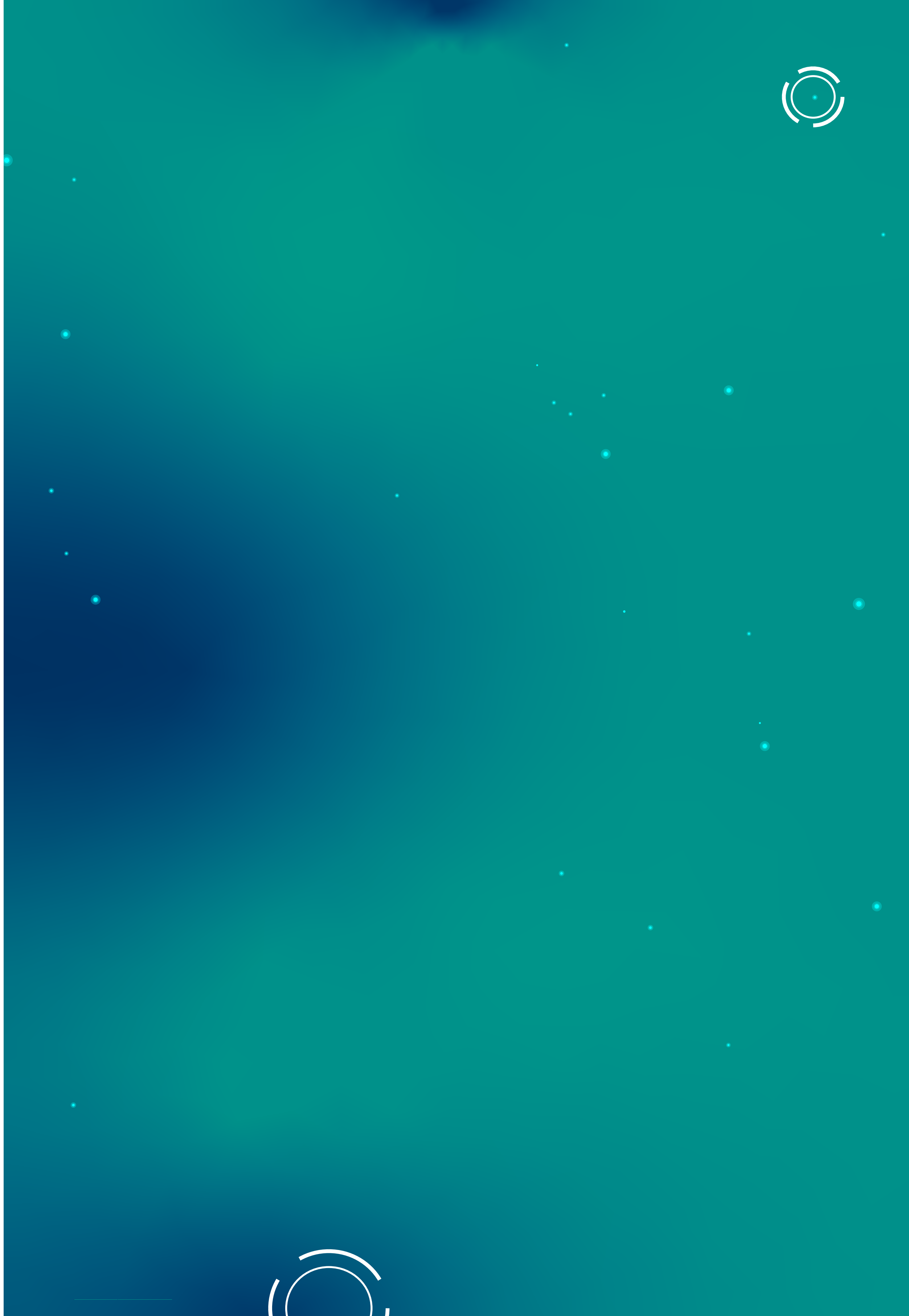
Data subjects have the right not to be subject to automated decision-making that has legal consequences or otherwise significantly affects the data subject. [38]

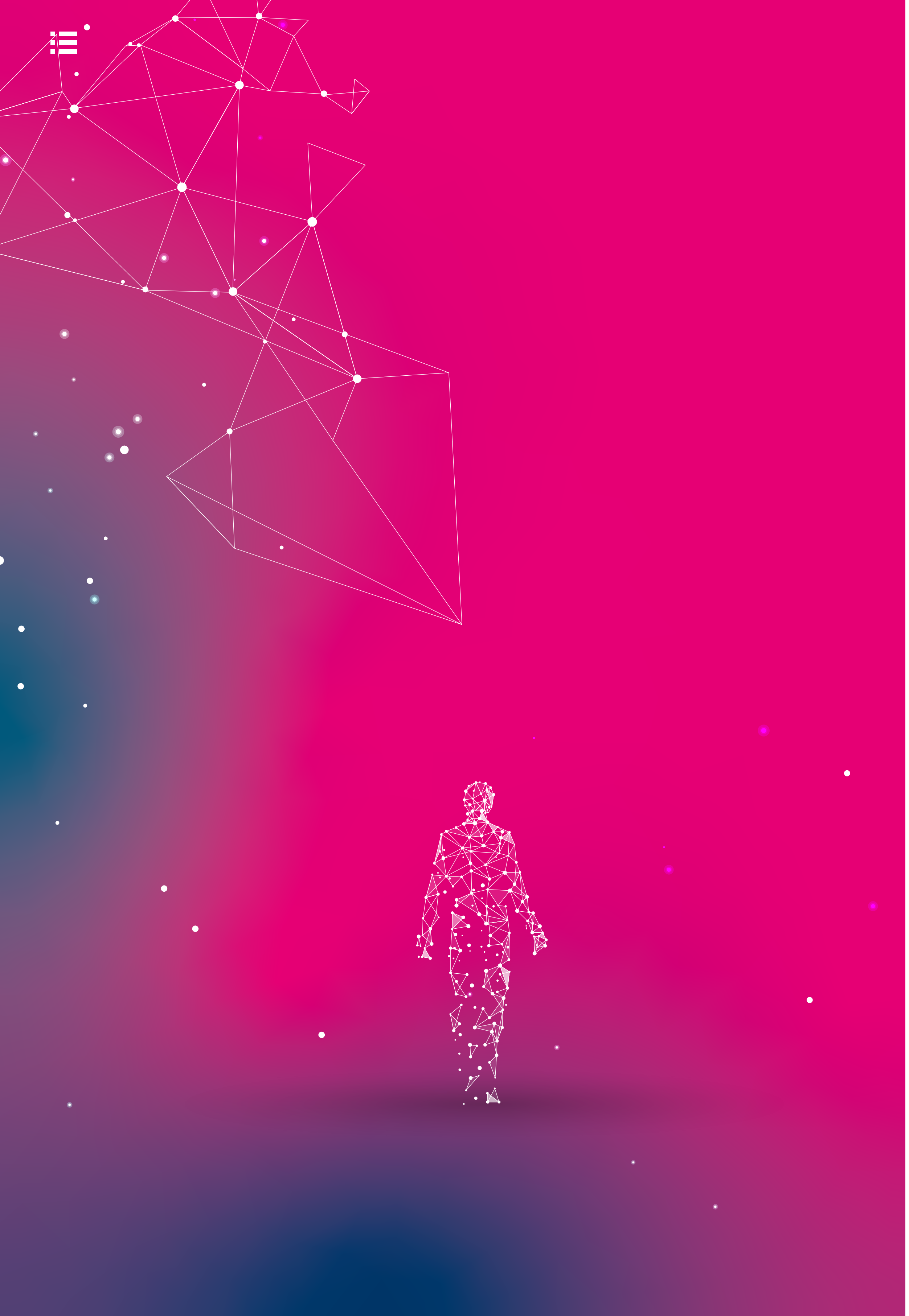
9. Responsibility

In the development and application of complex AI systems where many components and (sub) service providers play a role and where behaviour cannot always be traced back to specific components or service providers, measures must be taken to ensure that delineation of responsibilities is clear.

10. Audit

Before using an AI system, it must be determined how (resources, process) the relevant aspects of the information processing can be verified by means of an audit.





Annex 2 - AIIA roadmap

Step 1 **Is it useful to do an AIIA?**

Determine on the basis of the following screening questions whether it is useful to do an AIIA.

Is the AI used in a new (social) domain?	□ Yes	□ No
Does the application take place on a sensitive (social) terrain or subject?	□ Yes	□ No
Is a new form of AI technology used?	□ Yes	□ No
Does the AI have a high degree of autonomy?	□ Yes	□ No
Does the AI make complex decisions?	□ Yes	□ No
Is the AI used in a complex environment?	□ Yes	□ No
Are sensitive personal data used?	□ Yes	□ No
Does the AI make decisions that have a significant impact on persons or entities or have legal consequences for them?	□ Yes	□ No
Are the results of the AI application no longer (fully) understandable?	□ Yes	□ No

If the answer to one or more of these questions is 'Yes' then it makes sense to do an AIIA. Go to Step 2.

Step 2 **Describe the AI application**

Answer the following questions about the intended use of AI.

1. What is the purpose of the application?
2. Which AI technology/technologies are used to achieve the goal?
3. Which data are used to achieve the goal?
4. Which actors (suppliers, end-users, other stakeholders) play a role in the application?

Step 3 **Describe the benefits of the AI application**

Describe the positive aspects / benefits of the application by answering the following questions:

1. What are the benefits for the organisation?
2. What are the benefits for the individual?
3. What are the benefits for society as a whole?

Step 4 **Is the goal and the way the goal is reached ethical and legally safe?**

Describe the influence the application has on human and social values. If values are negatively influenced by the application (e.g. privacy risks or negative environmental effects), it must be substantiated how these risks are reduced and if there is residual risk, why this is accepted. Think of values such as:

1. Human dignity
2. Autonomy (freedom)
3. Responsibility
4. Transparency

5. Fairness
6. Democracy and rule of law
7. Safety
8. Privacy and data protection
9. Sustainability

Note: Whether an application is ethical, apart from the goal, is also highly dependent on the design of the preconditions (Step 5).

Step 5 **Is the application reliable, safe and transparent?**

Describe the preconditions for the ethical deployment of AI (reliability, safety, transparency) by answering the following questions:

1. Which measures have been taken to guarantee the reliability of the acting of the AI?

2. Which measures have been taken to guarantee the safety of the AI?
   - How is the safety of the AI in relation to the outside world guaranteed?
   - How is the (digital) safety of the AI itself guaranteed?

3. Which measures have been taken to guarantee the transparency of the acting of the AI?
   - Is the functioning of the AI (the logic of decision making) clear / public?
   - Is it clear what the consequences are of the deployment of AI (in particular the consequences for end users)?
   - Have measures been taken to be able to account for the application (*accountability*)?

Step 6 **Considerations and assessment**

On the basis of the above (steps 3, 4 and 5 in particular), weigh up whether the application as a whole is ethical. The following aspects can be included in this assessment:

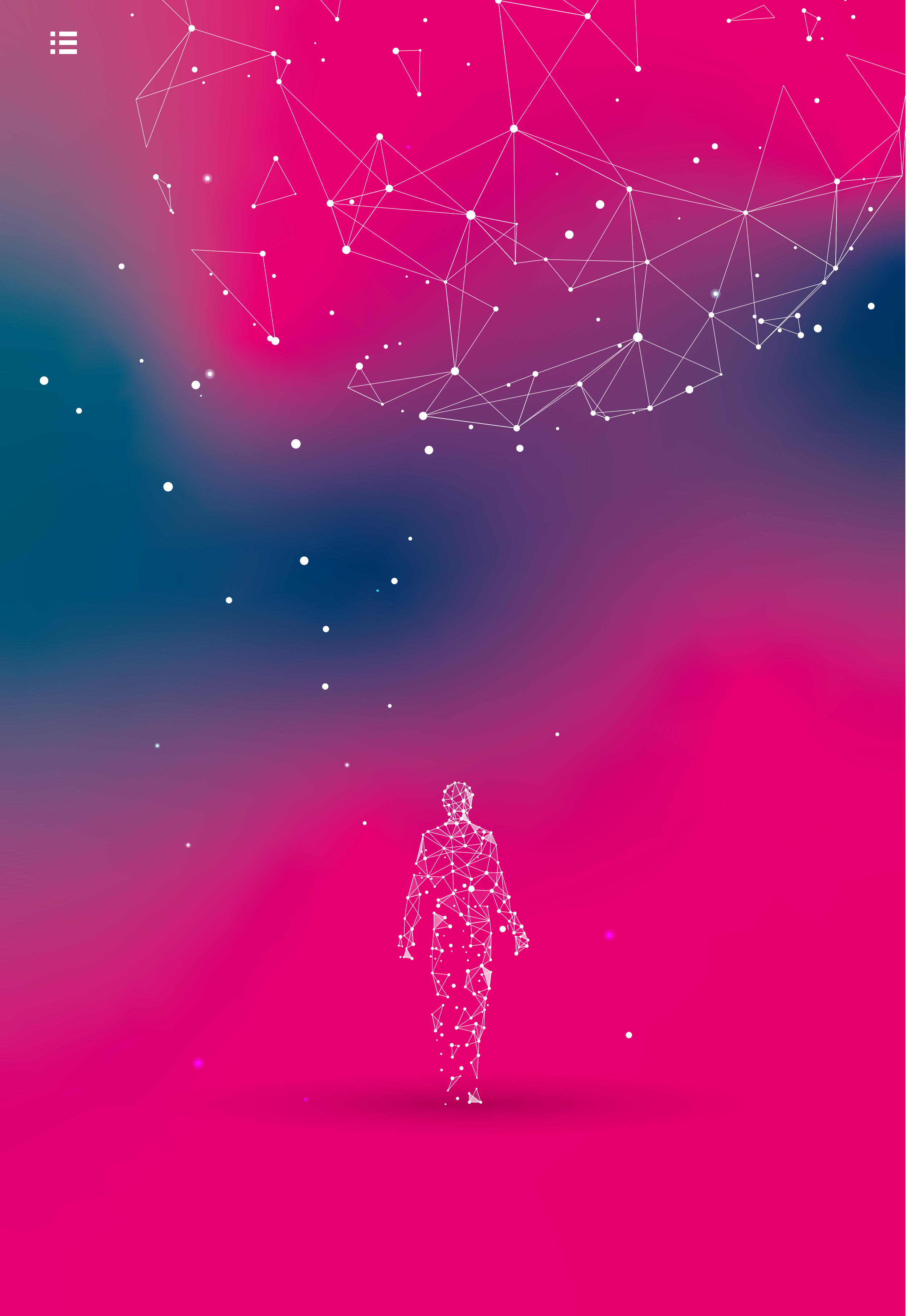
1. Is the application in proportion?
2. Can the same goal be achieved with less drastic means (subsidiarity)?
3. Is the choice for the application positive *sum* or *zero sum*
4. What are residual risks and why are they acceptable?
5. Will further use be taken into account (*downstream responsibility*)?

Step 7 **Documentation and accountability**

Record the answers to the above questions so that the choices can be accounted for, both internally and externally.

Step 8 **Assess periodically**

Evaluate in case of changes to the application and/or periodically if the above conclusions still apply.



Comments

1 https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Artificial-Intelligence-Automation-Economy.PDF(OECD, 2018)

2 Practical tools are, among others, the Ethical Data Assistant (https://dataschool.nl/deda/), the AI Ethics Framework (https://www. migarage.ai/ethics-framework/), the AI NOW Algorithmic Impact Assessment (https://ainowinstitute.org/aiareport2018.pdf) and the Princeton Dialogues on AI and Ethics (https://www.migarage.ai/ ethics-framework/)..

3 Values do not have a fixed definition that can be converted into a code, but are dependent on various cultural, historical and social factors. Within this AIIA, where values are discussed where possible, it is made as clear as possible what is meant by a certain value in the context of the AIIA.

4 The assessment distinguishes between the user of AI (the organisation that uses AI for services, the employee who works with AI when carrying out work), the developer (technology and platform parties, cloud service providers), the end user (who directly experiences the consequences of decisions and actions of the AI system, such as the driver of a self-driving car or the citizen who is faced with a decision taken by AI) and the stakeholders (the broader circle of parties who are affected by the deployment of AI: such as social and political organisations, professionals - and branch organisations).

5 The examples below are extreme examples and form a simplification of thinking within these ethical trends.

6 In many cases, the consideration of the deployment of AI within an organisation and the public debate is of a logical nature: if the result of the use of AI has a positive effect on the interests mentioned in the AIIA, or in a weighing of interests it is justified to accept certain risks of AI, the application is regarded as ethical and legitimate..

7 Virtues are qualities of a person who are considered morally good. The four cardinal virtues are prudence, fairness, moderation, and courage..

8 If it is to be expected that you will process personal data when using an AI, it is advisable to combine this step with the consideration of whether a DPIA is necessary.

9 Also see article 9 General Data Protection Regulation

10 Also see article 22 General Data Protection Regulation

11 When we speak about decision making by AI, we mean the action of the AI to arrive at the optimal outcome for the goals and values as defined by man. So although the AI makes decisions in order to arrive at an optimal outcome, this is based on the goals and the associated objective functions as defined by the user. The outcome can also be an advice, whereby a person makes the actual decision in the end..

12 The question whether an AI should have an ethical awareness of course depends to a great extent on the context and complexity of the deployment of AI.

13 Asimov's 'Three Laws of Robotics' are a popular example of such an hierarchy.

14 Quality can relate to the data itself (for example, are the data consistent and complete) but can also relate to substantive qualities such as truthfulness. Synthetic data are data generated by the computer. These are data that are not 'real', but that reflect a data set with 'real' data as close as possible. Synthetic datasets are used, among other things, to prevent testing with real personal data .

15 See for instance: IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017). For social cost-benefit analyses, you can connect to the *General Guideline for social cost-benefit analysis of the CPB and PBL and / or the Guide for social cost-benefit analysis in the digital government.*

16 This is not an exhaustive list. What values and interests are affected differs, of course, per application.. It is up to the organisation itself to determine whether other values and interests are at stake. The values and interests mentioned also strongly depend on each other. Therefore these are not interests that must be weighted in isolation.

17 https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/

18 Just as knowledge of the anatomy of an organism or the functioning of cells has only limited predictive value for predicting behaviour.

19 The ethical principles are based on: European group on ethics in science and new technologies. (2018). Statement on artificial intelligence, robotics and autonomous systems. Brussels: European Commission. Retrieved 05 01, 2018, from https://ec.europa.eu/research/ege/pdf/ege_ ai_ statement_2018.pdf

20 **Explanatory note to EGE (2018):** The principle of human dignity, understood as the recognition of the inherent human state of being worthy of respect, must not be violated by 'autonomous' technologies. This means, for instance, that there are limits to determinations and classifications concerning persons, made on the basis of algorithms and 'autonomous' systems, especially when those affected by them are not informed about them. It also implies that there have to be (legal) limits to the ways in which people can be led to believe that they are dealing with human beings while in fact they are dealing with algorithms and smart machines. A relational conception of human dignity which is characterised by our social relations, requires that we are aware of whether and when we are interacting with a machine or another human being, and that we reserve the right to vest certain tasks to the human or the machine**.**

21 **Note to the EGE (2018):** The principle of autonomy implies the freedom of the human being. This translates into human responsibility and thus control over and knowledge about 'autonomous' systems as they must not impair freedom of human beings to set their own standards and norms and be able to live according to them. All 'autonomous' technologies must, hence, honour the human ability to choose whether, when and how to delegate decisions and actions to them. This also involves the transparency and predictability of 'autonomous' systems, without which users would not be able to intervene or terminate them if they would consider this morally required

22 **Note to EGE (2018):** The principle of responsibility must be fundamental to AI research and application. 'Autonomous' systems should only be developed and used in ways that serve the global social and environmental good, as determined by outcomes of deliberative democratic processes. This implies that they should be designed so that their effects align with a plurality of fundamental human values and rights. As the potential misuse of 'autonomous' technologies poses a major challenge, risk awareness and a precautionary European Group on Ethics in Science and New Technologies 17 approach are crucial. Applications of AI and robotics should not pose unacceptable risks of harm to human beings, and not compromise human freedom and autonomy by illegitimately and surreptitiously reducing options for and knowledge of citizens. They should be geared instead in their development and use towards augmenting access to knowledge and access to opportunities for individuals. Research, design and development of AI, robotics and 'autonomous' systems should be guided by an authentic concern for research ethics, social accountability of developers, and global academic cooperation to protect fundamental rights and values and aim at designing technologies that support these, and not detract from them.

23 **Note to EGE (2018):** AI should contribute to global justice and equal access to the benefits and advantages that AI, robotics and 'autonomous' systems can bring. Discriminatory biases in data sets used to train and run AI systems should be prevented or detected, reported and neutralised at the earliest stage possible. We need a concerted global effort towards equal access to 'autonomous' technologies and fair distribution of benefits and equal opportunities across and within societies. This includes the formulating of new models of fair distribution and benefit sharing apt to respond to the economic transformations caused by automation, digitalisation and AI, ensuring accessibility to core AI-technologies, and facilitating training in STEM and digital disciplines, particularly with respect to disadvantaged regions and societal groups. Vigilance is required with respect to the downside of the detailed and massive data on individuals that accumulates and that will put pressure on the idea of solidarity, e.g. systems of mutual assistance such as in social insurance and healthcare. These processes may undermine social cohesion and give rise to radical individualism.

24 **Toelichting EGE (2018):** Key decisions on the regulation of AI development and application should be the result of democratic debate and public engagement. A spirit of global cooperation and public dialogue on the issue will ensure that they are taken in an inclusive, informed, and farsighted manner. The right to receive. education or access information on new technologies and their ethical implications will facilitate that everyone understands risks and opportunities and is empowered to participate in decisional processes that crucially shape our future. The principles of human dignity and autonomy centrally involve the human right to self-determination through the means of democracy. Of key importance to our democratic political systems are value pluralism, diversity and accommodation of a variety of conceptions of the good life of citizens. They must not be jeopardised, subverted or equalised by new technologies that inhibit or influence political decision making and infringe on the freedom of expression and the right to receive and impart information without interference. Digital technologies should rather be used to harness collective intelligence and support and improve the civic processes on which our democratic societies depend.

25 See Advisory Council on International Affairs (2017), The will of the people? Erosion of the democratic constitutional state in Europe

26 See: https://www.theguardian.com/news/series/cambridge-analytica-files

27 **Note to EGE (2018):** Rule of law, access to justice and the right to redress and a fair trial provide the necessary framework for ensuring the observance of human rights standards and potential AI specific regulations. This includes protections against risks stemming from 'autonomous' systems that could infringe human rights, such as safety and privacy. The whole range of legal challenges arising in the field should be addressed with timely investment in the development of robust solutions that provide a fair and clear allocation of responsibilities and efficient mechanisms of binding law. In this regard, governments and international organisations ought to increase their efforts in clarifying with whom liabilities lie for damages caused by undesired behaviour of 'autonomous' systems. Moreover, effective harm mitigation systems should be in place.

28 **Note to EGE (2018):** Security, safety, bodily and mental integrity: Safety and security of 'autonomous' systems materialises in three forms: (1) external safety for their environment and users, (2) reliability and internal robustness, e.g. against hacking, and (3) emotional safety with respect to human-machine interaction. All dimensions of safety must be taken into account by AI developers and strictly tested before release in order to ensure that 'autonomous' systems do not infringe on the human right to bodily and mental integrity and a safe and secure environment. Special attention should hereby be paid to persons who find themselves in a vulnerable position. Special attention should also be paid to potential dual use and weaponisation of AI, e.g. in cybersecurity, finance, infrastructure and armed conflict.

29 Also see under 'Step 4', point 2: Which measures have been taken to guarantee the safety of the AI

30 **Note to EGE (2018):** Data Protection and Privacy: In an age of ubiquitous and massive collection of data through digital communication technologies, the right to protection of personal information and the right to respect for privacy are crucially challenged. Both physical AI robots as part of the Internet of Things, as well as AI softbots that operate via the World Wide Web must comply with data protection regulations and not collect and spread data or be run on sets of data for whose use and dissemination no informed consent has been given. 'Autonomous' systems must not interfere with the right to private life which comprises the right to be free from technologies that influence personal development and opinions, the right to establish and develop relationships with other human beings, and the right to be free from surveillance. Also in this regard, exact criteria should be defined and mechanisms established that ensure ethical development and ethically correct application of 'autonomous' systems. In light of concerns with regard to the implications of 'autonomous' systems on private life and privacy, consideration may be given to the ongoing debate about the introduction of two new rights: the right to meaningful human contact and the right to not be profiled, measured, analysed, coached or nudged.

31 **Note to EGE (2018):** Sustainability: AI technology must be in line with the human responsibility to ensure the basic preconditions for life on our planet, continued prospering for mankind and preservation of a good environment for future generations. Strategies to prevent future technologies from detrimentally affecting human life and nature are to be based on policies that ensure the priority of environmental protection and sustainability.

32 Personal data may only be processed for legitimate purposes. This means that when artificial intelligence is used, it must first be determined with what purpose the data for / by artificial intelligence are processed. This must also be transparent for the outside world, more specifically those involved.

33 Once data have been collected for the legitimate purpose described above, the data may also only be processed for this purpose. The only exception to this rule is when the new purpose is compatible with the original overall purpose.

34 No more data may be processed than is necessary for the purpose of the processing (data minimization). The use of datasets by / for artificial intelligence must therefore be limited to what is necessary for the proper functioning of the artificial intelligence for the purpose of the specified goal to which the artificial intelligence is used. Data minimization does not always mean 'as little data as possible'. The artificial intelligence must have enough data to function correctly.

35 he data must be correct and up-to-date. Incorrect or outdated data must be modified or deleted..

36 Personal data may not be kept longer than necessary for the purpose of the processing. Data that no longer serve the processing purpose must be anonymized or delete

37 The confidentiality, integrity and availability of personal data in the use of personal data for / by artificial intelligence must be guaranteed with appropriate technical and organisational measures. In addition to these general principles, article 22 GDPR is also relevant in the context of artificial intelligence.

38 When an artificial intelligence makes decisions without human intervention (algorithmic decision-making), this is not permitted if this has legal consequences or the parties involved become significantly different in their rights. The GDPR and the Dutch GDPR Implementation Act make some specific exceptions to this general prohibition. For example, if there is explicit permission from the person concerned, or the decision-making is necessary for the conclusion of an agreement, then the decision-making is permitted. In addition, specific exceptions can be created in national or European law.