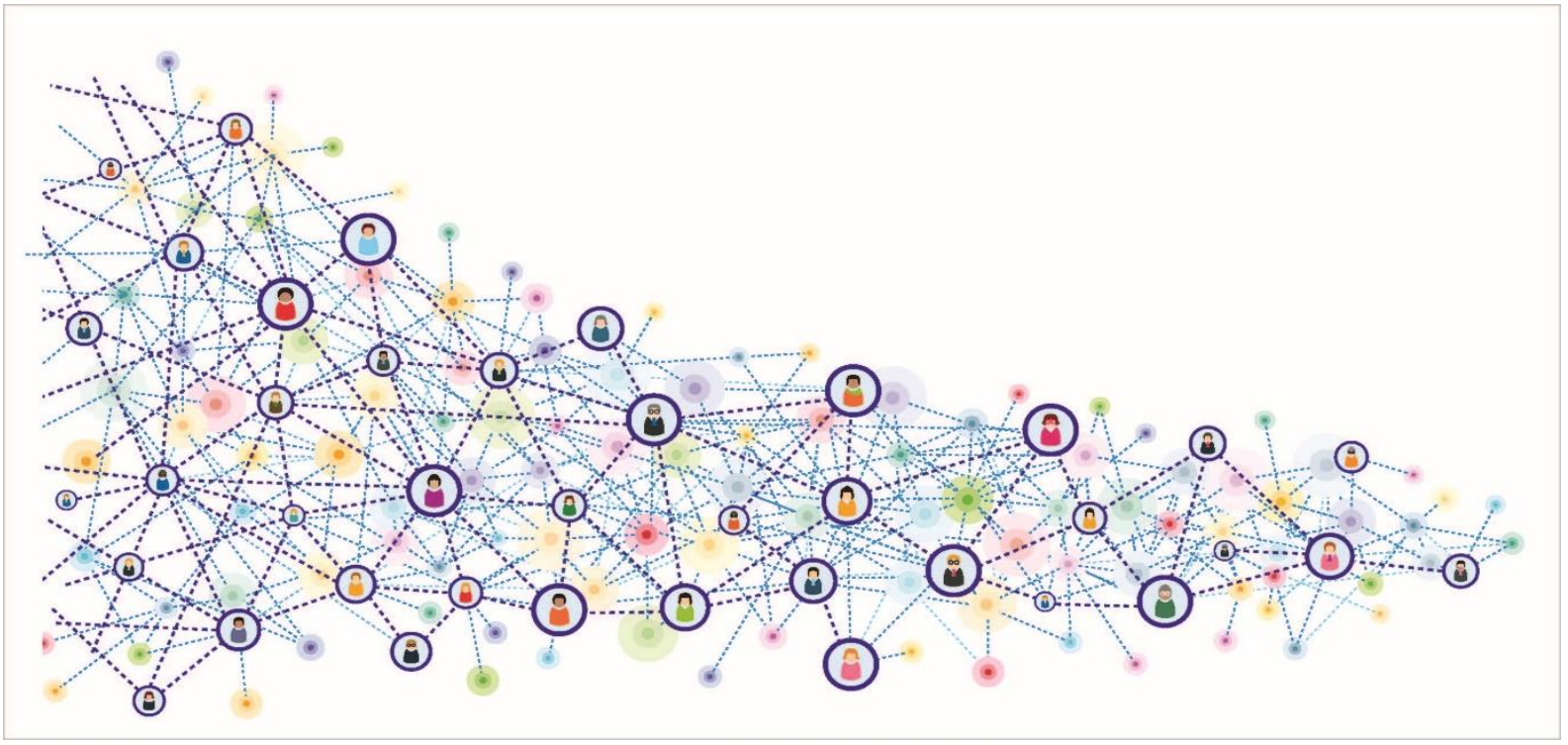
SHERPA

Shaping the ethical dimensions of smart information systems– a European perspective (SHERPA)

Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach



Main authors: Philip Brey, Björn Lundgren, Kevin Macnish, and Mark Ryan.
Other contributors: Andreas Andreou, Laurence Brooks, Tilimbe Jiya, Renate Klar, Dirk Lanzareth, Jonne Maas, Isaac Oluoch, and Bernd Stahl.
Acknowledgment: We would like to thank the participants of the workshop in July 2019 and those who provided feedback on our guidelines.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme Under Grant Agreement no. 786641

Executive summary

This report contains ethical guidelines for the technological development of artificial intelligence (AI) and big data systems. It is a Deliverable of the SHERPA project, an EU Horizon 2020 project on the ethical and human rights implications of AI and big data. The guidelines differ from others in that they are directly related to design and development practices. They are intended to be actionable guidelines for systems and software development, rather than abstract principles that have no direct application in practice. We call such guidelines *operational*, meaning ready for use. Applying these guidelines in development practices would result in more ethical AI and big data products.

In constructing *Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach*, we have incorporated input from a wide diversity of stakeholders, SHERPA partners, and insights from other guidelines. In a survey of potential guidelines we found over 70 matching documents, which were reduced to 25 suitable guidelines that we built on. After an introductory section, we devote Section 2 of this report ("High-Level Requirements") to present and discuss the high-level requirements that form the point of departure for this report. Our requirements are directly based on the guidelines of the EU's High-Level Expert Group on Artificial Intelligence (HLEG AI), with minor adaptations to improve coherence and fitness for operationalization. This results in the following seven requirements that mirror those of the HLEG AI: human agency, liberty, and dignity; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; individual, societal, and environmental wellbeing; and accountability. For each, we specify three to four sub-requirements that constitute a first step towards operationalization.

In Section 3 ("Applying ethics to development models for AI and big data systems"), we discuss models for the development of AI and big data systems, and how ethical principles could be made part of these models. While different development models include similar phases and practices (e.g., defining requirements, collecting data, evaluating the design), we use one particular development model, CRISP-DM, to present our operational (or "low-level") ethical principles. CRISP-DM is widely used for the development of data analytics and data-intensive AI systems. We also briefly discuss a currently popular approach for software development, Agile, but do not present a full operationalization of ethical principles for Agile at this point.

The CRISP-DM model identifies six major phases in the development process: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase has four to five sub-phases. Our objective is to develop operational requirements that are based on the high-level requirements and tied to different development phases and sub-phases. In Section 3, we provide some general guidelines for implementing ethical requirements in CRISP-DM. In Section 4, we provide operational guidelines for the seven requirements that were presented in Section 2.

In Section 5, we present and discuss ethical guidelines for special topics in AI and big data. By special topics, we mean AI / big data systems, applications, data types, or application domains that require special consideration. We present ten such special topics, ranging from the processing of medical data, to AI systems that recognize and produce emotions, to the application of AI and big data in defence. In our

model, special topics should be identified and taken into account at the Business Understanding phase in CRISP-DM.

The guidelines that we present in this report are operational in the sense that they are, in our view, ready to be used by ethics officers or managers, who have a responsibility for ensuring the implementation of ethical practices within their organizations. They are perhaps not directly usable by system developers. A further step that is required, but not contained in this report, is the training of developers in this new framework, and the assignment of different roles and responsibilities to them for ensuring that the ethical requirements are met. This may also require the development of training materials and operational guides for professionals with different roles in the development process. We intend to produce further implementation documents in the EU Horizon 2020 SIENNA project (www.sienna-project.eu).

1. Introduction

These guidelines, on the ethical *development* of artificial intelligence (AI) and big data systems, are part of a set of two (with separate guidelines for ethical *use*). These guidelines have been created by the SHERPA project, which has focused on the ethical, legal, and social issues arising from the development and use of AI and big data systems. They are intended to be implemented in your organization by a manager, and preferably (where one exists), by an ethics officer.[1] These guidelines can also be useful in ethical research assessment, with the reservation that these guidelines focus on achieving an ethical AI and big data system and the impact of such systems, not the process as such (ethical research boards are often concerned with the process). Applying these guidelines in development practices, or for research assessment, would result in more ethical AI and big data products or research.

In constructing these guidelines, we incorporated input from a wide diversity of stakeholders, SHERPA partners, and insights from other guidelines. In a survey of potential guidelines we found over 70 matching documents, which were reduced to 25 suitable guidelines that we built on, to construct *Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach*.[2] In particular, these guidelines are built closely on the EU's High-Level Expert Group on Artificial Intelligence (AI HLEG). Our aim has been to build on their fundamental values, but we seek to go further in producing guidelines that are more operational and directly useful in development practices.

> When reading these guidelines, it is important to keep in mind that when we refer to **users**, we are referring to organisations that deploy and use these AI and big data systems. This is distinct from a customer/individual using these technologies, who we will refer to as the **end-user**. When we talk of an AI and big data system, we will often refer to it as **the system**. And we will talk about stakeholders as individuals that have a stake in and/or can be affected by a system.

These guidelines begin by briefly describing the different types of requirements, starting with the top values (Section 2). Next, we describe how the ethical analyses can be mapped onto and integrated with IT development methods. We illustrate this with the so-called 'CRISP-DM' and 'Agile' methods in Section 3. After this analysis of how to integrate ethics into development methods, we turn to our specified ethical requirements in Section 4. Although these build on the analysis from the previous section, they do not depend on it directly and can be read as a standalone set of guidelines for development of these systems. In Section 5 we address special ethical issues related to these systems that are not captured in the general guidelines, and how our guidelines may provide recommendations for these topics.

[1] In the closely related SIENNA project (https://www.sienna-project.eu/) we are developing tools that can be used by a broader set of people within the organisation (such as engineers).

[2] The requirement included eight criteria: 1. Language: The document should be in English, or have an official translation in English; 2. Date: The document should be from 2012 or later, because of the pace of developments in AI; 3. Ethics focus: The document, or at least a large part of it, should have a clear ethical focus; 4. AI or Big Data focus: The document should have a focus on AI and/or Big Data; 5. Breadth: The document focuses on ethical issues for AI and/or Big Data in general, not solely on certain applications or techniques of AI or Big Data (such as self-driving cars or robots); 6. Guidance: The document should provide clear guidelines, norms or proposals for behaviour; 7. Level of operationalization: The document should be more extensive than a short list of principles, and it should provide context, operationalization and guidance for implementation; 8. Recognition and endorsement: The document is widely known, cited and/or used, and/or endorsed by important industry sectors, multinationals, organisations or governments.

Finally, these guidelines are complemented by more substantial materials from our full report. In that report is a glossary, which may be of use in reading the guidelines. We have made that glossary available in our online workbook.[3]

[3] https://www.project-sherpa.eu/workbook/

2. High-Level Requirements

We distinguish between high-level, intermediate level, operational, and specific operational guidelines or requirements. High-level requirements are abstract general principles or values. Many proposed sets of ethical guidelines for AI are of this general nature. Intermediate-level guidelines are more specific, providing more concrete conditions that must be fulfilled. Operational guidelines are tied to specific practices, while specific operational guidelines prescribe specific actions to be taken. In this report, we move from high-level to operational guidelines for the development of AI and big data.

In this Section we will briefly describe these high-level requirements to provide an insight into the fundamental principles and values behind the specific requirements. Readers who are familiar with the AI HLEG will notice that our high-level requirements are based directly on its high-level requirement, with some minor changes intended to improve their coherence and fitness for operationalization.

SHERPA High-level requirements and sub-requirements
1 Human agency, liberty and dignity: Positive liberty, negative liberty and human dignity
2 Technical robustness and safety: Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3 Privacy and data governance: Including respect for privacy, quality and integrity of data, access to data, data rights and ownership
4 Transparency: Including traceability, explainability and communication
5 Diversity, non-discrimination and fairness: Avoidance and reduction of bias, ensuring fairness and avoidance of discrimination, and inclusive stakeholder engagement
6 Individual, societal and environmental wellbeing: Sustainable and environmentally friendly AI and big data systems, individual wellbeing, social relationships and social cohesion, and democracy and strong institutions
7 Accountability: auditability, minimisation and reporting of negative impact, internal and external governance frameworks, redress, and human oversight

Table 1 [Development]: SHERPA High-level requirements

Below we briefly explain the high-level requirements and their sub-requirements.

2.1 Human Agency, Liberty and Dignity

Because we value the ability for humans to be autonomous and self-governing (*positive liberty*), humans' freedom from external restrictions (*negative liberties*, such as freedom of movement or freedom of association), and because we hold that each individual has an inherent worth and that we should not undermine the respect for human life (*human dignity*), we need to ensure that AI and big data systems do not negatively affect human agency, liberty, and dignity.

2.2 Technical Robustness and Safety

Because we value humans, human life, and human resources, it is important that the system and its use is safe (often defined as an absence of risk) and secure (often defined as a protection against harm, i.e., something which achieves safety). Under this category we also include the quality of system decisions in terms of their accuracy, reliability, and precision.

2.3 Privacy and Data Governance

Because AI and big data systems often use information or data that is private or sensitive, it is important to make sure that the system does not violate or infringe upon the right to privacy, and that private and sensitive data is well-protected. While the definition of privacy and the right to privacy is controversial, it is closely linked to the importance of an individual's ability to have a private life, which is a human right. Under this requirement we also include issues relating to quality and integrity of data (i.e., whether the data is representative of reality), and access to data, as well as other data rights such as ownership.

2.4 Transparency

Because AI and big data systems can be involved in high-stakes decision-making, it is important to understand how the system achieves its decisions. Transparency, and concepts such as explainability, explicability, and traceability relate to the importance of having (or being able to gain) information about a system (transparency), and being able to understand or explain a system and why it behaves as it does (explainability).

2.5 Diversity, Non-discrimination and Fairness

Because bias can be found at all levels of the AI and big data systems (datasets, algorithms, or users' interpretation), it is vital that this is identified and removed. Systems should be developed with an inclusionary, fair, and non-discriminatory agenda. Including people from diverse backgrounds (e.g., different ethnicities, genders, disabilities, ideologies, and belief systems), stakeholder engagement, and diversity analysis reports and product testing, are ways to include diverse views into these systems.

2.6 Individual, Societal and Environmental Wellbeing

Because AI and big data systems can have huge effects for individuals, society, and the environment, systems should be trialed, tested, and anomaly-detected, to ensure the reduction, elimination, and reversal of harm caused to individual, societal and environmental wellbeing.

2.7 Accountability

Because AI and big data systems act like agents in the world, it is important that someone is accountable for the systems' actions. Furthermore, an individual must be able to receive adequate compensation in the case of harm from a system (redress). We must be able to evaluate the system, especially in the situation of a bad outcome (audibility). There must also be processes in place for minimisation and reporting of negative impact, with internal and external governance frameworks (e.g., whistleblowing), and human oversight.

3. Applying ethics to development models for AI and big data systems

In this section, we discuss how ethics can be integrated into development methods. We illustrate this in detail by focusing on two such methods, but the important requirements for how to integrate ethics do not necessarily depend on the chosen method.

We consider the responsible and ethical development of AI and big data systems to be the outcome of three factors:

1. Responsible development models and methods for the system;
2. Responsible corporate structure and policy in AI and big data industry;
3. Support for responsible development by society (e.g., by governmental institutions, educational institutions, professional organisations, clients).

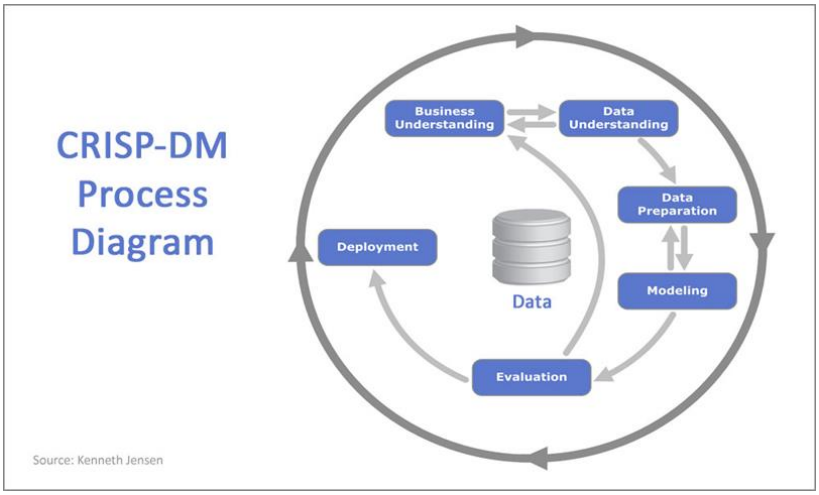
In this section, we will focus on the first of these. This means that we will mainly focus on technological methods for responsible AI.

In this section, we describe the well-known CRISP-DM model.[4] CRISP-DM is currently the most widely used model for the development of data analytics and data-intensive AI systems. CRISP-DM stands for **Cross-industry standard process for data mining**. We will describe the six development steps, and do so in a way that helps prepare a discussion of how ethical considerations may come into play at different steps.

We will also discuss the Agile framework, which is a response to traditional plan-driven approaches, which are unable to adapt to the changing wishes of customers or new discoveries in technology. The Agile model contains principles that must be followed to satisfy the 'Agile' criterion, which will be discussed further in Section 3.2.

3.1 CRISP-DM Model

CRISP-DM is built out of six steps or phases in the development process. These are intended to be sequential but also iterative; developers may go back and forth between different phases at different points in time (as illustrated in the diagram to the right). Below, we describe the six steps and present our requirements for how to integrate ethical considerations into this process.



[4] Shearer, Colin, "The CRISP-DM model: the new blueprint for data mining", *Journal of data warehousing,* Vol. 5, No. 4, 2000, pp. 13-22.; Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth, "CRISP-DM 1.0: Step-by-step data mining guide", *SPSS inc,* Vol. 16, 2000.

3.1.1 Business Understanding

At this stage, business objectives are defined and translated into project objectives and a project plan. It involves four sequential tasks:

1. *Determine business objectives.* What is it that the customer wants to accomplish? This involves defining the primary business objective, related business questions, and criteria for a successful or useful outcome.
2. *Assess situation.* This involves more detailed fact-finding about what is needed to realize the project. It includes (1) *inventory of resources* (required and available personnel, data, computing resources and software); (2) *requirements, assumptions, and constraints*: requirements include schedule of completion, comprehensibility and quality of results, security, and legal issues; assumptions include assumptions about the data that can be verified during data mining, and assumptions about the business related to the project; constraints include constraints on availability of resources and technological constraints; (3) *risks and contingencies*: risks or events that may delay the project or cause failure, and corresponding contingency plans. This step also includes a cost-benefit analysis.
3. *Determine data mining goals*. This is the translation of business objectives into technical terms: what must the system be able to do to contribute to the business objectives? Data mining objectives are defined, and also data mining success criteria.
4. *Produce project plan.* This is the initial plan for realizing the data mining goals and hence the business goals. It lists the various stages to be executed in the project, together with their duration, resources required, inputs, outputs, and dependencies. It also includes an initial assessment of tools and techniques.

Requirement 1a: *Inclusion of ethics requirements and ethical assessment of business objectives*

To integrate ethics into the business understanding phase, start in task 2, include the seven requirements for ethical and trustworthy AI in the list of requirements and test the business objectives formulated in task 1 against the seven ethics requirements (in practice you will also need to look to the specific requirements in Section 4; also, test whether any of the special issues are likely to be involved, and if so, the guidelines for the special issues should be involved). This step is needed to establish tensions between the business objectives and ethics requirements. Sometimes there can be a basic incompatibility between the objectives of a system and ethics requirements. For example, the objective may be to engage in covert surveillance of people (violating principles of privacy and autonomy), or to engage in politically driven censorship of news feeds (violating principles of freedom of information and societal wellbeing (democracy)). Possible outcomes of this assessment are:

1. The business objectives are compatible with the ethics requirements. Proceed to next step.
2. The business objectives are inherently incompatible with ethics requirements. The development of the system should be terminated.
3. The business objectives are incompatible with ethics requirements, but modifications of the business objectives are possible to ensure compatibility. Modify business objectives and proceed to next step.
4. It is unclear whether business objectives are compatible with ethics requirements. Cautiously proceed to the next step, and keep monitoring closely.

As part of the ethical assessment, specific ethical issues that could be at play in the project in relation to the objectives should also be listed. For example, the ethical assessment may uncover specific risks to locational privacy, to psychological wellbeing, or to democratic institutions.

Requirement 1b: *Ethical assessment of resources, requirements and constraints*

During task 2, test the inventory of resources and other requirements and constraints against the ethics requirements for possible tensions (e.g., it may be found that the requirements of transparency and accountability cannot be met with available resources for the established business objectives). Make modifications to resources and to other requirements and constraints to reduce tensions with ethics requirements. Also specify which ethical issues may be at play, as in Requirement 1a. To perform this task you will need to make a proper evaluation of the costs involved in satisfying ethical requirements.

Requirement 1c: *Expanded cost-benefit assessment*

The cost-benefit assessment that is undertaken as part of the risk assessment in task 2 should be expanded to not only consider costs and benefits to the business, but also (included or separately) costs and benefits to stakeholders and society at large.

Requirement 2: *Ethical assessment of data objectives*

In task 3, test the data objectives against the seven ethics requirements. Explanation: even if the business objectives are compatible with the ethics requirements, the data objectives may be formulated in a way that is not compatible (e.g., it may propose a segmentation of people into social categories that was not referred to in the business objectives and that does not fit well with principles of fairness and equality). Outcomes of the assessment are the same as the four-step process in Requirement 1a.

Requirement 3a and 3b: *Stakeholder analysis (a) or involvement (b) in the business understanding phase*

Inclusion of ethical criteria in the development process could benefit from a stakeholder analysis, in which direct and indirect stakeholders to the project are identified and their values and interests are assessed. This makes it easier to identify more specific ethical requirements, make ethical assessments, and assess possible tensions between objectives and requirements and ethical criteria. Going further, stakeholders could also be consulted or be involved in decision-making.

3.1.2 Data Understanding

At this stage, initial data collection takes place, and an initial study of the data is performed. It involves four sequential tasks:

1. *Collect initial data.* Collect the data (or acquire access to the data) that is listed in the project resources.
2. *Describe data.* Examine the "gross" or "surface" properties of the acquired data (such as format and quantity), and evaluate whether the data satisfies the relevant requirements.
3. *Explore data.* In preparation for further steps, answer data mining questions that concern patterns in the data (e.g., distribution of key attributes, relationships between pairs of attributes, properties of

significant sub-populations, simple statistical analyses), through queries, visualization, and reporting techniques.

4. *Verify data quality.* Examine the quality of the data, including completeness, correctness, and missing variables.

Requirement 4a: *Ethical data collection and assessment*

To integrate ethical requirements into this phase, start by evaluating the data collection choice (task 1, above). Make necessary changes (if appropriate changes are not possible to perform, you may need to return to phase 1 to re-evaluate the business objectives). Follow the four-step process established in Requirement 1a. At this stage, bias, discrimination, fairness and diversity, privacy, and data quality will be particularly important.

Requirement 4b: *Ethical data description, exploration, and verification*

To integrate ethical requirements into the rest of the tasks in this phase, evaluate the ethical consequences of describing, exploring, and verifying the data, and make changes if necessary. Follow the four-step process established in Requirement 1a. At this stage, issues relating to privacy, data quality, precision, accuracy, transparency, explainability, bias, discrimination, and fairness and diversity will be particularly important.

3.1.3 Data Preparation

This stage includes all activities needed to construct the final dataset that is fed into the model, from initial raw data. It involves the following five tasks, not necessarily performed sequentially:

1. *Select data.* Decide on the data to be used for analysis, based on relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types.
2. *Clean data.* Raise data quality to a required level, for example by selecting clean subsets of the data, insertion of defaults, and estimation of missing data by modeling.
3. *Construct data.* The construction of new data through the production of derived attributes, new records, or transformed values for existing attributes.
4. *Integrate data.* Combine data from multiple tables or records to create new records or values.
5. *Format data*. Make syntactic modifications to data that might be required by the modeling tool.

Requirement 5: *Selection bias and risk of new data*

To integrate ethical requirements into this phase, start by evaluating the ethical consequences of data selection (e.g., in relation to diversity or transparency) and make changes, if necessary. Next, make sure that construction or integration of (new) data does not yield any further ethical issues (e.g., relating to privacy, or accuracy and precision). In both cases, follow the four-step process established in Requirement 1a.

3.1.4 Modeling

In this phase, modeling techniques are selected, applied, and optimized. Since some techniques have requirements on the form of data, going back to the data preparation phase is often necessary. This phase involves four sequential tasks:

1. *Select modeling technique.* Based on the general choice of tool, if any, at the business understanding phase, select a specific technique, e.g., neural network generation with backpropagation, or decision-tree building with Python.

2. *Generate test design.* Generate a procedure or mechanism to test the model's quality and validity.
3. *Build model.* This involves running the modeling tool on the prepared data set to create one or more models. A data model is a structuring of the data that can be used to serve the business objectives set for the model.
4. *Assess model*. Generated data models are assessed relative to the defined success criteria, such as accuracy and generality. This is continued until it is believed that one has identified the best model(s).

Requirement 6: *Ethical assessment of modelling*

To integrate ethical requirements into this phase, ensure that ethical criteria are considered in the modelling stage, and that the selection of the model(s) are evaluated relative to these ethical criteria. Issues that may be particularly relevant are those relating to transparency, and safety and robustness. Follow the four-step process established in Requirement 1a.

3.1.5 Evaluation

After the building of the model(s) in phase 5, this phase subjects the model to a thorough evaluation and review, to ensure that it achieves the business objectives. It involves three sequential tasks:

1. *Evaluate results.* The model is subjected to a broader evaluation, evaluating it against the business objectives and success criteria, evaluating models and results not related to the business objectives but still relevant to consider, and optionally testing the model on test applications.
2. *Review process.* A review is undertaken of the development process, for quality assurance purposes.
3. *Determine next steps.* Possible next steps are identified, with pros and cons for each. If the outcomes of the previous two steps are positive, the team normally goes on to deployment.

Requirement 7: *Ethical assessment of project outcomes*

As part of the evaluation phase, starting in task 1 ("evaluate results"), an ethical assessment should be performed of the results. Possible outcomes are that ethical issues have been dealt with in a satisfactory way, that further development is needed, or that specific guidance for or restrictions on deployment and use need to be in place to mitigate ethical issues. Follow the four-step process established in Requirement 1a.

Requirement 8a and 8b: *Stakeholder analysis (a) or involvement (b) in the evaluation phase*

As part of the ethical assessment in the evaluation phase (Requirement 7), a stakeholder analysis could be performed, or stakeholders could be consulted or involved in the decision-making. A more far-reaching proposal is to either do stakeholder analysis or engage stakeholders for decisions at all phases of the development process. This guarantees that their interests and values are continuously taken into account.

3.1.6 Deployment

Deployment is the process of getting an IT system to be operational in its environment, including installation, configuration, running, testing, and making necessary changes. Deployment is usually not done by the developers of a system but by the (IT team of the) customer. Nevertheless, even if this is the

case, developers will have a responsibility to supply the customer with sufficient information for successful employment of the model. This will normally include a (generic) deployment plan, with necessary steps for successful deployment and how to perform them, and a (generic) monitoring and maintenance plan for maintenance of the system, and for monitoring the deployment and correct usage of data mining results.

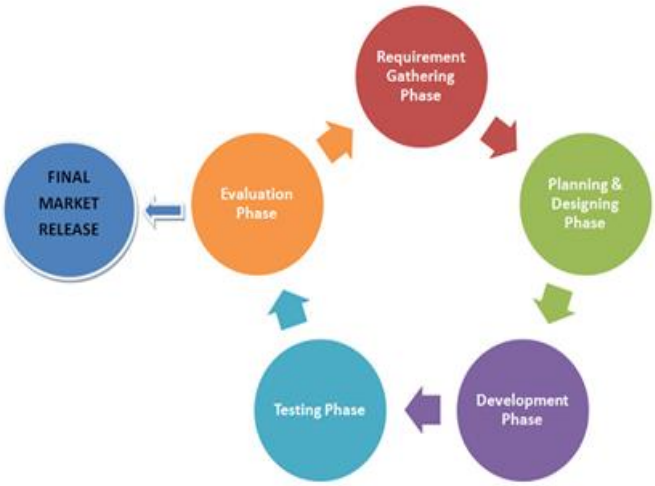
Requirement 9: *Test, communication, and final requirements*

> To integrate ethical requirements into this phase, it is important to make sure that all functions are working as intended, to communicate all relevant facts and limitations to the customer, and to ensure that the system includes ethically required functionality beyond the model (such as a mechanism for human oversight, audibility, or redress). Follow the four-step process established in Requirement 1a.

In Table 2 (at the end of subsection 3.2), we provide an overview of how high-level requirements align with phases in the CRISP-DM model. It is important to note that every value can be actualized in almost any phase, so the table is not meant to be read as presenting an absolute truth. It is meant to provide a reasonable degree of guidance.

3.2 The Agile Model

The Agile model is a response to the traditional plan-driven ('waterfall') approach. A plan-driven approach is not able to adapt to the changing wishes of customers or adjust according to new findings in similar technologies. Once a plan has been made, it needs to be executed in the way it was planned. Agile, on the other hand, has the ability to adjust its plan accordingly. The Agile model is a type of incremental model, which means that several cycles exist within the development of a software, facilitating adaptation to new wishes or desired changes. Because of this ability to adapt, the Agile model accumulates less risk than a plan-driven approach, while still delivering the same value for the customer.[5]



The Figure above provides an overview of the different phases in an iteration. An iteration is also called a 'sprint' in which the software is developed and/or adjusted. At the end of the sprint there is a working model that can be tested. This then provides new insights in how to further adjust the software. One sprint consists of several different phases: requirement analysis, design, develop, test, and discover.

Thanks to these sprints the Agile model is easily adjustable to changing requirements. If it turns out that the client is not pleased with a certain feature of the software, this can easily be adjusted. This dynamic

[5] TryQA, "What Is Agile Model – Advantages, Disadvantages and When to Use It?", *Try QA*, n.d. http://tryqa.com/what-is-agile-model-advantages-disadvantages-and-when-to-use-it/

process allows integration of changing demands from ethical requirements (e.g., relative to new functionality). During the next sprint, problematic ethical issues may be adjusted. Outlined below are the different phases and in what way ethics may be integrated into them. In all phases, stakeholder engagement can be applied.

3.2.1 Phase 1: Requirement Gathering

In this phase, requirements for the final product are analysed. These requirements are based either on the feedback from the client during the evaluation phase, or, in the first iteration, based on rough desires from the client. Potential ethical issues related to these requirements can be identified by applying a process similar to the ethical evaluation process performed in the business understanding phase of CRISP-DM.

3.2.2 Phase 2: Planning & Designing

Although the Agile model does little planning, each iteration does require at least some planning. The planning phase may sketch out how to avoid infringing on potential ethical values.

3.2.3 Phase 3: Development

During this phase of the iteration the product is developed. It is important that the ethical values are kept in mind and integrated during the development process.

3.2.4 Phase 4: Testing

The testing phase tests out the delivered product to see to what extent it meets the desired requirements. In this phase it should be tested whether the product remains ethical or whether it violates (one or multiple) ethical principles.

3.2.5 Phase 5: Evaluation

The evaluation phase has the potential to give a green light to the product. If all requirements are fulfilled it is easy for the developers to release the product to the client, who releases it to the market. While the Agile approach has the potential to adapt to new wishes or find problematic aspects in the product, it also has the potential to lead to "haphazard and harmful creations that are flung into the world before their potential impacts are assessed".[6] It is thus important that in this phase ethical values are evaluated as well as the desired requirements of the client.

[6] Alix, "Working Ethically At Speed", *Medium*, May 7, 2018. https://medium.com/@alixtrot/working-ethically-at-speed-4534358e7eed

	Business Understanding	Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
Human Agency	✓				✓	✓
Liberty	✓				✓	✓
Dignity	✓				✓	✓
Resilience to Attack	✓	✓	✓	✓	✓	✓
Fallback Plan	✓			✓	✓	✓
Accuracy		✓	✓	✓	✓	✓
Reliability		✓	✓	✓	✓	✓
Privacy and DP	✓	✓	✓	✓	✓	✓
Quality & Integrity of Data		✓	✓	✓		✓
Access to Data	✓	✓	✓			✓
Data Rights & Ownership	✓	✓	✓			✓
Traceability		✓	✓	✓	✓	✓
Explainability		✓	✓	✓	✓	✓
Communication	✓					✓
Avoidance & Reduction of Bias		✓	✓	✓	✓	✓
Fairness & Avoidance of Discrimination	✓	✓	✓	✓	✓	✓
Inclusive Stakeholder Engagement	✓				✓	✓
Environmentally Friendly Systems	✓				✓	✓
Individual Wellbeing	✓				✓	✓
Social Relationship & Cohesion					✓	✓
Democracy & Strong Institutions	✓			✓	✓	✓
Auditability	✓	✓	✓	✓	✓	✓
Minimisation & Reporting of Impact		✓	✓		✓	✓
Internal & External Governance		✓	✓	✓	✓	✓
Redress						✓
Human Oversight	✓	✓	✓	✓	✓	✓

Table 2 [Development]: CRISP-DM Model and the Ethical Requirements

4. Specific Operational Ethics Requirements

Following our general discussion of how to apply ethical criteria in the development process, we now turn to more specific ethics operational requirements. While the low-level requirements in this section have been mapped to CRISP-DM, the requirements as such do not depend on the application of the CRISP-DM model, and can be applied with any development method. However, for each requirement, where we mention which phases of the CRISP-DM model are the *most* relevant, that will only be useful if you can map your method onto the CRISP-DM model.

4.1 Human Agency, Liberty and Dignity

It is essential that any technology respects and promotes human liberty and dignity. We recommend the following three sub-requirements:

1. Ensure the protection of the stakeholders' human agency and positive liberty by keeping them informed, ensuring that they are neither deceived nor manipulated, and can meaningfully control the system;
2. Ensure the protection of the stakeholders' negative liberty by ensuring that they have the freedom to use the system and that they are not restrained in functionality and opportunity;
3. Ensure the protection of the stakeholders' human dignity by ensuring that the system is not used to directly or indirectly affect or reduce their autonomy or freedom, and does not violate their self-respect.

1 Human Agency

Requirement 10: *Potential for impact on autonomy.*
In the business understanding and evaluation phases, assess and ensure that:

- evaluation of the end-users' awareness about how the system may impact their autonomy is performed to determine if it is appropriate to make people aware of this impact, and if so, then ensure their awareness (e.g., if an end-user is using the system in a medical capacity, you need to ensure that the functionality of the system and the context in which it is used does not undermine their informed consent to any treatment options);
- the system does not harm individuals' autonomy (i.e., the freedom and ability to make one's own goals and influence the outcomes of those decisions);
- any interference the system has with the stakeholders' decision-making process (e.g., by recommending actions, decisions, or by how it presents stakeholders with options) is justified and minimised.

2 Negative Liberty

Requirement 11: *Fundamental rights*
In all phases, assess and ensure that:

- the system does not interfere with fundamental liberties of users or other stakeholders (including, e.g., freedom of movement, freedom of assembly, and freedom of speech).

3 Human Dignity

Requirement 12: *Respect for Human Dignity.*
In all phases, assess and ensure that:

- the system does not affect human dignity negatively (e.g., by treating individuals as means for other goals, rather than as goals in themselves; by disrespecting individuality, e.g., in profiling and data processing; by objectifying or dehumanizing individuals; or by causing harmful effects on human psychology or identity, e.g., by harming their self-control or their sense of self-worth, which may be rooted in the meaning-creation of various human activities such as work);
- the system is developed to promote human capacity (e.g., by enabling individual self-development) and humans' intrinsic value is respected in the design process and by the resulting system;
- any individual is aware whether they are interacting with an AI, particularly if they are interacting with an autonomous system.

4.2 Technical Robustness and Safety

It is essential that technical systems are robust, resilient, safe, and secure. We recommend the following three sub-requirements:

1. Ensure that the system is Secure and Resilient against attacks;
2. Ensure that the system is Safe in case of failure;
3. Ensure the accuracy, reliability, and reproducibility of the system.

1 Resilience to Attack and Security

Requirement 13: *Security, design, testing, and verification.*
In each phase, assess and ensure that:

- you have evaluated the possible security risks and that the system is protected against cybersecurity attacks both during the design process and when implemented;
- security is implemented into the system's architecture and that the security of the system is tested and, whenever possible, verified before, during, and after deployment;
- security measures are designed to benefit humans.

Requirement 14: *Resilience.*
In each phase, assess and ensure that:

- the system has protection against successful attacks, by assessing possible risks and ensuring extra protection (e.g., safe shut-down) relative to the severity and plausibility of those risks.

2 Fallback Plan and General Safety

Requirement 15: *Safety and verification.*
In the business understanding, modeling, and evaluation phases, assess and ensure that:

- those responsible for the development of the system have the necessary skills to understand how they function and their potential impacts;
- mechanisms to safeguard user safety and protect against substantial risks are implemented;
- the system is tested before, during, and after deployment, to remain safe and secure throughout its lifetime;
- safety measures are designed to benefit humans.

Requirement 16: *Fallback.*
In the business understanding, modeling, and evaluation phases, assess and ensure that:

- if the system fails it does so safely (e.g., by shutting down safely or going into a safe mode).

3 Accuracy, Reliability, and Reproducibility

Requirement 17: *Accuracy, reliability, and effectiveness*
In the data understanding, data preparation, modeling and evaluation phases, assess and ensure:

- the accuracy, reliability, and effectiveness of the system.

Requirement 18: *Reproducibility and follow-up.* In all phases, assess and ensure that:

- the security and safety objectives, results, and outcomes are actively monitored and documented during the design process and, whenever possible, after implementation;
- relevant data are available and reproducible for security and safety audits and/or external evaluations;
- failures and attacks are properly logged to allow for reproducibility and necessary adjustments.

4.3 Privacy and Data Governance

Privacy is an issue in AI- and big data-technology because systems may acquire, interpret, store, combine, produce and/or disseminate personal or sensitive information. This can be information that was entered during the data collection and preparation phases, information that is newly created during the model phase, or information that is recorded during use. Personal or sensitive information can also be at risk because it can be predicted from non-personal or non-sensitive data or information. Personal and sensitive information/data is subject to the General Data Protection Regulation (GDPR) in the EU, and accompanying ethical criteria. This requirement includes four sub-requirements:

1. Ensure the protection of and respect for the stakeholders' privacy;
2. Ensure the protection of the quality and integrity of data;

3. Ensure the protection of access to the data;
4. Ensure the protection of data rights and ownership.

1 Respect for Privacy

Requirement 19: *Clarify roles and responsibilities towards information use, security and privacy.*
In all phases (but especially in business understanding, data understanding, and data preparation), assess and ensure that:

- there are clear and precise descriptions of the roles and responsibilities of users toward information, media and network usage, security, and privacy;
- a common culture is established and encouraged that strongly promotes ethical behaviour for all individuals in the enterprise, and establishes a low tolerance threshold for unethical behaviours.

Requirement 20: *Develop cultures of security and privacy awareness.*
In all phases (but especially in business understanding, data understanding, and data preparation), assess and ensure that:

- a culture of security and privacy awareness is established and encouraged that positively influences desirable behaviour and actual implementation of security and privacy policy in daily practice;
- a validated log is maintained of who has access to any information that could have implications for security or privacy;
- sufficient security and privacy guidance is provided to the developing team during the development process, and to relevant stakeholders both during development and after deployment;
- security and privacy champions are indicated (including C-level executives, leaders in HR, and security and/or privacy professionals) and proactively support and communicate security and privacy programs, innovations and challenges;
- a culture is established and encouraged that facilitates awareness regarding user responsibility to maintain security and privacy practices;
- 'privacy by design' is a core part of the development process and that the end-product abides by these design principles.

Requirement 21: *Personal data use, reduction, and elimination.*
In all phases (but especially in business understanding, data understanding, and data preparation), assess and ensure that:

- alternatives that minimize or eliminate the use of personal data or sensitive data are considered and used whenever possible and, in line with the GDRP, that all personal data held is strictly necessary, reasonable and proportionate for the successful execution of business objectives;
- there are protections against the risk that previously non-sensitive and/or non-personal data may become sensitive or personal (e.g., through the use of aggregation technology).

Requirement 22: *Personal data storage.*
In all phases (but especially in business understanding, data understanding, and data preparation), asses and ensure that:

- any personal data collected is stored and treated with adequate protections, proportionate to the sensitivity of the data stored;
- providers of storage facilities/solutions provide a code of practice for how their network operates and how they store data.

Requirement 23: *Informed consent.*
In the data understanding and data preparation phases, assess and ensure that:

- data containing personal information is only collected if there is informed consent from the data subject or, if not, that there is an alternative legal basis for collecting personal data as set out in Articles 6(1) and 9(2) of the GDPR. Informed consent should include considerations of potential secondary use of data (i.e., use of the data for ends other than the primary end collected), and the potential for the creation of new personal data through (e.g., data set aggregation);
- if the data held are to be used for a secondary purpose (i.e., not envisioned in the original consent agreement), then further informed consent, or an alternative legal basis, is sought.

Requirement 24: *Creation of new personal data.*
In the data understanding, data preparation, and modeling phases, assess and ensure that:

- If needed, further informed consent is acquired (or, if not, that there is an alternative legal basis as set out in Articles 6(1) and 9(2) of GDPR) for the creation of new personal or sensitive information/data (e.g., through estimation of missing data, the production of derived attributes and new records, data integration, or aggregation of data sets);
- all newly created personal or sensitive information/data is given at least the same protection and attracts the same rights as previously collected or held personal or sensitive information/data.

Requirement 25: *Subsequent collection and/or creation of new personal data*.
In the data understanding, data preparation, and modeling phases, assess and ensure that:

- no new personal information is or can be collected or created during regular use of the system, unless necessary (e.g., for the function of the system or realization of the business objectives);
- if new personal information is collected or created, then limitations are properly imposed to protect individuals' privacy or sensitive information/data, and further informed consent is acquired, if needed.

Requirement 26: *Privacy awareness*.
In the deployment phase, assess and ensure:

- mechanisms allowing developers and users to flag issues related to privacy or data protection in the system's processes of data collection (including for training and operation) and data processing;
- mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable).

Requirement 27: *Data review and minimization*.
In the data understanding, data preparation, modeling, and deployment phases, assess and ensure that:

- consideration is given to develop the system or train the model with or without minimal use of potentially sensitive or personal data, and applied whenever possible (note that it is questionable whether any data is ever fully anonymized—see Requirement 34);
- potential measures to protect or enhance privacy (e.g., through encryption, anonymization, aggregation, or deletion) are used when possible and proportionate to the risk;
- an oversight mechanism is established for data collection, storage, processing, and use.

Requirement 28: *Alignment with existing standards*.
In every phase, assess and ensure that:

- the system is aligned with relevant and appropriate standards (e.g., ISO, IEEE) and/or widely adopted protocols for daily data management and governance.

Requirement 29: *Data Protection Officers.*
In all phases, ensure that:

- a Data Protection Officer (DPO), where one exists, is adequately involved in the development process.

2 Quality and Integrity of Data

Requirement 30: *Oversight of data quality*.
In the data understanding, data preparation, and modeling phases, assess and ensure that:

- there are processes to ensure the quality and integrity of all pertinent data, including means of verifying that data sets have not been compromised or hacked (if you are in control of the quality of the external data sources used, to assess to what degree you can validate their quality);
- a culture of shared responsibility for the organization's data assets is established and encouraged;
- the potential value of data assets is acknowledged, and that roles and responsibilities are clear for governance and management of data assets;
- the impact and risk of data loss is continuously communicated;
- employees understand the true cost of failing to implement a data quality culture.

Requirement 31: *Employment of protocols and procedures for data governance*.
In the business understanding, data understanding, and data preparation, assess and ensure that:

- appropriate protocols, processes, and procedures are followed to manage and ensure proper data governance;
- there are reasonable safeguards for compliance with relevant protocols, processes and procedures for your industry.

3 Access to Data

Requirement 32: *Oversight of access to data*.
In the business understanding, data understanding, and data preparation, assess and ensure that:

- persons who can access particular data under particular conditions are qualified and required to access the data, and that they have the necessary competence to understand the details of the data protection policy;
- there is an embedded oversight mechanism to log when, where, how, by whom, and for what purpose data was accessed, as well as for data collection, storage, processing, and use.

Requirement 33: *Availability of data*.
In the business understanding, data understanding, and data preparation, assess and ensure that:

- personal data is available to those to whom the data relate and that this process protects other individuals' privacy (e.g., through linking individual data to the informed consent process—see Requirement 23);
- there is an embedded process that allows individuals to remove their data from the system and/or correct errors in the data where these occur, and ensure that this process is available at any stage in the process (note that once data is correctly and fully anonymized it is no longer considered personal data, although there may be potential for re-identification through aggregation of data sets).
- if previously anonymized data is re-identified (see Requirements 24 and 25), then these data are made available once more (note, however, that it is questionable whether any data is ever fully anonymized—see Requirement 34).

Requirement 34: *Protection against re-identification.*
In the deployment phase, assess and ensure that:

- appropriate measures are in place to protect against de-anonymization or re-identification (de-anonymized or re-identification can be achieved, e.g. by linking to other possibly available data).

4 Data Rights and Ownership

Requirement 35: *Clarity on ownership of data.*
In the business understanding, data understanding, and data preparation, assess and ensure that:

- where the prevailing laws on ownership of personal data are unclear, ambiguous, or insufficient, that the ownership of the data and data sets are clear in any agreements with the providers of such data;
- the ownership of personal or sensitive information/data is clarified to the relevant party in the process of gathering informed consents (Requirement 24);
- agreements stipulate what the owner and (end-)users of the data are permitted to do with those data.

4.4 Transparency

The amount of transparency needed for a system is a function of (1) the severity of potential impacts of decisions taken or recommended by the system on humans and society; and (2) the importance of accountability for system errors or failures. Accountability is, for example, crucial in cases of systems that can strongly affect the rights and wellbeing of individuals. It allows them to get redress. The requirement of transparency is closely related to the requirement of accountability, in this regard. The requirement of transparency includes three sub-requirements:

1. Ensure that the system has a sufficient level of Traceability;
2. Ensure that the system has a sufficient level of Explainability;
3. Ensure that the relevant functions of the system are Communicated to stakeholders.

Note: The importance of transparency depends on the potential of a system to harm stakeholder interests or rights and the importance of redress. If a system performs harmless tasks, then it need not be transparent. But if a system can harm people, and especially if they should be able to appeal decisions made by a system, then this requires understanding and so transparency is more important (e.g., for systems that recommend punishments in the legal system).

1 Traceability

Requirement 36: *Traceability measures*.
In the data understanding, data preparation, modeling, and evaluation phases, assess and ensure that:

- measurements to ensure traceability are established through the following methods:
    - Methods used for designing and developing systems (rule-based AI systems: the method of programming or how the model was built; learning-based AI systems:

the method of training the algorithm, including which data was gathered and selected, and how this occurred);
    - Methods used to test and validate systems (rule-based AI systems: the scenarios or cases used in order to test and validate; learning-based model: information about the data used to test and validate);
    - Outcomes of the system (outcomes of or decisions taken by the system, as well as potential other decisions that would result from different cases, e.g., for other subgroups of users);
    - A series of technical methods to ensure traceability should be taken (such as encoding the metadata to extract and trace it when required). There should be a way of capturing where the data has come from, and the ability to construct how the different pieces of data relate to one another.

Requirement 37: *Responsibility for Traceability*.
In every phase, assess and ensure that:

- there is a "human in control" when needed, especially when the system may cause harmful outcomes (e.g., an AI playing a game like chess, which may have no harmful outcomes, would not necessarily require a human in control, unless there was the potential for negative effects);
- a balanced prioritisation for human control, related to the plausibility and/or severity of the outcome;
- there are measures to enable audit and to remedy issues related to governing the system and allow organisations using your technology the ability to identify when there is an issue or harm, and the ability to prevent these issues from occurring, and stop it when these issues are identified;
- there are appropriate remedial steps for detection and response mechanisms if something goes wrong, by closely liaison with the organisational user, or end-user.

2 Explainability

Requirement 38: *Training data*.
In the data understanding, data preparation, modeling, and evaluation phases, assess and ensure that:

- if possible, you can analyse your training data, that your data is representative, and value aligned;
- whenever possible, there is an ability to go back to each state the system has been in to determine or predict what the system would have done at time t and, whenever possible, determine which training data was used.
- in the event of a system malfunction or harm resulting from the system, as much transparency as is possible of your training data is made available, without violating privacy, to the appropriate authorities.

Requirement 39: *Explainable systems*.
In the data understanding, data preparation, modeling, and evaluation phases, assess and ensure that:

- you know to what degree the decisions and outcomes made by the system can be understood, including whether you have access to the internal workflow of the model;
- explainability is guaranteed (through technologies such as Explainable AI), when there is a greater emphasis within its use for explainability over performance, or when there is no trade-off between explainability and performance.

Requirement 40: *Explanations of rationale*.
In every phase, assess and ensure that:

- whenever possible, the process of, and rationale behind, the choices made by the system are explainable upon request to an organisational user and/or auditing body in situations where there is a potential and/or existent harm;
- the reasons for the collection and use of particular data sets are explainable upon request to organisational users and/or auditing bodies;
- in situations where the system-development organisations provide these technologies directly to the end-user, there is redress and explanations of how the system arrived at those decisions, if there is harm caused to the end-user by the system's decisions;
- decisions made about individuals are understandable in colloquial language terms for an ordinary (end-)user or stakeholder (e.g., 'You have been put into this category because of x, y, and z').

Requirement 41: *Trade-offs*.
In every phase, assess and ensure that:

- trade-offs between explainability/transparency and best performance of the system are appropriately balanced based on the systems context of application (e.g., in healthcare the accuracy and performance of the system may be more important than its explainability; whereas, in policing, explainability is much more crucial to justify behaviours and outcomes of law enforcement; and in other areas, such as recruitment, both accuracy and explainability are similarly valued).

3 Communication

Requirement 42: *Communication regarding interactions with the system.*
In the business understanding and deployment phases, assess and ensure that:

- it is communicated to, and presumably understood by, the (end-)users or other affected persons that they are interacting with a non-human agent and/or that a decision, content, advice or outcome is the result of an algorithmic decision, in situations where not doing so would be deceptive, misleading, or harmful to the user.

Requirement 43: *Communication with stakeholders*.
In the business understanding and deployment phases, assess and ensure that:

- a culture is established and encouraged in which open and structured communication is provided to stakeholders, in line with their requirements (including organisational users and end-users, if you are dealing directly with them).

- information to stakeholders, (end-)users, and other affected persons, about the system's capabilities and limitations, is communicated in a clear, understandable, and proactive manner, that enables realistic expectation setting;
- it is clear to stakeholders, (end-)users, and other affected persons, what the purpose of the system is and who or what may benefit from the product/service;
- usage scenarios for the product are specified and clearly communicated so that they are understandable and appropriate for the intended audience;
- in cases where stakeholders cannot be provided with certain data and answers, there is a full disclosure of that limitation, why there is a limitation, and also what they themselves do and do not know.

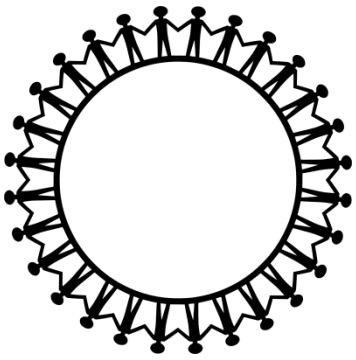
Requirement 44: *Communication within user and stakeholder community*.
In the business understanding and deployment phases, assess and ensure that:

- a culture is established and encouraged based on mutual trust, transparent communication, open and understandable terms, a common language, ownership, and accountability;
- an explanation, which all reasonable users and stakeholders can presumably understand, is given as to why the system took a certain choice resulting in a certain outcome;
- mechanisms are established to inform organisational users and end-users (if dealing directly with them) about the reasons and criteria behind the system's outcomes and, in collaboration with users, establish processes that consider users' feedback and use this to adapt the system;
- any potential or perceived risks are clearly communicated to the (end-)user (e.g., consider human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue).

4.5 Diversity, Non-discrimination, and Fairness

This requirement is important to prevent harmful discrimination against individuals or groups in society owing to a lack of diversity in the development process, in training data sets or in the parameters of algorithms used. It also aims to take a proactive approach and proposes that developers of these systems should aim to do good with their systems in relation to fairness, diversity, and non-discrimination. We distinguish three sub-requirements:

1. Ensure the avoidance of discrimination; and reduction of harmful bias;
2. Ensure fairness and diversity;
3. Ensure the inclusion and engagement of stakeholders.



Note: There are forthcoming standards on algorithmic bias from IEEE and ISO that will detail practical procedures for avoiding algorithmic bias on a more detailed level than is possible here.

1 Avoidance and Reduction of Harmful Bias

Requirement 45a: *Bias assessment in Planning.*

In the business understanding phase, assess and ensure that:

- the potential for harmful bias in the business understanding and requirements stage is evaluated and, if possible, avoided (e.g., some requirements may inadvertently favour particular groups in society over others, e.g., if you are using the system to hire a new candidate, there may be more gender- or ethnicity-specific characteristics entered into the criteria for assessment, which would have negatively biased results);
- developing teams receive unconscious bias training to assist developers to identify innate biases during the development of systems.

Requirement 45b: *Bias assessment in data analysis.*

In the data understanding phase, assess and ensure that:

- an evaluation is performed to determine the diversity and representativeness of users in the data, testing for specific populations or problematic use cases is performed, and that input, training, and output data is analysed for harmful bias;
- the potential for harmful bias in the data understanding stage is evaluated (e.g., some data sets may contain harmful biases if they consist solely of the behaviour of subclasses of all people, e.g., young white men, and if the system is deployed in situations where groups other than those in the data set will be affected) and, if possible, avoided (e.g., incorporate additional users' data that is not included in the data; look at the alternative or additional supply chains from the data that you are using; or in some cases, the datasets need to be discarded altogether).
- data from just one class is not used to represent another class, unless it is justifiably representative.

Requirement 45c: *Bias assessment in data preparation.*

In the data preparation phase, assess and ensure that:

- the potential for harmful bias in the data preparation stage is evaluated and, if possible, avoided (e.g., the cleaning of the data set may inadvertently remove data relating to certain minority or under-represented groups, leaving the data set as a whole biased);
- you have clearly established what kind of sample you need, what kind of sample you have taken, and that you articulate what it will be used for.

Requirement 45d: *Bias assessment in modeling.*

In the modeling phase, assess and ensure that:

- the potential for harmful bias in the modeling stage is evaluated and, if possible, avoided (e.g., some algorithms make assumptions about universal behaviours and characteristics which are untrue; many behaviours which are assumed to be universal are in fact culturally specific);
- a strategy or a set of procedures is established to avoid creating or reinforcing unfair bias in the system regarding the use of input data as well as for the algorithm's design, and that the strategy includes an assessment of the possible limitations stemming from the composition of the used data sets;

- there is in the design process an awareness of cultural bias to prevent or exacerbate any potential harmful bias.

Requirement 46: *Engagement with users to identify harmful bias*.
In the business understanding, evaluation, and deployment phases, assess and ensure that:

- a mechanism allows others to flag issues related to harmful bias, discrimination, or poor performance of the system and establish clear steps and ways of communicating on how and to whom such issues can be raised (i.e,. during the design, development, and deployment of the system);
- there is transparency about how the algorithms may affect individuals to allow for effective stakeholder feedback and engagement;
- the implementation of methods for redress and feedback from users at all stages of the system's life-cycle.

Requirement 47: *Anticipating harmful functional bias*.
In every phase, assess and ensure that:

- whenever possible, the potential of the system being used for harmful or illegal purposes is avoided, and that if the system can be used for unintended purposes, then consider potential implications of this likelihood and develop mitigation procedures in the event of potential ethical issues arising;
- the system is not designed for bad purposes and attempt to eliminate, whenever possible, ways that they can be misused (one way to do this is to use tried-and-tested general models, rather than building all models from scratch).

Requirement 48: *Decision variability*.
In the evaluation and deployment phases, assess and ensure that:

- a measurement or assessment mechanism, of the potential impact of decision variability on fundamental rights, is established based on an evaluation of the system's possibility for decision variability that can occur under the same conditions;
- variability is explained to the organisational user of the system and/or the end-user (if they are using it directly). For example, in medicine this should be explained to doctors that use it.

Requirement 49: *Avoiding harmful automation bias.*
In every phase, assess and ensure:

- an appropriate level of human control for the system (by including respective task allocations between the system and humans for meaningful interactions and appropriate human oversight and control);
- safeguards are embedded to prevent overconfidence in or overreliance on the system through education and training to be more aware of harmful bias in the system.

2 Ensuring Fairness and Diversity

Requirement 50: *Accessibility and Usability*.
In every phase (but especially in the business understanding and evaluation phases), assess and ensure that:

- the system is understandable and accessible to users of assistive technologies, users with special needs or disabilities, or groups otherwise at risk of exclusion;
- the system is usable by users of assistive technologies, users with special needs or disabilities, or groups otherwise at risk of exclusion (or if the system cannot be *used* properly, attempt to make improvements and ensure that any limitations are fully understood by these groups);
- you seek feedback from teams or groups that represent different backgrounds and experiences (including but not limited to users of assistive technologies, users with special needs, or disabilities), and that this process should be accommodating to include different variations and users;
- no persons or groups are disproportionately negatively affected by the system. Or if that cannot be ensured, then attempt to minimize the negative effects and ensure that these people and groups fully understand these negative effects before using the system, and that those at risk of being negatively affected are adequately represented in the design process by including feedback from those likely to be affected in the design of the system.

Requirement 51: *Intended use.*
In the modeling and evaluation phases, assess and ensure that:

- to the degree it is possible, function of the algorithm is appropriate (including legal compliance and risks) relative to an evaluation of the reasonability and unreasonability of the systems' inferences about individuals beyond bias.

Requirement 52: *Review process*.
In every phase (but especially the evaluation phase), assess and ensure that:

- knowledgeable professionals, both internal and external to the company, examine the development process and the product through a risk assessment procedure.

Requirement 53: *Distributing the system to organisational users.*
In the deployment phase, assess and ensure that:

- the user interface is clearly presented, including information about potential errors and the accuracy of the system (including the underlying certainty).

Requirement 54: *Whistleblowing*.
In every phase, assess and ensure:

- a process that enables employees to anonymously inform relevant external parties about unfairness, discrimination, and harmful bias, as a result of the system;
- that individual whistleblowers are not harmed (physically, emotionally, or financially) as a result of their actions.

3 Inclusionary Stakeholder Engagement

Requirement 55: *Diversity.*
In every phase (but especially in the business understanding and evaluation phases), assess and ensure:

- a process to include the participation of different stakeholders in the development, use, and review of the system;
- that efforts are made so that a wide diversity of the public, including different sexes, ages, and ethnicities, are represented;
- that this is applied within the organization, by informing and involving impacted workers and their representatives in advance.

Requirement 56: *Inclusion*.
In every phase of development, assess and ensure:

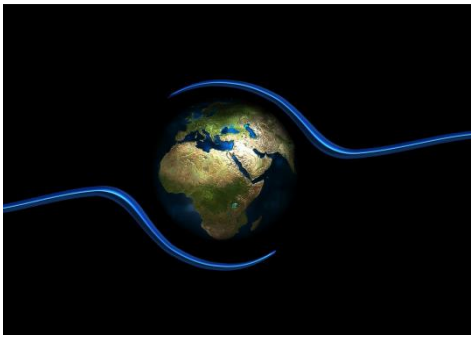
- an adequate inclusion of diverse viewpoints during the development of the system;
- that development is based on an acknowledgement that different cultures may respond differently, have different thought processes and patterns, and express themselves differently.

4.6 Individual, Societal, and Environmental Wellbeing

It is important that any system seeks to maximise positive benefits to society and the environment while limiting any potential harm as much as possible. We suggest the following four sub-requirements:

1. Ensure that the system promotes sustainability and environmentally friendliness;
2. Ensure the protection of individual wellbeing (including the development of human capabilities and access to social primary goods, such as opportunities for meaningful paid work);
3. Ensure the protection of societal wellbeing (the technology supports and does not harm rich and meaningful social interaction, both professionally and in private life, and should not support segregation, division and isolation); and
4. Ensure the protection of democracy and strong institutions to support democratic decision-making.

Note: Because wellbeing interacts with and depend on other values (such as autonomy and dignity), organisations need to ensure individual wellbeing through the promotion of all of the values outlined in the guidelines.



1 Sustainable and Environmentally-friendly Systems

Requirement 57: *Environmental impact*.
In the business understanding, evaluation, and deployment phases, assess and ensure:

- a mechanism to measure the ecological impact of the system's use (e.g., the energy used by data centres).
- where possible, measures to reduce the ecological impact of your system's life cycle;
- an adherence to resource-efficiency, sustainable energy-promotion, the protection of the non-human living world around us, and the attempt to ensure biodiversity and the healthy functioning of ecosystems (in particular, decisions made by the system that will directly affect the non-human world around us need to be carefully factored in, with strong emphasis on the impact on these ecological externalities, through a holistic ecosystem-focused outlook);
- transparency about ecological impact and, if possible, work with environmental protection organisations to ensure that the system is sustainable, and keep the ecological footprint proportionate to the intended benefit to humanity.

2 Individual Wellbeing

Requirement 58: *Individual wellbeing assessment*.
In the evaluation and deployment phases, assess and ensure that:

- the system is evaluated for its likely and potential impact on individual wellbeing (including consideration of the way in which the system will or could be used which may be detrimental to users or stakeholders). Particular care should be taken for detriments towards vulnerable groups through discussion with them, rather than assuming their needs.

Requirement 59: *Emotional attachment*.
In the evaluation phase, assess and ensure that:

- if the system is developed to interact directly with humans, evaluate whether it encourages humans to develop unwanted attachment and unwanted empathy towards the system or detrimental addiction to the system, and if so take appropriate action to minimize such effects;
- the system clearly communicates that its social interaction is simulated and that it lacks human capacities such as "understanding" and "feelings";
- the system does not make humans believe it has consciousness (e.g., through expressions that simulate emotions).

3 Societal Wellbeing

Requirement 60: *Societal impact assessment.*
In the evaluation phase, assess and ensure that:

- the system's likely and potential impact on social relationships and social cohesion (including consideration of the way in which the system will or could be used which may be detrimental to groups of users or groups of stakeholders) is not inappropriate;

- social benefits are determined through social metrics, not simply measurements in terms of GDP (e.g., liveability indexes).

Requirement 61: *Engagement with stakeholder community*.
In the evaluation and deployment phases, assess and ensure that:

- the broader societal impact of the AI system's use beyond the individual (end-)users (such as potentially indirectly affected stakeholders) is evaluated;
- the social impacts of the system are well understood (e.g., assess whether there is a risk of job loss , deskilling of the workforce, or changes to occupational structure) and record any steps taken to counteract such risks;
- a culture is established and encouraged to ensure timely communication of IT change requests to affected groups, and consult the affected groups regarding implementation and testing of changes;
- stakeholders are involved throughout the system's life cycle, and foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.

4 Democracy and strong institutions

Requirement 62: *Mitigation of impacts on democracy*.
In the evaluation and deployment phases, assess and ensure:

- an evaluation of whether the system is intended, or could be used for, supporting, organizing or influencing political processes, including political messaging and communication, and if so, take measures to ensure that the system supports democratic processes and protects against interventions that manipulates, misleads or excludes voters and distorts democratic processes;
- compliance with higher authorities of AI development and implement an ethical officer to ensure corporate social responsibility within the company;
- that external ethics audits are carried out to guarantee that system development is not harming democratic processes.

4.7 Accountability

Any system, and those who design it, should be accountable for the design and impact of the system. We identify five sub-requirements here:

1. Ensure that systems with significant impact are designed to be auditable;
2. Ensure that negative impacts are minimised and reported;
3. Ensure internal and external governance frameworks;

4. Ensure redress in cases where the system has significant impact on stakeholders;
5. Ensure human oversight when there is a substantial risk of harm to human values.

Note: accountability may also relate to IT governance, not just IT management, since boards of directors have final accountability and may want to assure proper accountability at lower levels.

1 Auditability

Requirement 63: *Engagement and reporting.*
In every phase, assess and ensure that:

- incidents are identified and reported on a correct and timely basis and implement appropriate internal and external escalation paths;
- incidents are responded to and resolved immediately;
- a culture of proactive problem management (detection, action and prevention), with clearly defined roles and responsibilities, is established and encouraged;
- a transparent and open environment for reporting problems is established and encouraged, by providing independent reporting mechanisms and/or rewarding people who bring problems forward;
- there is an awareness of the importance of an effective control environment;
- a proactive risk- and self-aware culture is established and encouraged, including commitment to self-assessment, continuous learning, and independent assurance reviews;
- auditability is built into the system;
- performance indications are identified and regularly report on the outcomes, in relation to the auditing system.

Requirement 64: *Compliance as culture*.
In every phase, assess and ensure that:

- a compliance-aware culture is established and encouraged, including disciplinary procedures for noncompliance with legal and regulatory requirements;
- a culture that embraces internal audit, assurance findings, and recommendations (based on root cause analysis) is established and encouraged;
- leaders take responsibility to ensure that internal audit and assurance are involved in strategic initiatives and recognize the need for (and value of) audit and assurance reports;
- mechanisms that facilitate the system's auditability (such as ensuring traceability and logging of the AI system's processes and outcomes);
- in applications affecting fundamental rights (including safety-critical applications) the system can be audited independently;
- the developing team attempts to learn to avoid situations requiring accountability in the first place, by ensuring ethical best practices.

Requirement 65: *Code of ethics*.

In all phases (but starting in the business understanding phase), assess and ensure that:

- an ethical culture of internal auditing through an appropriate code of ethics, or clear appeal to widely accepted industry standards, is established and encouraged;
- a code of ethics exists, which identifies accountability structures, encourages regular auditing for ethical assurance and improvements, and has accountability procedures to ensure that the code of ethics is being followed.

2 Minimising and reporting negative impacts

Requirement 66: *Reporting Impacts*.

In the business understanding, evaluation, and deployment phases, assess and ensure that:

- a risk assessment is conducted, which takes into account different stakeholders that are (in)directly affected by the system and the likelihood of those impacts;
- training and education is provided to help develop accountability practices (including teachings of the potential legal framework applicable to the system);
- if possible, that an 'ethical AI review board' or a similar mechanism is established to discuss overall accountability and ethics practices, including potentially unclear grey areas;
- processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks, or biases in the system, is established.

Requirement 67: *Minimising negative impact*.

In the business understanding, evaluation, and deployment phases, assess and ensure:

- a process for minimisation of negative impacts (such as external guidance and/or an auditing processes to oversee ethics and accountability), in addition to internal initiatives;
- that audit controls are built into the system to check performance, record decisions made about the purpose and functioning of the system (including reporting on the impacts in general, not just occurrences of negative impacts);
- an attempt to predict the consequences/externalities of the system's processing.

3 Internal and External Governance Frameworks

Requirement 68: *Impact on business*.

Assess and ensure that:

- there is an ability to evaluate the degree to which the system's decision influences the organisation's decision-making processes, why this particular system was deployed in this specific area, and how the system creates value for the organization and the general public;
- a clear rationale is established by your organization about why you are designing and creating the system, and the intended purpose that it will serve.

Requirement 69: *Identify interests and values at risk*.
In the evaluation and deployment phases, assess and ensure:

- a mechanism to identify relevant interests and values implicated by the system and potential trade-offs between them, before deployment and during the life-cycle of the system, which should include considerations regarding how trade-offs were decided and documented;
- the establishment of values and interests at risk, through stakeholder analysis, product testing, discussion groups, external workshops, and a range of diversity and inclusion sessions.

Requirement 70: *Install systems to allow for internal complaint*.
In the evaluation and deployment phases, assess and ensure:

- the existence and advertisement (through the companies) of a clear complaints and whistleblowing system (directing employees to a suitable contact venue and setting out the process for registering both anonymous and identifiable complaints);
- that employees are aware of a zero tolerance policy for any recriminations for whistleblowing or the registering of internal complaints.

Requirement 71: *Internal Auditor*.
In the evaluation phase, assess and ensure that:

- the internal auditor(s) within the company is audited to guarantee that it is not abusing their role within the organisation;
- an internal ethics advisor has the same degree of independence and security as is now envisaged for the DPO under GDPR. Alternatively (or in addition) we encourage organisations to develop sectoral solutions (e.g., an ethics council for their sector; startups and microbusinesses may not have the resources to put an ethicist on the payroll, so an alternative, such as Ethics-as-a-Service or external ethics auditing, may be implemented instead).

4 Redress

Requirement 72: *Redress mechanisms*.
In the deployment phase, assess and ensure that:

- the contextual meaning of accountability is clear for different roles in the development chain (e.g., data scientists, CDOs, board members, business managers), including what form of sanctions are in place for whom, and which roles should take personal responsibility, with redress mechanisms in case of negative impacts;
- a set of mechanisms that allows for redress in case the occurrence of any harm or adverse impact is established;
- where possible, embed mechanisms to provide information to (end-)users/third parties about opportunities for redress.

5 Human Oversight

Requirement 73: *Avoiding automation bias.*
In the data understanding, data preparation, modeling, evaluation, and deployment phases, assess and ensure:

- an appropriate level of human control for the system and use case, including respective task allocations between the system and humans for meaningful interactions and appropriate human oversight and control;
- safeguards are embedded to prevent overconfidence in or overreliance on the system for work processes.

Requirement 74: *Responsibility*.
In all phases, assess and ensure that:

- the "human in control", and the moments or tools for human intervention, are clearly identified;
- there are measures to enable audit and to remedy issues related to governing AI autonomy;
- there is a human-in-the-loop to control the system, to ensure and protect the autonomy of human beings;
- detection and response mechanisms are appropriate in the event of something going wrong.

5. Special Topics for Consideration

This section gives an overview of ethical issues concerning specific types of data, functions, techniques, systems, and application areas. For each section it presents a number of requirements to be taken, complimentary to the requirements provided in section 3 and 4.

5.1 Processing of images, video, speech and textual data

The recording, processing, and analysis of images, video feeds, speech and texts raise special ethical issues, especially when these media represent persons and their behaviours. Speech and text are studied and analysed in the field of Natural Language Processing (NLP). The field of computer vision is concerned with the analysis of images and video feeds. Both fields nowadays heavily involve machine learning techniques. These fields can involve special issues of privacy and fairness that need to be considered. First, it is possible through analytics methods to uncover or conjecture personal information of the speaker, author or depicted person, including socio-economic categories such as age, gender and ethnicity, but also possibly social class, sexual orientation, health, mood, and other forms of personal information. They could also be used for identification. Analytics in these fields are therefore potentially privacy-invasive, and also involve conjectures that may turn out to be false but could nevertheless be the basis of subsequent actions. Another concern lies in possible bias. It has been shown, for example, that some video analytics techniques result in much higher fault rates for women than for men or for people of colour as compared to white people. Tagging of persons and situations may also be prejudicial, as when a fast-moving person is labelled as a potential criminal.

Requirements:

- Investigate whether the system produces, intentionally or unintentionally, new personal information, especially concerning socioeconomic qualities, moods, behaviours, intentions, personality, and identity. If so, determine whether this new information is needed, how sensitive or potentially harmful it is, whether it requires informed consent, whether it is sufficiently warranted based on the available evidence, and whether its use can be limited to intended applications. Take appropriate measures to protect privacy;
- Investigate whether the system contains algorithmic bias in its depiction of social groups, in containing disproportionate error rates for certain social groups, in over- or underrepresenting certain social groups, or in providing less functionality for certain social groups.

5.2 Merging of Databases

The combination of different sets of information may disclose sensitive information that violates privacy when the different sets are put together. This is a potential risk of merging databases. It may reveal new personal information, and it may lead to identification that was previously not possible. Data mining techniques may deanonymize anonymized data and create new personal information that was not contained in the original data set. If data subjects gave informed consent for the processing of personal information in the original data sets for particular purposes, they did not necessarily by extension also give permission for the merging of data sets and for data mining that reveals new information. New information produced in this way may also be based on probabilities or conjectures, and therefore be false, or contain biases in the portrayal of persons.

Requirements:

- Establish or adopt an explicit protocol to determine what is fair use of an individual's data, particularly relating to its use during database merging;
- Identify what new personal information is created, whether this new information is needed, how sensitive or potentially harmful it is, whether it requires informed consent, whether it is sufficiently warranted based on the available evidence, and whether its use can be limited to intended applications. Take appropriate measures to protect privacy;
- Consider whether the newly produced information is biased in its depiction of social groups, in containing disproportionate error rates for certain social groups, in over- or underrepresenting certain social groups, or in providing less functionality for certain social groups;
- Different guidelines may be needed for data that is used in the public interest and data that is used commercially.

5.3 Systems that make or support decisions

AI systems sometimes merely produce information, but at other times they either make or recommend decisions that then lead to consequences in the actual world. Embedded AI, AI embedded in software or hardware systems, allows such systems to operate autonomously to make their own decisions and perform their own actions. It may, for example, drive a robot to autonomously select and shoot at a target, or a self-driving car to choose what trajectory to follow when a crash is unavoidable. Other systems merely recommend decisions to be made by human beings. This particularly applies to decision support systems, which are information systems that support organizational decision-making. They usually serve higher and middle management.

Systems that make or support decisions raise special issues about responsibility: who is responsible for the decisions that are subsequently carried out? Another worry is transparency and explainability: how can people still understand the grounds or reasons for the decisions that are made? Relatedly, how can meaningful human control be maintained, if at all, for systems that operate (semi)autonomously? These systems also raise special issues about autonomy: to what extent are people still autonomous if machines make decisions for them? There are also corresponding concerns about safety and accuracy.

Requirements:

- For fully autonomous systems, consider whether they can be justified based on considerations of responsibility, transparency, autonomy, safety and accuracy, and meaningful human control;
- For decision-support systems, make the same consideration, taking into account the division of labour between the machine and the human user. Does the machine ultimately support human decisions that are still autonomously taken, or do human users tend to unquestioningly follow the recommendations of the machine?
- For fully autonomous systems, do risk assessments implement clear procedures of what they can and cannot do, do proper testing, and take proper precautions to ensure safety?

5.4 Tracking, behaviour analytics, facial recognition, biometrics and surveillance

In the Ethics Guidelines report of the High-Level Expert Group on AI, the identification and tracking of individuals using AI is mentioned as a critical concern, especially when this is done in mass surveillance. It considers involuntary and automated methods of identification used by public and private entities, including facial recognition, automated voice detection, and other biometric and behavioural detection methods, and the tracking and tracing of individuals across different locations. AI can be used, amongst others, to identify voices in a crowd,[7] lip-read what individuals are saying,[8] track people's activities across space,[9] and recognize people through gait recognition or facial recognition.

Although there are legitimate and important applications of automated identification and tracking, there are ethical problems with using these techniques for targeted or mass surveillance, because of possible negative implications for privacy, autonomy, liberty and fairness. Uses beyond law enforcement (e.g., tracking consumers and employees) are morally controversial because they often do not have the public's interest in mind. But also, law enforcement applications may be morally problematic (cf. the Chinese social credit system). On a societal level, surveillance techniques risk creating the self-fulfilling prophecy: locations where more crime is detected will be monitored more thoroughly, identifying more crime, thus resulting in the placement of even more surveillance technologies. On an individual level, people may experience a chilling effect, and people (including) criminals may be lead to adopt behaviours considered "normal" by the standards of the system. These technologies can also contain biases that disadvantage certain social groups.

Requirements:

- Identify what new personal information is created or processed, whether this new information is needed, how sensitive or potentially harmful it is, whether it requires informed consent, whether it is sufficiently warranted based on the available evidence, and whether its use can be limited to intended applications. Take appropriate measures to protect privacy;
- Investigate whether the system contains algorithmic bias in its depiction of social groups, in containing disproportionate error rates for certain social groups, in over- or underrepresenting certain social groups, or in providing less functionality for certain social groups.

5.5 Processing of medical data

As systems are deployed through various devices (from sensors to RFID chips and video feeds), diagnostic data (images, blood tests, vital signs monitors) as well collected from structured and unstructured data sources (from consultation notes to patient prescriptions and payment records), the amount of data that healthcare professionals and data companies have at their disposal necessitates attention. With applications in early disease detection, identifying the spread of diseases as well as development of

[7] Tung, Liam, "Google AI Can Pick out a Single Speaker in a Crowd: Expect to See It in Tons of Products", *ZDNet*, April 13, 2018. https://www.zdnet.com/article/google-ai-can-pick-out-a-single-speaker-in-a-crowd-expect-to-see-it-in-tons-of-products/

[8] Condliffe, Jamie, "AI Has Beaten Humans at Lip-reading", *Technology Review*, November 21, 2016. https://www.technologyreview.com/s/602949/ai-has-beaten-humans-at-lip-reading/

[9] Kitchin, Rob, "Getting smarter about smart cities: Improving data privacy and data security", Data Protection Unit, Department of the Taoiseach, Dublin, Ireland, 2016, p. 5.

healthcare robotics and wearables, developers need to be aware of a number of issues that can emerge from the use of AI and big data systems in the healthcare domain, especially with regard to medical data.

The aim of most AI and big data systems in the domain of medicine is to make a transition from population-based healthcare to personalised medicine programs, by using the various data sources, data collecting devices, and data analytics to make medical recommendations using each patient's data records. This is becoming possible as medical records contain data including demographic information, information from laboratory tests, imaging and diagnostics data, as well as clinical notes and prior interventions.[10] Companies that offer storage, analysis and processing of biomedical information include Amazon Web Services, Cisco Healthcare Solutions, DELL Healthcare Solutions, GE Healthcare Life Sciences, IBM Healthcare and Life Sciences, Intel Healthcare, Microsoft Life Sciences and Oracle Life Sciences.[11] The increasing involvement of data processing and storage companies that have access to patient information invites a number of ethical concerns that developers need to be aware of.

As patient information becomes transferred across different hospitals and data companies, the security and privacy of this data needs to be ensured at each stage/site of transfer.[12] This means that while for processing purposes greater interconnection may mean better analysis, from an ethical standpoint this interconnectivity presents two further points of concern: firstly, a weakness in one site/stage may carry over to other sites/stages, and secondly, increased interconnectivity can make it more difficult to identify which parties access data and at what point in time patient data is made use of. These points of concern can lead to reduced traceability and accountability, as well as the viability of patients having sufficient information to consent to who has access to their data, and knowledge of where their data is being stored/processed. Moreover, while patient information may appear anonymized through aggregation, re-identification techniques can be used without patients being informed,[13] especially if the data is of high research or public health importance.

Requirements:

- Determine what medical data is sensitive and how it can be used. For example, sensitive data is any data that reveals: Racial or ethnic origin; political opinions; religious or philosophical beliefs; trade union membership; genetic data; biometric data for the purpose of uniquely identifying a natural person; data concerning health or a natural person's sex life and/or sexual orientation;
- Processing of such data is prohibited according to the GDPR unless explicit consent has been given by the data subject or for overriding reasons such as specified in the GDPR. Legal guidelines are contained in the GDPR (https://gdpr-info.eu/art-9-gdpr/). However, additional ethical guidelines could be provided for systems development or organizational use;
- For sensitive medical information, impose appropriate safeguards for its processing, distribution, merging with other data sources, and reidentification, and take appropriate measures to protect privacy;

[10] Peek, N., J. H. Holmes, and J. Sun, "Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics", *Yearbook of medical informatics,* Vol. 23, No. 1, 2014, pp. 42-47., p. 43.
[11] Costa, Fabricio F., "Big data in biomedicine", *Drug discovery today,* Vol. 19, No. 4, 2014, pp. 433-440., p. 437.
[12] Costa, Fabricio F., op. cit., p. 438; Bellazzi, Riccardo, "Big data and biomedical informatics: a challenging opportunity", *Yearbook of medical informatics,* Vol. 23, No. 1, 2014, pp. 8-13., p. 10.
[13] Rumbold, John M.M., and Barbara K. Pierscionek, "A critique of the regulation of data science in healthcare research in the European Union", *BMC medical ethics,* Vol. 18, No. 27, 2017, pp. 1-11.

- Patients should have a right to know who has their data, where it is, and when it is accessed. It should be clearly communicated, and accessible to patients, what research questions/tasks healthcare professionals and data companies want to have answered when acquiring patient data, and there should be transparency and explainability in the kinds of inferences that are drawn from their medical data;
- There should be a means of ensuring that at each stage of processing a trace can be identified between e.g. hospitals and data companies of when, and why specific data was used, to ensure greater accountability and intelligibility. This means of tracing should also allow for any findings to be made knowable to the patient as well as limiting who has access to the findings.

5.6 Covert and deceptive AI and big data systems

For reasons of autonomy, transparency, liberty, wellbeing, and fairness, serious limits should be imposed on AI systems that are covert or deceptive. ***Covert AI systems*** are AI systems that are not easily identifiable as such. They include systems that human beings interact with without knowing them to be AI systems, either because they come across as computer-mediated human beings, or as regular machines or software programs. They also include AI systems that quietly perform activities in the background that affect the interests of the individuals present (e.g., recording and analysing them, or influencing their behaviours).

Deceptive AI is AI that is programmed to provide false and misleading information, and to trick and deceive people. Since about 2010, deceptive AI systems have been under development. In the military, deceptive AI is considered compatible with military law. The use of deceptive AI outside of the military could be considered morally problematic. It affects autonomy, can lead to individual and societal harms, and undermines trust. Such AI systems pose the greatest threats to those in society that are susceptible to deception and manipulation. Such groups include, for example, the elderly, those with health problems (specifically mental health), those with a low level of comprehension of the language, children, or individuals with cognitive disabilities or social disorders.

Requirements:

- Human beings should always know if they are directly interacting with another human being or a machine. It is the responsibility of AI practitioners that this is reliably achieved, by ensuring that humans are made aware of – or able to request and validate the fact that – they are interacting with an AI system (for instance, by issuing clear and transparent disclaimers);
- For AI that is not interactive or cannot be mistaken for a human being, it is recommended that it is communicated to users that the information system or embedded system that is used makes use of AI, and how the AI algorithm operates;
- The use of deceptive AI beyond defence applications requires a strong justification and an extensive assessment in terms of its impacts on legal and human rights, and an overall cost-benefit analysis.

5.7 AI and big data systems that can recognize or express emotions

AI systems may interact with humans using spoken or written natural language, and may use an on-screen appearance of an animated person or avatar. Without an avatar, they may still take on an identity as if

they were a person (e.g., Alexa, Siri). These systems are called conversational agents. AI may also be embedded in robots that resemble humans in their appearance and movements. The recognition and expression of emotions may result in better interaction with human users, but also raises ethical issues. The recognition and processing of human emotions may infringe on human autonomy, freedom and privacy. The expression of emotions by machines may lead to unwanted attitudes and beliefs in humans, who may be deceived or manipulated and develop unwanted attachments.

Requirements:

- When machines recognize, process or express emotions, an ethical impact assessment should be done that covers impacts on legal and human rights, social relations, identity, and beliefs and attitudes. Stakeholders should be involved. There should be a clear benefit to the emotion abilities that should be weighed against the ethical considerations;
- When machines express emotions, there should be pre-emptive statements that one is interacting with a machine and there should be built-in distinguishability from humans.

5.8 AI and big data systems with applications in media and politics

The domains of media and politics require special ethical concerns because of the importance of free speech and of democratic institutions. The use of AI and big data systems in media includes applications in marketing, telecommunications, social media, publishing, information service companies and entertainment companies. These applications contain structured and unstructured text, audio, video and image data which are mined by analytics techniques to reveal patterns, opinions, and attitudes, and to generate data and content, for example in the form of trending topics, data visualisations, personalised ads, and value-added services such as location/content recommendations for public interest and consumption. Companies working in media sectors have an incredible amount of data that they can access, analyse and make decisions on, which affect and influence individual and group behaviour. These decisions are based on the data that these same individuals and groups produce, whether knowingly or unknowingly. Ethical issues in digital media include privacy and surveillance, autonomy and freedom (including free speech), fairness and bias, and effects on social cohesion (relating to the formation of filter bubbles and echo chambers).

When this level of tracking, monitoring and messaging is performed for political purposes, it contains risks of political manipulation of voters through psychologically exploitative microtargeting and distribution of fake news as part of misinformation campaigns.[14] Media companies are also in a position to determine what kind of political speech they allow and under what conditions, and to which third parties they give access to their platforms, giving them responsibility for political discourse and democratic processes.[15]

[14] Lepri, Bruno, Jacopo Staiano, David Sangokoya, Emmanuel Letouzé, and Nuria Oliver, "The tyranny of data? the bright and dark sides of data-driven decision-making for social good", in Tania Cerquitelli, Daniele Quercia, and Frank Pasquale (eds.), *Transparent data mining for big and small data*, Springer, Cham, 2017, pp. 3-24., p. 11.

[15] Helbing, Dirk, Bruno S. Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hagner, Yvonne Hofstetter, Jeroen van den Hoven, Roberto V. Zicari, Andrej Zwitter, "Will democracy survive big data and artificial intelligence?", *Towards Digital Enlightenment*, Springer, Cham, 2019, pp. 73-98., p. 7.

Requirements:

- In the development of digital media, ethical impact assessments should be done that covers impacts on legal and human rights, issues of fairness and bias, and effects on social cohesion and democracy. Stakeholders should be involved, and a careful balancing of relevant values should take place;
- Political and ideological speech should in principle not be abrogated, but should be subjected to assessments of falsehood and hate speech before publication. In case of violation of policies, speech should either not be published or it should be published with a warning;
- Readers/users should be approached based on principles of informed consent, and information offered to them should come with relevant disclaimers, opt-out mechanisms, and opportunities to see how they are profiled.

5.9 AI and big data systems in defence

The deployment of AI and big data systems in defence contexts occurs in a wide range of applications. These include: conventional military defence (e.g. development of military AI), counter-nuclear proliferation, counter-chemical/biological WMD, counter-terrorism, and cybersecurity as well as counter-intelligence. These applications have data sources that range from human actors, geospatial tools (e.g. mapping and satellite data), measurement and signature sensing tools (i.e. for identifying distinctive features of emitters), as well as online data.[16] Within combat, AI will likely be used in combat in two ways. First, AI will be used in a 'hybrid' way, assisting soldiers in targeting or communication in ways that nonetheless retain significant control by the human. In these cases, the human will retain meaningful control, though the AI will control, direct, or automate some elements of the humans' interaction with the battlespace. Second, AI might be used to direct genuinely 'autonomous' weapon systems that will have full control throughout the decision chain to use deadly force where human oversight is indirect and unreliable.

Ethical issues in defence pertain to the fundamental interests of persons: life, health, and property. They also concern the conditions under which different technologies and applications allow for confirmation of doctrines of 'a Just war'. In addition, they raise rights issues for soldiers who use these technologies. Autonomous and semi-autonomous weapons systems, and AI systems in defence generally, raise issues of responsibility and accountability: should AI systems be able to make autonomous decisions about life and death? Who is ultimately accountable for these decisions, and do systems allow for enough meaningful human control for humans to be accountable?

Requirements:

- For new, AI-enabled weapons systems, an ethical impact assessment should be done that includes careful consideration of the effects on 'Just war' policies, risks for new arms races and escalation, risks for soldiers and civilians, and ethical considerations concerning rights and fairness;

[16] Brewster, Ben, Benn Kemp, Sara Galehbakhtiari, and Babak Akhgar, "Cybercrime: attack motivations and implications for big data and national security", in Babak Akhgar, Gregory B. Saathoff, Hamid R. Arabnia, Richard Hill, Andrew Staniforth, and Petra Saskia Bayerl (eds.), *Application of big data for national security: a practitioner's guide to emerging technologies*, Butterworth-Heinemann, 2015, pp. 108-127.

- AI-enabled weapons systems should allow for meaningful human control in targeting and the use of force, and a clear delineation of responsibility and accountability for the use of force;
- New technologies for enhancing soldiers' readiness and ability, especially those that are invasive or work on the body, should be carefully considered for their consequences for the individual rights and wellbeing of soldiers;
- AI-enabled technologies for surveillance and cyberwarfare should be subjected to an ethical impact assessment that assesses their consequences for individual rights and civil liberties, safety and security risks, and impacts on democracy and politics, and the possibility of meaningful human control, weighed against their intended benefits.

5.10 Ethically aware AI and big data systems

Ethically aware AI and big data systems are studied and developed in the field of machine ethics, which aims to develop machines with the ability to ethically assess situations and act on these assessments. Ethically aware AI is AI that is programmed to avoid unethical behaviour, or, even to be able to apply ethical principles and adjust conduct as a result. The obvious benefit of ethically aware AI is that such AI systems may behave more morally. An added benefit may be that they are capable of giving moral reasons for their actions, thus enhancing explainability and transparency. There are however several issues that arise with ethically aware AI.

Firstly, ethically aware AI may be considered problematic due to the nature of ethics. Ethics is not an algorithmic exercise of applying systematically ranked moral principles to situations.[17] There are incoherencies and inconsistencies in ethical theories that humans can deal with, but computers (so far) cannot. Moral reasoning also requires moral intuitions and common sense, which AI does not have naturally, and there are issues of value pluralism and value conflict that computers cannot easily deal with. This makes it difficult to implement ethical theories into AI systems. We can build ethics into a system but that is different from ensuring that the system complies with ethical principles.

Secondly, there is the possibility of system failure and corruptibility. Machines may draw the wrong ethical inferences, with potentially disastrous effects. Third, ethically aware AI may limit human responsibility by suggesting that moral responsibility can be delegated to machines (Cave et al., 2019). Fourth, ethically aware systems could be conceived by some as moral patients, that can experience harm and have certain rights.

Requirements:

- In developing ethically aware systems, the limitations of artificial ethics should be carefully assessed, as well as risks of system failure and corruptibility, limitations to human responsibility, and risks of attributions of moral status;
- Users should be made aware that AI systems are ethically aware and what this implies;
- Ethics should be in line with the culture in which it is embedded;

[17] Brundage, Miles, "Limitations and risks of machine ethics", *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 26, No. 3, 2014, pp. 355–372.

- Compliance certification (external) and internal audit should be ensured.