SHERPA

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

D3.2 Guidelines for the development and use of SIS

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.

29/11/2019

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

Document Control

Deliverable	D3.2 Guidelines for the development and use of SIS
WP/Task Related	WP 3: Responsible Development of SIS
Delivery Date	28 November 2019
Dissemination Level	Public
Lead Partner	University of Twente
Main Authors	Philip Brey, UT; Björn Lundgren, UT; Kevin Macnish, UT; Mark Ryan, UT
Contributors	Andreas Andreou, AES; Laurence Brooks, DMU; Tilimbe Jiya, DMU; Renate Klar, EUREC; Dirk Lanzareth, EUREC; Jonne Maas, UT; Isaac Oluoch, UT; and Bernd Stahl, DMU.
Reviewers	Josephina Antoniou, UCLanCY; Bernd Stahl, DMU; Thamar Zijlstra, NEN
Abstract	This report provides two sets of guidelines for responsible development and use of SIS, along with supporting documentation
Key Words	guidelines, codes, conduct, CRISP-DM, AGILE, COBIT, ethics, smart information systems, big data

Revision History

Version	Date	Author(s)	Reviewer(s)	Notes
1	31/7/2019	M Ryan	P Brey, K Macnish	First draft
2	12/9/2019	M Ryan, B Lundgren	P Brey, K Macnish	Second draft
3	28/9/2019	B Lundgren	P Brey, K Macnish	Amended in response to stakeholder feedback
4	11/11/2019	B Lundgren, P Brey	K Macnish	Third draft including special topics of interest
5	23/11/2019	B Lundgren	T Zijlstra, J Antoniou, B Stahl	Amended in response to Quality Assurance process
6	27/11/2019	B Lundgren	P Brey, K Macnish	Amended in response to editorial comments
7	25/8/2020	K Macnish	B Stahl	Amended in response to reviewer comments

Executive Summary

This report contains two sets of ethical guidelines – one for the technological *development* and one for the *use* – of artificial intelligence and big data systems, a glossary, two annexes, and a list of references. 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 other existing guidelines in that they are directly related to design and development practices. They are intended to be actionable guidelines for systems and software development and use respectively, 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* or *use* practices would result in more ethical AI and big data products or uses of such products. The development 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). Following the guidelines would result in more ethical AI and big data research.

The guidelines have been drafted following an analysis of over 70 sets of potentially suitable ethical guidelines for AI and big data, which was reduced to 25 suitable guidelines. These 25 guidelines were then subject to a closer analysis, and our final product is closely aligned with the ethical guidelines of the EU High-Level Expert Group on Artificial Intelligence (HLEG AI). A complete analysis can be found in Annex 2.

Following the initial analysis, draft guidelines were submitted to the SHERPA Stakeholder Advisory Board and other interested stakeholders identified through the project for comment. Once initial comments were collected and incorporated into the documents, two workshops were held to engage stakeholders with the guidelines and envision them as to the value of an ethics by design approach. Feedback from the workshops was incorporated into the documents, which were then re-sent to stakeholders for a further round of comments. These were again incorporated before a final version was sent to stakeholders for approval. In this way, the three rounds of stakeholder comment and the workshop ensured that the guidelines were subject to significant stakeholder engagement. As such, they reflect real-world concerns, practices, and solutions which may not have been obvious to the original drafters of the guidelines. Stakeholder engagement is developed in Annex 1.

After this executive summary and the introduction of this report, we present the two sets of guidelines, starting with the development guidelines. Since these guidelines are intended for separate dispersion, both guidelines contain an executive summary and an introduction. After their introductory sections, we devote section 2 (“High-Level Requirements”) to present and discuss the high-level requirements that form the point of departure of this report. In the two sets of guidelines, we distinguish 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, specifying more concrete conditions that must be fulfilled. Operational guidelines are guidelines tied to specific practices and specific operational guidelines prescribe specific actions to be taken. In the two guidelines, we move from high-level to operational guidelines for the development of AI and big data.

Our high-level requirements are directly based on the guidelines of the 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 well-being; and accountability. For each, we also specify three or four sub-requirements that constitute a first step towards operationalisation. Because section 2 is devoted to high-level requirements they are similar for both sets of guidelines.

Next, the two guidelines diverge because we turn to present independent models for development and for use of AI and big data systems. In the development guidelines (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 we discuss 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 for 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 moreover has four to five subphases. Our objective is to develop operational requirements that are based on the high-level requirements that are tied to different development phases and subphases. In section 3, we provide some general guidelines for implementing ethical requirements in CRISP-DM. In section 4, we provide specific operational guidelines for the seven requirements that were presented in section 2.

In the use guidelines (Section 3, "Models for the ethical use of AI and big data systems in organisations"), we discuss models for the deployment and use of information systems in organisations, and we discuss how ethical principles for AI and big data could be made part of these models. Different deployment and use models include similar phases and practices (e.g., acquisition and design, deployment and implementation, normal use, evaluation). We use a combination of the commonly-used COBIT and ITIL models for the management and governance of information technology in organisations, and use the different practices and phases that they present to implement operational (or "low-level") ethical principles for AI and big data.

Our combined COBIT/ITIL model identifies six major phases in the deployment and process: IT governance, IT management strategy, Acquisition and design, Deployment and implementation, Service operation, and Monitoring, evaluation and improvement. For each phase, we propose operational requirements that are based on the high-level requirements and sub-requirements. In section 3, we provide some general guidelines for implementing ethical requirements in our model. In section 4, we provide specific operational guidelines for the seven requirements that were presented in section 2.

Both guidelines end with a presentation and discussion of ethical guidelines for special topics in AI and big data. By special topics, we mean AI or big data systems, applications, data types, or other 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 the model used in the development guidelines, special topics should be identified and taken into account at the Business Understanding phase in CRISP-DM. In the model used in the use guidelines, special topics should be included in the IT management strategy as part of the ethics requirements, and should be tested for in the Acquisition and design stage as well as in successive stages.

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 officer or managers in organizations that have a responsibility for ensuring the implementation of ethical practices within the organization. They are not, perhaps, directly usable by system developers. A further step that is still required, and not contained in this report, is the training of developers, of IT staff, and users, 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).

After the two sets of guidelines this report ends with a glossary and two annexes. The glossary is included to help with the technical terms used in the guidelines. It comprises of a number of currently available glossaries in the area, combining a mixture of research ethics, information technology, and our own classification of terms. Next, we have included two annexes. First, in developing these guidelines we used stakeholder participation, both in the form of a two-day workshop (arranged in July 2019) and in the form of a survey. Details about this is included in the first annex. Second, we include the aforementioned comparative analysis of the 25 guidelines that matched our selection criteria. Finally, at the end of the report is a reference list.

Table of Contents

Introduction

This report contains two sets of ethical guidelines – one for the technological *development* and one for the *use* – of artificial intelligence and big data systems, a glossary, two annexes, and a reference list. These guidelines are created by the SHERPA project, which has focused on the ethical, legal, and social issues arising from the development and use of these AI and big data systems. They are intended to be implemented in organizations by a manager or, preferably and 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 a system, not the process as such (ethical research boards are often concerned with the process). Applying these guidelines in development and use practices, or for research assessment, would result in more ethical AI and big data products, use, 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 in turn was reduced to 25 suitable guidelines that we built on, to construct the guidelines.[2] In particular, these guidelines build closely on EU's High-Level Expert Group on Artificial Intelligence. Our aim has been to build on their fundamental values but 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 whenever 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, whom we refer to as the 'end-user' throughout the text. Also, when we talk of an AI and big data system, we typically refer to it as 'the system'. Furthermore, we talk about stakeholders as individuals that have a stake in and/or can be affected by a system.

After this introduction, we present the two sets of guidelines, starting with the development guidelines. Both sets of guidelines start with an executive summary and an introduction. Next, the guidelines briefly describe the different types of requirements, starting with the top level values (section 2). These values are the same in both guidelines. Next, the guidelines diverge. In the development guidelines, 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. In the use guidelines, we describe how the ethical analyses can be mapped and related to IT management and governance frameworks (we illustrate this using the 'COBIT' and 'ITIL' models — section 3). After this analysis of how to integrate ethics

[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.

into development models and management and governance frameworks, we present our specific ethical requirements (section 4). Although these will build on the analysis from the previous section, they do not depend on it directly and can be read as standalone sets of guidelines for development and use of these systems. Lastly, in both sets of guidelines, we address special ethical issues related to these systems that are not captured in the general guidelines and make recommendations for these topics (section 5).

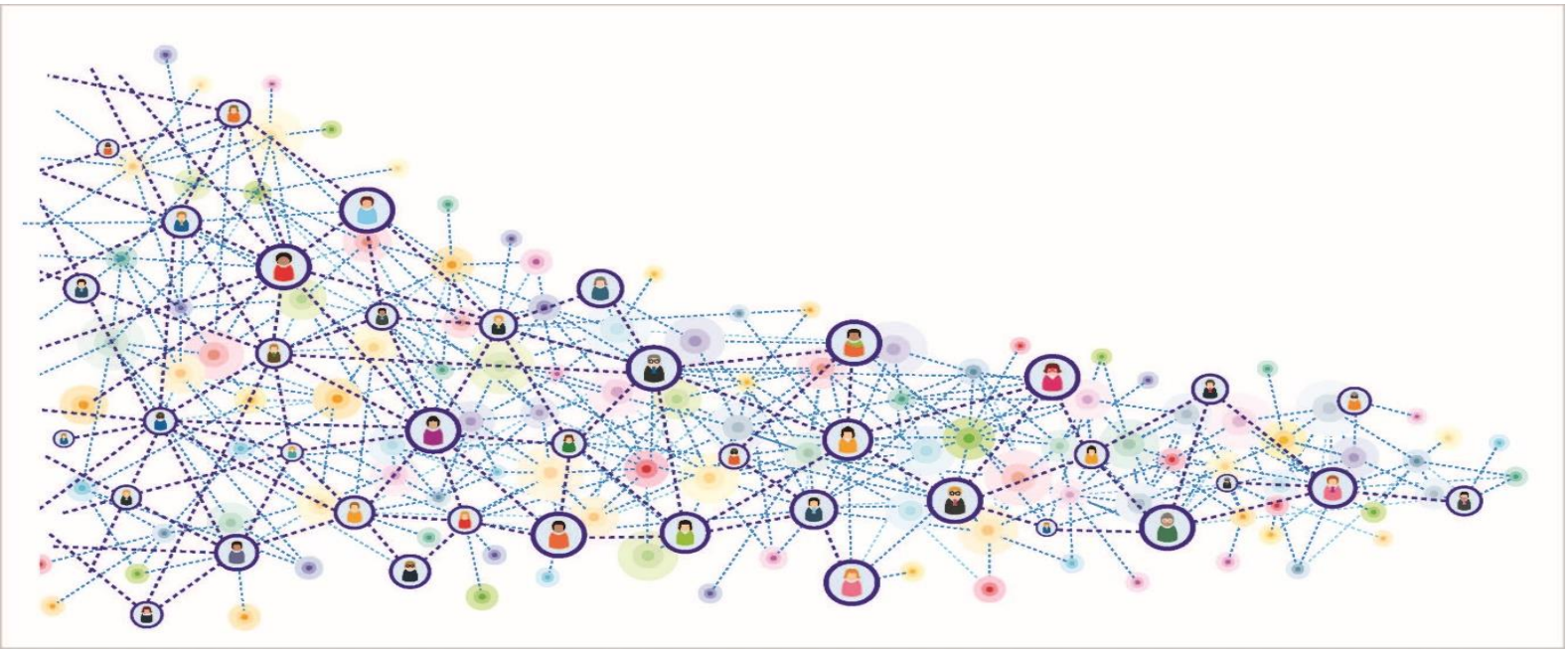
Finally, after the two sets of guidelines this report ends with a glossary, two annexes, and a reference list. The glossary contains explanations of technical terms used in the guidelines. It is comprised of a number of currently available glossaries in the area, combining a mixture of research ethics, information technology, and our own classification of terms. Next, we have included two annexes. first, in developing these guidelines we used stakeholder participation, both in the form of a two-day workshop (arranged in July) and in the form of a survey. Thus, in the first annex, we include the survey and an analysis thereof. Second, we include the aforementioned comparative analysis of the 25 guidelines that matched our selection criteria. Lastly, we include a list of references.

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

- part of the D3.2 ethics by design approach to creating ethical guidelines aimed at developers of AI and big data systems



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).

Contents

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.[3] 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*.[4] 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.

---

[3] 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).

[4] 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.[5]

[5] 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).

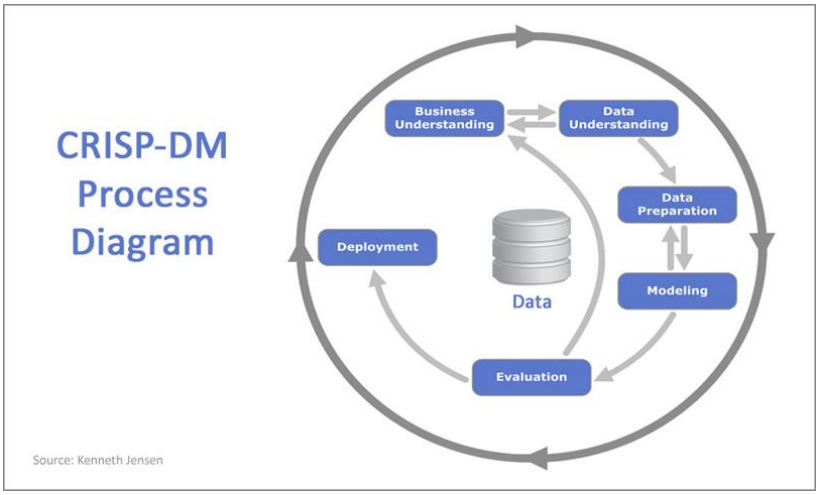
In this section, we will focus on the first of these. This means that we will mainly focus on technological methods for responsible AI.

In this section, we describe the well-known CRISP-DM model.[6] 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.



[6] 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

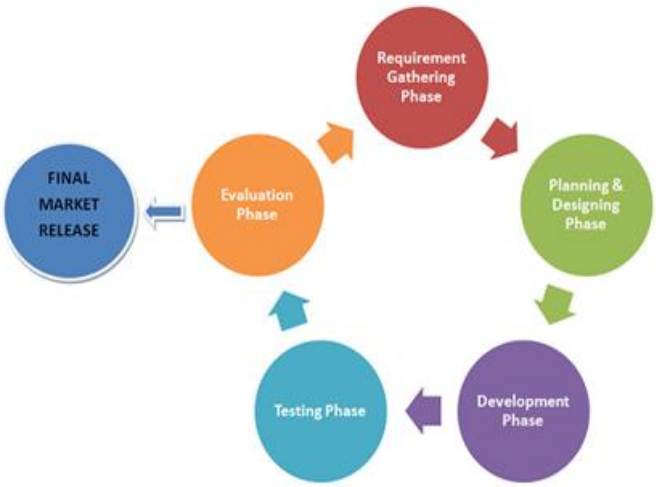
Deployment is the process of getting an IT system to be operational in its environment, including installation, configuration, running, testing, and making necessary changes. Deployment is usually not done by the developers of a system but by the (IT team of the) customer. Nevertheless, even if this 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.[7]

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 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

[7] 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/

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".[8] It is thus important that in this phase ethical values are evaluated as well as the desired requirements of the client.

[8] 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.

4.1.1 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.

4.1.2 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).

4.1.3 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.

4.2.1 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.

4.2.2 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).

4.2.3 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.

4.3.1 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.

4.3.2 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.

4.3.3 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).

4.3.4 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).

4.4.1 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.

4.4.2 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).

4.4.3 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:

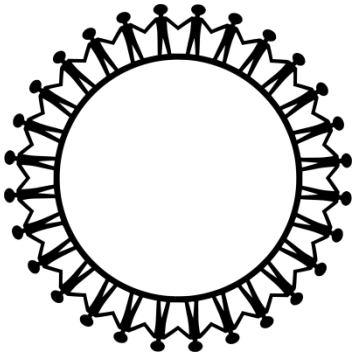
- a culture is established and encouraged based on mutual trust, transparent communication, open and understandable terms, a common language, ownership, and accountability;
- an explanation, which all reasonable users and stakeholders can presumably understand, is given as to why the system took a certain 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.

4.5.1 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.

4.5.2 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.

4.5.3 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.

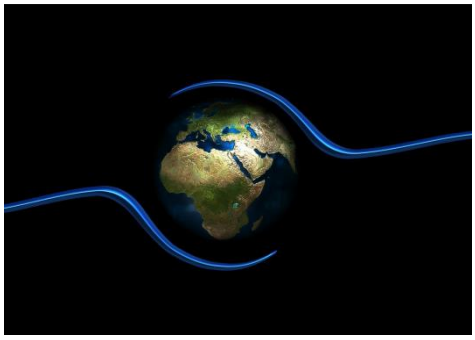
4.6 Individual, Societal, and Environmental Wellbeing

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.

4.6.1 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.

4.6.2 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).

4.6.3 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.

4.6.4 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.

4.7.1 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.

4.7.2 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.

4.7.3 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).

4.7.4 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.

4.7.5 Human Oversight

Requirement 73: 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,[9] lip-read what individuals are saying,[10] track people's activities across space,[11] 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 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.

[9] 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/

[10] 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/

[11] 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.

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.[12] 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.[13] 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.[14] 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,[15] 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;
- 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

[12] 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.

[13] Costa, Fabricio F., "Big data in biomedicine", *Drug discovery today,* Vol. 19, No. 4, 2014, pp. 433-440., p. 437.

[14] 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.

[15] 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.

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.[16] 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.[17]

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;

[16] 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.

[17] 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.

- 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. [18] 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;
- 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

[18] 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.

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.[19] 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;
- Compliance certification (external) and internal audit should be ensured.

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

SHERPA

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

Guidelines for the Ethical Use of AI and Big Data Systems

- part of the D3.2 ethics by design approach to creating ethical guidelines aimed at users of AI and big data systems

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 deployment and use of artificial intelligence (AI) and big data systems in organizations. 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 practices of deployment, implementation and use. They are intended to be actionable guidelines for organisations that use these systems, rather than abstract principles that have no direct application in practice. We call such guidelines *operational*, meaning ready for use. Applying these guidelines in practice would result in more ethical use of AI and big data technologies.

In constructing *Guidelines for the Ethical Use of AI and Big Data Systems*, 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, which 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 ("Models for the ethical use of AI and big data systems in organisations"), we discuss models for the deployment and use of information systems in organisations, and how ethical principles for AI and big data could be made part of these models. Different deployment and use models include similar phases and practices (e.g., acquisition and design, deployment and implementation, normal use, evaluation). We use a combination of the COBIT and ITIL models for the management and governance of information technology in organisations, and use the different practices and phases they present to implement operational (or "low-level") ethical principles for AI and big data.

Our combined COBIT/ITIL model identifies six major phases in the deployment and process: IT governance, IT management strategy, Acquisition and design, Deployment and implementation, Service operation, and Monitoring, evaluation and improvement. For each phase, we propose operational requirements that are based on the high-level requirements and sub-requirements. In Section 3, we provide some general guidelines for implementing ethical requirements in our model. 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 included in the IT management strategy as part of the ethics requirements, and should be tested for in the Acquisition and design stage, and successive stages.

The guidelines 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. The guidelines are perhaps not directly usable by system operators. A further step that is required, but not contained in this report, is the training of IT staff and users 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 deployment and use process. We intend to produce further implementation documents in the EU Horizon 2020 SIENNA project (www.sienna-project.eu).

Contents

1. Introduction

These guidelines on the ethical *use* of artificial intelligence (AI) and big data systems, are part of a set of two (with separate guidelines for their ethical *development*). 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.[20] Applying these guidelines in practice would result in more ethical use of AI and big data technologies.

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 Use of AI and Big Data Systems*.[21] 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 and related to IT management and governance frameworks, and illustrate this by using the so-called 'COBIT' and 'ITIL' models in Section 3. After this analysis of how to integrate ethics into governance 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 and can be read as a standalone set of guidelines for how to use these systems. In Section 5 we address some of the most pressing special issues related to these systems, and how our guidelines may provide recommendations for these topics.

[20] 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).

[21] 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.[22]

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 requirements, 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:

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

Auditability, minimisation and reporting of negative impact, internal and external governance frameworks, redress, and human oversight

Table 1 [Use]: 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 we should not undermine 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 deployed and used with an inclusionary, fair, and non-discriminatory agenda. Requiring the developers to include 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 in 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 well-being.

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 impacts, with internal and external governance frameworks (e.g., whistleblowing), and human oversight.

3. Models for the ethical use of AI and big data systems in organisations

In this section, we discuss how ethics can be integrated into governance and management in organisations in such a way that the deployment and use of AI and big data systems take ethical criteria into account. We illustrate this by discussing two popular models for IT management and governance. However, the ethical guidelines do not depend on these particular models, which only serve as examples. The responsible and ethical deployment and use of AI and big data systems in organizations is the outcome of three factors:

1. Responsible IT management;
2. Responsible IT governance;
3. Support from other stakeholders and society at large (e.g., IT suppliers, governmental institutions, educational institutions, professional organizations, clients).

We focus on the first two of these factors. First, we discuss responsible IT governance, using the COBIT 19 model. COBIT is a good-practice framework for IT governance and management created by ISACA, an international professional association focused on IT governance. It is the most widely used framework of its kind. Second, we discuss responsible IT management, using both the COBIT 19 and the ITIL model. ITIL is the most widely used reference framework for IT service management. It is owned by AXELOS, a joint venture between Capita and the UK Cabinet Office.

There is agreement in the industry that IT management and IT governance should be distinguished from each other. ***IT governance*** is focused on strategic decision-making concerning the role of IT in the organization, whereas ***IT management*** concerns the operational excellence of IT services in the organization:

IT governance is typically the responsibility of the board of directors of a company, under the leadership of the chairperson – although in large organizations, specific governance responsibilities may be delegated to other units. It ensures that balanced and agreed enterprise objectives are defined, based on an assessment of stakeholder needs and options; that direction is set by prioritization and decision-making; and that performance and compliance are monitored against agreed-on objectives.

IT management is focused on planning, building, running and monitoring IT systems, services and activities, in alignment with IT governance, to achieve enterprise objectives. It is usually the responsibility of the executive management, under the leadership of the CEO. Often, the executive management will institute a board of business managers and IT managers to oversee the IT department, with responsibility for the overall IT management strategy and its alignment with corporate governance.

3.1 IT Governance and Ethics of AI and big data systems

The COBIT model[23] defines five objectives for the governance of IT by the directorate of an organization, that:
1) jointly ensures that there is an overall governance framework for IT in place that aligns IT management strategy with overall corporate strategy and objectives;
2) ensures effective oversight of IT-related processes that ensures adequate and sufficient business and IT-related resources;
3) accounts for strategic risks;
4) ensures engagement of stakeholders, and
5) ensures that IT services are delivered efficiently and effectively.

COBIT 2019 establishes a role for ethics in IT governance. It proposes, as part of the establishment of the overall governance framework for IT, that directors "[a]lign the ethical use and processing of information and its impact on society, the natural environment, and internal and external stakeholder interests with the enterprise's direction, goals and objectives", that they "[d]irect that staff follow relevant guidelines for ethical and professional behavior and ensure that consequences of noncompliance are known and enforced", and that they "[i]dentify and communicate the decision-making culture, organizational ethics and individual behaviors that embody enterprise values" and "[d]emonstrate ethical leadership and set the tone at the top."[24] Based on this we derive the following requirement:

Requirement 1: The board of directors should direct in its IT governance framework that IT management adopts and implements relevant ethical guidelines for the IT field, and should monitor conformity with this directive. There should be an appointed representative at each level of the organisation, including the board of directors, who are 'ethics leaders' or 'ethics champions', and who should meet regularly to discuss ethical issues and best practice within the organisation. The ethics leader from the board of directors should be responsible for the ethical practice of the whole organisation.

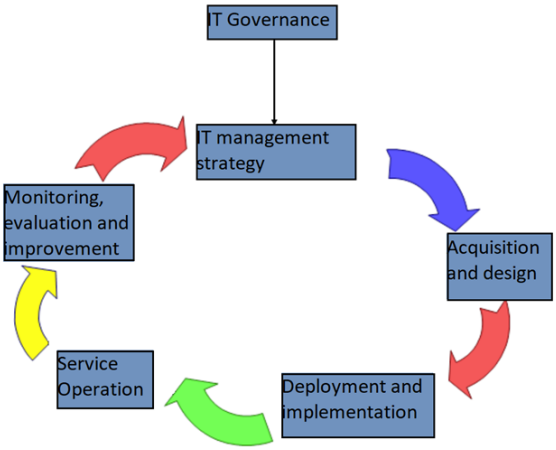
At the strategic level, there is no need to address which specific guidelines should be adopted for which IT-related purpose, although an engaged board can opt to issue more specific directives. To the extent that an organization has adopted broader ethics guidelines, corporate values, or corporate social responsibility strategies, as part of its overall corporate strategy, the board may also direct that these values and principles are adopted and implemented at the IT management level.

[23] ISACA (n.d.a), "COBIT 2019 Framework: Introduction and Methodology", *COBIT 2019 Framework: Introduction and Methodology*, n.d. http://www.isaca.org/COBIT/Pages/COBIT-2019-Framework-Introduction-and-Methodology.aspx; ISACA (n.d.b), "COBIT 2019 Framework: Governance and Management Objectives", *COBIT 2019 Framework: Governance and Management Objectives*, n.d. http://www.isaca.org/COBIT/Pages/COBIT-2019-Framework-Governance-and-Management-Objectives.aspx

[24] ISACA (n.d.b), op. cit., pp. 30-33.

3.2 IT Management and Ethics of AI and Big Data Systems

The two most frequently used reference frameworks for IT management are ITIL, which specifically focuses on IT service management,[25] and COBIT, which covers IT governance and management. Their perceptions of the overall IT management lifecycle and their segmentation of the different components of IT management is similar. Each identifies the activity of developing an overall IT management strategy, in relation to the IT governance strategy, as a necessary first step. Each then identifies the acquisition or design of IT systems and services as a next step in the development of IT services, followed by deployment and implementation. (In COBIT, design/acquisition and implementation are grouped together as one process.) Each then sees the regular operation of established IT systems and services as a next step in the cycle, and each defines a continuous activity of monitoring, assessment and improvement of IT services.



Process	In COBIT	In ITL
IT management strategy	Align, plan, and organize	Service strategy
Acquisition and design	Build, acquire, and implement	Service design
Deployment and implementation	Build, acquire, and implement	Service transition
Service operation	Deliver, service, and support	Service operation
Monitoring, evaluation, and improvement	Monitor, evaluate, and assess	Continual service improvement

Table 2 [Use]: IT life cycle in the COBIT and ITIL models.

We will now consider how to apply ethical considerations to the use of AI and big data systems for each of these five processes. We will do so with special reference to the COBIT model, since it already defines various points at which it recommends the inclusion of ethics considerations (see Table 3, next page).

[25] AXELOS, "ITIL® Foundation, ITIL 4 Edition", *ITIL® Foundation, ITIL 4 Edition*, TSO (The Stationery Office), n.d. https://www.tsoshop.co.uk/Business-and-Management/AXELOS-Global-Best-Practice/ITIL-4/?CLICKID=002289

	IT Management Strategy	Acquisition and Design	Deployment and Implementation	Service Operation	Monitoring, Assessment and Improvement
Human Agency	✓	✓	✓		✓
Liberty	✓	✓	✓		✓
Dignity	✓	✓	✓		✓
Resilience to Attack	✓	✓	✓	✓	✓
Fallback Plan	✓	✓	✓	✓	✓
Accuracy		✓	✓	✓	✓
Reliability		✓	✓	✓	✓
Privacy and Data Protection	✓	✓	✓		✓
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 3 [Use]: COBIT Model and the Ethical Requirements

3.2.1 IT Management Strategy

The establishment of an overall IT management strategy is recommended before the establishment of IT services. This strategy will also cover the organization of the IT department(s) and any supporting activities. In the COBIT model, it includes design of the overall IT management system; determination and communication of management objectives, decisions, policies and procedures; design and implementation of the organizational structure and management processes (including roles and responsibilities of units and staff members and their accountability), and definition of target skills and competencies. It also covers a strategic plan and road map; a common architecture for the IT function; the institution, management and monitoring of a portfolio of programs and services; management of budget and costs; human resources management; management of services and service levels; management of data; management of quality requirements; management of risks and information security, and management of relations with stakeholders and vendors.

Within the management strategy activities, COBIT includes the communication of codes of ethics to relevant audiences, and the inclusion of specific requirements in role and responsibility descriptions regarding adherence to codes of ethics, as well as the creation of policies to drive IT control expectations

on ethics. From this follows our second requirement, which is specific about the ethical guidelines that are at issue:

Requirement 2: The IT management strategy should include the adoption and communication to relevant audiences of ethics guidelines for AI and big data systems, define corresponding ethics requirements within role and responsibility descriptions of relevant staff, and include policies for the implementation of the ethics guidelines and monitoring activities for compliance and performance.

The strategy could, for example, include the institution of an ethics officer or an ethics committee, or the assignment of specific ethics responsibilities to different staff, such as the compliance manager, supplier manager, information security manager, applications analyst and/or IT operations manager. A company may only have one officer, so there is a need to embed ethical practice and understanding within the organisation. Individuals should be able to raise concerns with the 'ethics leader' within their department, or have the option to discuss them with an ethics leader at a different level in the organisation, the ethics officer, or an externally-appointed affiliate. There should be the possibility to escalate concerns at all levels within the organisation.

The IT management strategy includes the development of training programs to meet organizational and process requirements. This could include training programs for ethics awareness and ethical conduct for staff, including end-users. This brings us to:

Requirement 3: The IT management strategy should include the design and implementation of training programs for ethical awareness, ethical conduct, and competent execution of ethical policies and procedures, and these programs should cover the ethical deployment and use of the system. More generally, IT management should encourage a common culture of responsibility, integrating both bottom-up and top-down approaches to ethical adherence.

Various tasks within the IT management strategy will themselves be affected by ethics requirements generally, and AI and big data systems ethics guidelines specifically. The importance of and need for ethical guidelines must be discussed and highlighted for all members of the team, particularly the ethics leader at each level of the organisation. Requirement 4 reflects this:

Requirement 4: Consider how the implementation of the AI and big data systems ethics guidelines, and other IT-related ethics guidelines, affects the various dimensions of IT management strategy, including overall objectives, quality management, portfolio management, risk management, data management, enterprise architecture management, stakeholder relationship management. Ensure proper adjustment of these processes. There will be different levels of risk involved, depending upon the application, so the levels of risk need to be clearly articulated to allow different responses from the organisation's ethical protocols.

Also, make an evaluation of whether any of the special issues (from Section 5) are likely to be involved. If so, the guidelines for the special issues should be involved.

3.2.2 Acquisition and Design

This is the process of acquisition and/or design of IT solutions. The decision to acquire and implement a new IT solution will normally be made by IT management, within its IT management strategy. It will be the expression of a business objective that should be met with an IT solution, which may or may not be accompanied by further specifications. In the acquisition and design phase, the IT department will first investigate possible solutions and specify and analyze requirements, in consultation with stakeholders. It will then decide either to do in-house development, involve an external developer or vendor, or engage in a combination of these options.

Our concern is with IT solutions that involve AI and big data systems. Our requirements are as follows:

Requirement 5: The business objective should be tested against the ethics guidelines for the system, and system ethics criteria should be included in the requirements for the IT solution.

Requirement 5a: If in-house development is chosen, then the design team should follow development methods that include the system ethics requirements, such as specified in our document "Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach".[26]

Requirement 5b: If the IT solution is custom-built, then give preference to a developer who uses development methods that include the system ethics requirements. If this is not possible, include the ethics requirements in the specification given to the developer. The organisation should ensure that adequate due diligence is followed by the company to which they are outsourcing the systems' development, and ethical practices should be a procurement requirement.
Requirement 5c: If an off-the-shelf solution is acquired, then compare different solutions provided by different vendors for compliance with the system's ethical requirements, if possible also through testing. Whenever possible, similar steps should be taken to Requirement 5b. The organisation should identify potential bias and risks associated with vendors to identify those most in-line with their ethical protocols. They should also build a relationship of trust with the vendor to ensure confidence in their technologies and their ethical practices.

Requirement 5d: Stakeholder analysis or (better) stakeholder consultation should take place in 5a – c in order to identify direct and indirect stakeholders to the IT solution and to identify and take into account their values and interests. There needs to be clear identification of the relevant stakeholders, both internal and external to the organisation, as there will be different requirements for varying stakeholders.

Requirement 5e: An ethical impact assessment of the IT solution and its intended role in the organization should be considered before a final decision on deployment is made.

3.2.3 Deployment and Implementation

This is the process of deploying the IT solution into the user environment, and planning and implementing required changes in the business context to ensure its successful implementation. In COBIT 19, this corresponds with Build, Acquire and Implement processes BAI05-BAI11. In ITIL, this corresponds with the service transition phase. This includes the development of an implementation plan; the preparation and commitment of stakeholders for business change; the planning for business process, system and data

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

conversion; the development and implementation of an operation and use plan; the configuration of the IT solution and its embedding into IT infrastructure; the testing of the IT solution in its new environment; the implementation of any necessary organizational changes and any necessary new policies; the training of relevant stakeholders; the testing of acceptance by stakeholders, and post-implementation review:

Requirement 6: Account for the ethical guidelines for the system in the implementation plan, and monitor and assess proper implementation of the ethics guidelines during the deployment and implementation process. There should be a requirement to communicate the ethical principles to vendors and throughout the different levels of the company.

Requirement 6a: Establish and implement operation and use plans and policies that support compliance with the ethics guidelines for the system.

Requirement 6b: Update data, access, security, and risk management policies and procedures that apply to the new system to account for ethics requirements.

Requirement 6c: Conduct a stakeholder analysis or (better) consult stakeholders in carrying out 6a-b. There needs to be clear identification of the relevant stakeholders, both internal and external to the organisation, as there will be different requirements for varying stakeholders.

Requirement 6d: In training for operation and use of the system, including new ethics policies and procedures, and pay attention to ethical aspects in communication around the introduction of the new system.
Requirement 6e: Monitor the implementation of ethics guidelines for the system throughout the implementation phase, identify issues and risks, and make adjustments where needed.

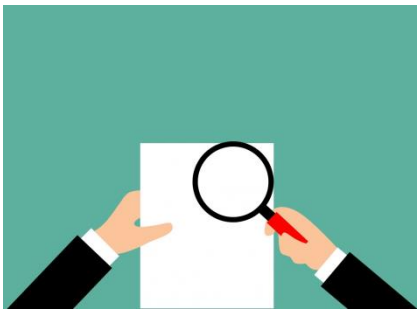
3.2.4 Service Operation

This is the regular operation of an IT solution after implementation. It includes the delivery of effective and efficient IT services, safeguarding access for authorized users, fulfilling user requests, solving service outages and other incidents that affect the quality of service, and reducing the impact of incidents. It involves regular maintenance, operational information security and access management, and the maintenance of information integrity and the security of information assets:

Requirement 7: IT operations personnel operate the system according to established procedures that include ethical aspects, verify and ensure that end-users use the system according to user policies that include ethical requirements, are vigilant about ethical issues in operation and use, and consult with senior staff on issues that are morally problematic or ambiguous. The organisation should try to ensure transparency within the organisation and identify ways to escalate issues if there are concerns from staff members. The board of directors and management should ensure that there are ways to raise conflicts and issues and a feeling of empowerment to do so.

We assume that at this point, operations and use policies are already in place and include ethical aspects, and that personnel have been trained to handle ethics requirements in their jobs.

3.2.5 Monitoring, Assessment and Improvement



This is a continuous process within the organization that includes the monitoring of performance, conformance, and compliance with external requirements, auditing and assurance processes, and the development and implementation of improvement plans:

Requirement 8: Include compliance with ethics guidelines for the system in monitoring goals and metrics, and propose improvements if monitoring shows compliance to be below target. There should be an ethics program within the company, with ethics metrics or goals, to see how many systems have been through an ethics impact assessment. Different levels of 'ethics leaders' need to come together to establish these metrics collectively and determine how to ensure and develop the organisation's ethical agenda.

Requirement 9: Include conformity with ethics guidelines for the system in assurance initiatives, and select qualified assurance providers that are familiar with the ethics guidelines or otherwise capable of including them in their assurance activities.[27] There should be built-in audit controls so that the system can tell you the decisions it is making and show how it is identifying potential discrimination and functioning.

[27] IEEE and ISO standards for (aspects) of ethics of AI and big data are currently forthcoming, and IEEE is also working on certification. The organization could consider adopting these standards and the certification scheme.

4. Specific Operational Ethics Requirements

Following our general requirements for applying ethical criteria in the management and governance processes, we now turn to more specific ethics requirements. This continues from the above requirements on how to include ethics into the COBIT and ITL models, and for each requirement we will connect it to the five phases above. However, the requirements do not depend on either model and can be used irrespective of your governance or management method. But if you want to make full use of the operational nature of the requirements, you need to consider how to map the five phases to your governance or management method.

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.

4.1.1 Human Agency

Requirement 10: Potential for impact on autonomy.

- In all phases (except the service phase), 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, ensure their awareness (e.g., if an end-user is using the system in a medical capacity, then 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 stakeholder's decision-making process (e.g., by recommending actions, decisions, or by how it presents stakeholder's with options) is justified and minimised.

4.1.2 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).

4.1.3 Human Dignity

Requirement 12: Respect for Human Dignity.
In all phases (except the service phase), 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 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.

4.2.1 Resilience to Attack and Security

Requirement 13: Security, design, testing, and verification.
In all phases, assess and ensure that:

- you have evaluated the possible security risks and that the system is protected against cybersecurity attacks during use;
- the security of the system is tested and, whenever possible, verified before, during, and after deployment;
- security measures benefit humans.

Requirement 14: Resilience.
In all phases, 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.

4.2.2 Fallback Plan and General Safety

Requirement 15: Safety and verification.
In all phases, assess and ensure that:

- your organisation has the necessary skills to understand how the system functions and its potential impact;
- evaluate possible risks of the system and ensure that mechanisms to safeguard user safety and protect against substantial risks are implemented before deployment;
- the system is tested before, during, and after deployment, to remain safe and secure throughout its lifetime;
- safety measures benefit humans.

Requirement 16: Fallback.
In all 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).

4.2.3 Accuracy, Reliability, and Reproducibility

Requirement 17: Accuracy, reliability, and effectiveness.
In every phase (except the management phase), assess and ensure:

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

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

- the security and safety objectives, results and outcomes are actively monitored and documented during use and, whenever possible, that the developer supplies such documentation for the design process;
- that 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 at 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.

4.3.1 Respect for Privacy

Requirement 19: Clarify roles and responsibilities towards information use, security and privacy.
In all phases (but especially the management phase), 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 behaviour.

Requirement 20: Develop cultures of security and privacy awareness.
In all phases, 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;
- whenever possible, 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.

Requirement 21: Personal data use, reduction, and elimination.
In all phases (except the service phase), assess and ensure that:

- alternatives that minimize or eliminate the use of personal 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 aggregation technology);

Requirement 22: Personal data storage.
In the acquisition and design, deployment and implementation, and monitoring phases, assess 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 acquisition and design, deployment and implementation, and monitoring 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 deployment and implementation and monitoring phases, assess and ensure that:

- Assess the creation of new personal and/or sensitive data, for example, through estimation of missing data, the production of derived attributes and new records, data integration, or aggregation of data sets. Assess how potentially privacy-sensitive this new information is and ensure a further informed consent if needed, or seek an alternative legal basis as set out in Articles 6(1) and 9(2) of GDPR. Ensure that all newly created personal or sensitive information/data is given at least the same protection as previously collected or held personal or sensitive information/data.

Requirement 25: Subsequent collection and/or creation of new personal data.
In the acquisition and design, deployment and implementation, and monitoring 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 of 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 a further informed consent is acquired, if needed.

Requirement 26: Privacy awareness.
In the acquisition and design or deployment and implementation phases, assess and ensure:

- processes that allow users to flag issues related to privacy or data protection in the system's processes of data collection and processing;
- processes 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 acquisition and design or deployment and implementation phases, assess and ensure:

- whenever possible, ways to use the system without or with minimal use of potentially sensitive or personal data (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 the acquisition and design or deployment and implementation phases, assess and ensure that:

- the system is not deployed unless it 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 Privacy Officers.
In all phases, ensure that:

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

4.3.2 Quality and Integrity of Data

Requirement 30: Oversight of data quality.
In the acquisition and design, deployment and implementation, and monitoring phases, assess and ensure that:

- there are processes to ensure the quality and integrity of all pertinent data, including, if possible, means of verifying that data sets have not been compromised or hacked;
- 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 management, deployment and implementation and service phases, 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 type of organization.

4.3.3 Access to Data

Requirement 32: Oversight of access to data.
In the deployment and implementation and monitoring phases, 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 oversight process 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 all phases (except the service phase), 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 a 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 should be 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 acquisition and design, deployment and implementation, and monitoring phases, 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).

4.3.4 Data Rights and Ownership

Requirement 35: Clarity on ownership of data.
In the acquisition and design and deployment and implementation phases, 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, users, and end-user 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).

4.4.1 Traceability

Requirement 36: Traceability measures.
In the acquisition and design and deployment and implementation phases, assess and ensure that:

- before purchasing or deploying a system, that the development companies should attempt to ensure that they design them to ensure traceability through the following methods:
    - Methods used for designing and developing the system (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 the system (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 the acquisition and design, deployment and implementation, and monitoring phases, assess and ensure that:

- before purchasing or deploying the system, there is a "human in control" when needed, and the moments or tools for human intervention 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 end-users 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;

- before purchasing or deploying the system, ensure detection and response mechanisms if something going wrong, and closely liaise with end-users about appropriate remedial steps thereafter.

4.4.2 Explainability

Requirement 38: Training data.
In the acquisition and design phase, assess and ensure that:

- whenever possible, communicate with the developers or suppliers of the system to inquire about what the system is being trained on, what the training data is, and ensure that it complies with relevant ethical standards.

Requirement 39: Explainable systems.
In the acquisition and design phase, assess and ensure that:

- before purchasing or deploying the system, evaluate the extent to which the decisions and outcomes made by the system can be understood, including whether you have access to the internal workflow of the model;
- prioritize, whenever possible, systems that increase decisional transparency (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 acquisition and design, deployment and implementation, and service phases, assess and ensure that:

- before purchasing or deploying the system, that the process of, and rationale behind, the choices made by the system are explainable upon request to an end-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 auditing bodies;
- there is redress and explanations of how the system arrived at those decisions, if there is harm caused to them 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 the acquisition and design phase, assess and ensure that:

- before purchasing or deploying the system, trade-offs between the explainability/transparency and best performance of the system are appropriately balanced based on the context of use (e.g., in healthcare the accuracy and performance of the system may be more important that 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).

4.4.3 Communication

Requirement 42: Communication regarding interactions with systems.

In the deployment and implementation phase, assess and ensure that:

- it is communicated to, and presumably understood by the end-users 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 end-user.

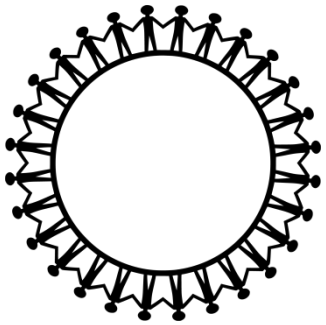
Requirement 43: Communication with stakeholders.
In all phases (except the service phase), assess and ensure that:

- a culture is established and encouraged in which open and structured communication is provided to stakeholders, in line with their requirements.
- 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 the purpose of the system 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 end-user and stakeholder community.
In the management, acquisition and design, and deployment and implementation 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;
- before purchasing or deploying the system, you will be able to provide an explanation which all reasonable end-users and stakeholders can presumably understand, as to why the system took a certain choice resulting in a certain outcome;
- there is a process to inform end-users about the reasons and criteria behind the system's outcomes, and establish processes that consider users' feedback and, in collaboration with developers, use this to adapt the system;
- any potential or perceived risks are clearly communicated to the end-user. 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 when organisations use AI and big data systems. It also aims to take a proactive approach and proposes that organisations 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. Although this mostly pertains to development issues, it will be highly relevant in the Acquisition and Design phase.

4.5.1 Avoidance and Reduction of Harmful Bias

Requirement 45a: System bias assessment.
In the management and acquisition and design phases, assess and ensure that:

- before purchasing or deploying the system, an evaluation of the diversity and representativeness of users in the data is performed, testing for specific populations or problematic use cases, and that input, training, output data, and the model, is analysed for harmful bias (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; some data sets may contain harmful biases if they consist solely of the behaviour of subclass of all people, e.g., young white men, and the system will be deployed in situations where groups other than those in the data set will be affected; 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; or 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) and, if possible, avoided (e.g., the organisation may incorporate additional users' data that is not included in the data or request the developers of the system to do so; consider alternative system developers that are not using unfair data; or the datasets being used may need to be discarded altogether);
- before purchasing or deploying the system, implement unconscious bias training to assist developers to identify innate biases during the development of system, or demand transparency from your AI supply chain that allows you to evaluate the system biases;
- before purchasing or deploying the system, data from just one class is not used to represent another class, unless it is justifiably representative.

- before purchasing or deploying the system, you have clearly established what kind of sample the system needs, what kind of sample you have taken, and that you articulate what it will be used for;

Requirement 45b: Use bias assessment.
In management and deployment and implementation phases, assess and ensure that:

- a strategy or a set of procedures is established to avoid creating or reinforcing unfair bias during the use of the system regarding the use of input data, and that the strategy is based on an assessment of the possible limitations stemming from the composition of the used data sets;
- use of the system is guided by an awareness of cultural bias to prevent or exacerbate any potential harmful bias.

Requirement 46: Engagement with users to identify harmful bias.
In the deployment and implementation phase, assess and ensure that:

- a process allows others to flag issues related to harmful bias, discrimination, or poor performance of the system, and establish clear steps and ways of communicating how and to whom such issues can be raised, during the deployment of systems;
- transparency to end-users and stakeholders about how the algorithms may affect individuals to allow for effective stakeholder feedback and engagement;
- when possible, implementation of methods for redress and feedback from end-users at all stages of the system's life-cycle (e.g., in collaboration with the developing company).

Requirement 47: Anticipating harmful functional bias.
In the acquisition and design, deployment and implementation, and monitoring phases, assess and ensure that:

- whenever possible, the potential of the system being used for different ends than those for which it was intended is avoided, and that if the system can be used for other ends, then consider potential implications of this likelihood, and develop mitigation procedures in the event of potential ethical issues arising;
- the system is not used for bad purposes and attempt to eliminate, whenever possible, misuse of the system (one way to do this is to request that the developer of the system use tried-and-tested models from trustworthy organisations).

Requirement 48: Decision variability.
In the deployment and implementation phase, assess and ensure that:

- a measurement or assessment process 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 end-user (e.g., in medicine this should be explained to doctors that use it).

Requirement 49: Avoiding harmful automation bias.
In all phases, 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 to prevent overconfidence in or overreliance on the system through education and training to be more aware of harmful bias in the system.

4.5.2 Ensuring Fairness and Avoidance of Discrimination

Requirement 50: Accessibility and Usability.
In the acquisition and design and deployment and implementation 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, e.g., in collaboration with the developers, and ensure that any limitations are fully understood by these groups);
- in the deployment and implementation phase, that you seek to involve or consult with people from teams or groups that represent different backgrounds and experiences (including but not limited to users of assistive technologies, users with special needs, 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 any negative implications are evaluated and that, whenever possible, adjustments are made to ensure that negative implications do not disproportionally affect some specific groups or individuals.

Requirement 51: Intended use.
During the acquisition and design and deployment and implementation 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.
During the acquisition and design and deployment and implementation phases, assess and ensure that:

- a knowledgeable professional, internal and external to the company, examines the product and its use through a risk assessment procedure.

Requirement 53: Distributing the system to end-users.
During the deployment and implementation phase, assess and ensure that:

- the end-user receives information about potential errors and the accuracy of the system (including the underlying certainty).

Requirement 54: Whistleblowing.
During all phases, 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.

4.5.3 Inclusionary Stakeholder Engagement

Requirement 55: Diversity.
In the acquisition and design, deployment and implementation, and monitoring phases, assess and ensure:

- a process to include the participation of different stakeholders in the 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;
- if this is applied within your organization, then inform and involve impacted workers and their representatives in advance.

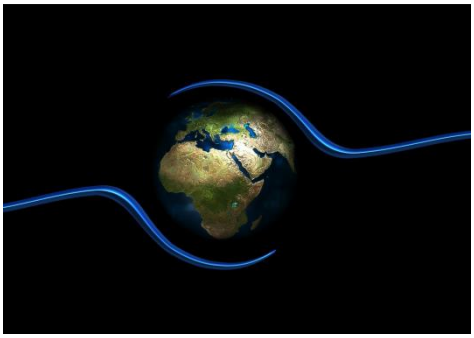
Requirement 56: Inclusion.
During the deployment and implementation and monitoring phases, assess and ensure:

- an adequate inclusion of diverse viewpoints during the use of the system;
- that deployment 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 environmental 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.

4.6.1 Sustainable and Environmentally-Friendly Systems

Requirement 57: Environmental impact.
In all phases (but especially during and after deployment), assess and ensure:

- a process 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 needs to be carefully factored in, with strong emphasis on the impact on these ecological externalities, through a holistic ecosystem-focused outlook);
- that your organisation is transparent about ecological impact and, if possible, work with environmental protection organisations to ensure the use of your systems are sustainable, and keep their ecological footprint proportionate to the intended benefit to humanity.

4.6.2 Individual Wellbeing

Requirement 58: Individual wellbeing assessment.
During the acquisition and design and deployment and implementation phases, assess and ensure that:

- you contribute, whenever possible, to increasing the knowledge of how the system may affect individual wellbeing;
- 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 vulnerable groups through discussion with them, rather than assuming their needs.

Requirement 59: Emotional attachment.
In the acquisition and design, deployment and implementation, and monitoring phases, 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;
- it is clearly communicated that the system's social interaction is simulated and that it has no capacities of "understanding" and "feeling";
- the system does not make humans believe it has consciousness (e.g., through expressions that simulate emotions).

4.6.3 Societal Wellbeing

Requirement 60: Societal impact assessment.
During acquisition and design, deployment and implementation, and monitoring phases, assess and ensure that:

- a 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 deployment and implementation and monitoring phases, assess and ensure that:

- societal impact of the AI system's use beyond the individual end-users (such as potential 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 or 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.

4.6.4 Democracy and Strong Institutions

Requirement 62: Mitigation of impacts on democracy.
During deployment and implementation and the monitoring 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 use of the system supports democratic processes and protects against interventions that manipulates, misleads or excludes voters and distorts democratic processes;
- compliance with higher authorities to ensure corporate social responsibility within the company;
- that external ethics audits are carried out to guarantee that usage of the system 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.

4.7.1 Auditability

Requirement 63: Engagement and reporting.
In all phases, 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;
- deployment and use of the system does not interfere with the auditability of the system;
- performance indications are identified and regularly report on the outcomes, in relation to the auditing system.

Requirement 64: Compliance as culture.
In all phases, 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;
- processes that facilitate the system's auditability (such as ensuring traceability and logging of the system's processes and outcomes);
- in applications affecting fundamental rights (including safety-critical applications), the system can be audited independently;
- your organisation 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, 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.

4.7.2 Minimising and Reporting Negative Impacts

Requirement 66: Reporting Impact.
During the deployment and implementation and monitoring phases, assess and ensure that:

- a risk assessment is conducted, which takes into account different stakeholders (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, 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, are established.

Requirement 67: Minimising negative impact.
In the acquisition and design, deployment and implementation, and monitoring 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;
- a process to determine how risks and benefits are balanced, while communicating the possible side-effects and their probability/uncertainty (which is linked to communication within the Transparency requirement);
- an attempt to predict the consequences/externalities of the system's use.

4.7.3 Internal and External Governance Frameworks

Requirement 68: Impact on business.
In the management and acquisition and design phases, 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 using the system, and the intended purpose that it will serve.

Requirement 69: Identify interests and values at risk.
Assess and ensure:

- a process 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.

4.7.4 Redress

Requirement 70: Redress mechanisms.
In the deployment and implementation phase, assess and ensure:

- a set of processes that allows for redress in case of the occurrence of any harm or adverse impact;
- where possible, processes to provide information to end-users/third parties about opportunities for redress.

4.7.5 Human Oversight

Requirement 72: 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 processes 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,[28] lip-read what individuals are saying,[29] track people's activities across space,[30] 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 endanger risk creating the self-fulfilling prophecy: locations where more crime is detected will be monitored more thoroughly, thus identifying more crime, 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 led 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 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

[28] 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/

[29] 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/

[30] 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.

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.[31] 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.[32] 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.[33] 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,[34] 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;
- 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;

[31] 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.

[32] Costa, Fabricio F., "Big data in biomedicine", *Drug discovery today,* Vol. 19, No. 4, 2014, pp. 433-440., p. 437.

[33] 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.

[34] 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.

- 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.[35] 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.[36]

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;

[35] 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.

[36] 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.

- 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.[37] 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;

- 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;

[37] 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 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.[38] 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;
- Compliance certification (external) and internal audit should be ensured.

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

Glossary

The glossary of terms comprises several currently available glossaries in the area, combining a mixture of research ethics, information technology, and our own classification of terms. Glossary terms are in bold. Following the terms, in square brackets are abbreviations, and in parenthesis is the reference to the source or originating author of that term, although alterations may have been made from the original source to account for context. Where there is no source, the definitions are our own.

Accountability (PRO-RES 2019)
Taking responsibility for your actions rather than trying to shift responsibility (or blame) elsewhere. This involves being able to explain the reasons behind your actions when necessary, and being prepared to discuss your actions and their consequences. It implies a willingness to accept and act on criticism of your actions where that is justified. Accountability is a central principle of Indigenous research axiology [see also Axiology and Transparency].

Agent (IEEE 2017)
An intelligent being who acts by will, from intention, whether for its own ends or those of other agents.

Agency (IEEE 2017)
The performance of actions by an agent acting from intention to meet particular ends.

Aggregation (Lean Methods Group 2019)
A process of searching, gathering, and presenting data.

Algorithm (Council of Europe 2019)
A finite suite of formal rules (logical operations and instructions) allowing one to obtain a result from input elements. This suite can be the object of an automated execution process and rely on models designed through machine learning.

Anonymise (EUPATI 2019)
The removal of personally identifiable information (such as names or addresses of clinical trial participants) so that people using the data cannot identify participants. Anonymised data contain no information that could reasonably be used by anyone to identify individuals, even by cross-checking the data against other sources of information. Anonymous data are data that have been anonymised.

Anonymity (PRO-RES 2019)
Having identifying details withheld from third parties. N.B. Committees or boards that grant ethics approval usually require researchers to maintain participants' anonymity. However, this is not always ethical in itself. For example, Kristen Perry (2011), who worked with Sudanese refugees in America, found they became upset and angry when she told them she would use pseudonyms for them in her publications. On further investigation, she found that the repressive majority regime in Sudan would force name changes on people from the minority. While anonymity is appropriate in many, if not most, cases, some research participants may have compelling reasons for wanting to be named. These should be recognized and respected.

Anomaly (Lean Methods Group 2019)
A rare or unexpected item or event in a dataset that does not conform to other items in the dataset and does not match a projected pattern or expected behaviour. Anomalies are also called outliers, exceptions, surprises or contaminants and they often provide critical and actionable information.

Artificial Intelligence [AI] (Council of Europe 2019)
A set of sciences, theories and techniques whose purpose is to reproduce by a machine the cognitive abilities of a human being. Current developments aim, for instance, to be able to entrust a machine with complex tasks previously delegated to a human. However, the term artificial intelligence is criticized by experts who distinguish between "strong" AI (which are able to contextualize very different specialized problems completely independently) and "weak" or "moderate" AI (which perform extremely well in their field of training). According to some experts, "strong" AI would require advances in basic research to be able to model the world as a whole and not just improvements in the performance of existing systems.

Artificial Neural Networks [ANN] (AI Trends 2019)
Learning models based on biological neural networks present in the brains of animals. Based on the activity of neurons, ANNs are used to solve tasks that would be too difficult for traditional methods of programming.

Assistive Technology (ICT4LT 2019)
Computer software or devices used by people with special needs to enable them to access the services provided by ICT, e.g. computer programs, email, and the World Wide Web. Technologies under this heading include Text To Speech (TTS) screen readers for the unsighted or partially sighted, alternative keyboards and mice for people who have problems in hand-eye coordination, head-pointing devices, speech recognition software, and screen magnification software.

Audit (PRO-RES 2019)
A type of research that aims to reveal the efficiency, effectiveness or quality of an institution, community, or other entity by examining its use of resources and assets relative to costs and outputs. Although this stance has been contested (Gawande 2007), audit is generally not thought to require ethics regulation because it is being conducted for internal organizational processes of accountability or quality assurance. However, it presents similar ethical issues to other types of research and these should be taken into account. Additional ethical problems can arise when audit moves from being a tool of measurement to a tool of manipulation, as in the 'audit culture' where audit becomes 'a central organizing principle in the governance and management of human conduct' (Shore 2008).

Autonomy (PRO-RES 2019)
"The capacity of a person to govern him or herself, on the basis of reasoned decisions and free from controlling influences by others. Autonomy is widely held to involve the capacity for reason and understanding, a degree of self-control, and freedom from coercion and manipulation" (Hughes et al 2010). As a principle, autonomy is, in general, highly valued in European and other Western countries, while collectivity is valued more highly in the global South (Kara 2018). The implication of this for ethical research practice in Western settings is that the autonomy of participants and other stakeholders should be respected. In other settings, researchers may need to make a judgement about the implications of

respecting collective values, as expressed by community leaders, on hearing views from other members, who may be marginalized or disadvantaged by current arrangements.

Bias (PRO-RES 2019)
"A tendency, inclination, or prejudice toward or against something or someone" (Psychology Today 2019). Biases can be positive, such as a bias towards caring for others or away from crime. However, biases tend to be based on assumptions rather than evidence or logical thought. This means that bias can lead to poor judgement and discriminatory action. A combination of individual biases and other factors can lead to structural biases, such as the well-known publication bias towards positive research findings. There are two main ways in which individual researchers can tackle bias: Debiasing and Reflexivity.

Big Data (Council of Europe 2019)
Refers to a large heterogeneous data set.

Black-box (Rouse 2019)
Refers to a system whose behavior has to be observed entirely by inputs and outputs. Black box testing assesses a system solely from the outside, without the operator or tester knowing what is happening within the system to generate responses to test actions. Even if the internal structure of the application under examination can be understood, the tester chooses to ignore it.

Chatbot (Council of Europe 2019)
Conversational agent that engages in dialogue with its user.

Classification (Data Science Glossary 2019)
Identification which of two or more categories an item falls under. Deciding whether an email message is spam classifies it among two categories (spam or not-spam), and analysis of data about movies might lead to classification of them among several genres.

Coefficient (Data Science Glossary 2019)
When graphing an equation such as y = 3x + 4, the coefficient of x determines the line's slope.

Computation (IEEE 2017)
Computation is the integration of numerical simulation, mathematical modeling, algorithm development and other forms of quantitative analysis to solve problems that theorization, experimentation, and/or observation cannot.

Consciousness (IEEE 2017)
The state or ability to be aware of self and environment.

Conflict of Interest (PRO-RES 2019)
Conflicts of interest occur when personal, financial, political and academic concerns coexist and the potential exists for one interest to be illegitimately favoured over another that has equal or even greater legitimacy, in a way that might make other reasonable people feel misled or deceived. Conflicts of interest reside in situations not behaviour and may arise even when there has been no misconduct. Researchers caught in a conflict of interest risk appearing negligent, incompetent, or deceptive. Conflicts of interest

also exist at an institutional level, where research organizations accept funding from sources that may appear to compromise the independence and integrity of their research.

Confidentiality (PRO-RES 2019)
A legal and ethical obligation imposed on the recipient of information provided by another person (the provider) not to use that information for any purpose other than that for which the information was provided. The obligation can arise from a relationship or from a contract and applies to all information that is provided and that is not publicly available, irrespective of whether the information identifies the provider. The relationships that are recognised to involve the obligation are typically between professionals and clients or patients. The obligation can be superseded by the need to use the information to serve the public interest, such as the protection of children, prevention of the spread of infectious disease, the conduct of court proceedings and the investigation of serious criminal offences.

Consent (PRO-RES 2019)
Consent refers to the agreement to do something without coercion, force, or threat. This should be based on an understanding of the research project and its aims. In theory consent should be freely given or withheld, as the potential participant prefers. In practice some research, such as national censuses in some countries, is mandatory and those who do not participate may be punished (Kara 2018). The granting of consent is often treated as an event, where in fact it is more of a process that is negotiated and renegotiated as research progresses. Indeed, in some cases renegotiation may be essential, such as in research with people with cognitive impairment or in longitudinal research. (See also Dignity and Respect).

Correlation (Data Science Glossary 2019)
The correlation coefficient is a measure of how closely the two data sets correlate. For example, if sales go up when the advertising budget goes up, they correlate. A correlation coefficient of 1 is a perfect correlation, .9 is a strong correlation, and .2 is a weak correlation. This value can also be negative, as when the incidence of a disease goes down when vaccinations go up. A correlation coefficient of -1 is a perfect negative correlation.

Corruption (PRO-RES 2019)
In practice, corruption can take the form of bribery, nepotism or misappropriation. Although corruption is the focus of international policies and strategies for its elimination, such as the UN Convention against Corruption (United Nations 2003) and the European Commission's (2019) summary of its policy, it is a complex concept and difficult to define with precision. Common elements include the exercise of a public duty for a benefit provided to the duty holder by a person who gains a reciprocal benefit from the wrongful exercise of the duty of that duty holder; dishonest or fraudulent conduct by those in power, typically involving bribery. Other definitions focus on the abuse of a trust, generally involving public power, for private gain usually in the form of money or on exclusion from an opportunity to participate in open, competitive, and fair political and economic processes (Johnston, 1996). Another recurrent element is that corruption is consciously unfair or discriminatory and permits persons holding power to decide without competition, and through covert considerations, who gets what he or she wants or needs (Rotberg, 2017).

Data (Lean Methods Group 2019)

A quantitative or qualitative value. Common types of data include sales figures, marketing research results, readings from monitoring equipment, and user actions on a website, market growth projections, demographic information and customer lists.

Data Aggregation (Lean Methods Group 2019)
The process of collecting data from multiple sources for the purpose of reporting or analysis.

Database (Council of Europe 2019)
A database is a "container" storing data such as numbers, dates, or words, which can be reprocessed by computer to produce information.

Data Governance (Lean Methods Group 2019)
The policies, rules, processes and behaviour that determine how powers are exercised concerning data. Ideally, governance should ensure data integrity is maintained and that best practices are met.

Data Integrity (Lean Methods Group 2019)
The measure of trust in the accuracy, completeness, timeliness, and validity of the data.

Data Manipulation (EUTAPI 2019)
The process of taking data and manipulating (reformatting) them to be easier to read or better organised.

Data Merging (EUTAPI 2019)
A process that involves combining data from different sources, and providing users with a single view of these data.

Data Mining (EUTAPI 2019)
The practice of searching through existing large sets of data to find useful patterns or trends.

Data Ownership (Techopedia 2019)
Data ownership is the act of having legal rights and control over data. It defines and provides information about the rightful owner of data assets and the acquisition, use and distribution policy implemented by the data owner.

Data Science (Council of Europe 2019)
A broad grouping of mathematics, statistics, probability, computing, data visualization to extract knowledge from a heterogeneous set of data (images, sound, text, genomic data, social network links, physical measurements, etc.). The methods and tools derived from artificial intelligence are part of this family.

Deep Learning (AI Trends 2019)
A subset of machine learning that uses specialized algorithms to model and understand complex structures and relationships among data and datasets.

De-identification (Resnik 2015)
The stripping of personally identifying information from data.

Developers
Individuals and groups of individuals designing, developing, and essentially creating computer software and application.

Dignity (PRO-RES 2019)
The intrinsic importance and value that a person has, that warrants respect from other people for that person; having a state or quality of being worthy of honour or respect. This reflects two historical and conceptual origins: the theological idea of inherent human value and the humanist idea of the respect owed to a rational being.

Discrimination (IEEE 2017)
Differentiation for the purpose of separating persons to determine entitlements, rights, or eligibility.

Diversity (PRO-RES 2019)
The inclusion of a wide range of people and groups in an activity. The term may refer to the inclusion of people with different ethnic heritages, socio-economic origins, genders, sexual orientations, geographical locations, demographic characteristics, or other socio-cultural markers of difference. Diversity may be a consideration in considering both research participation and research teams. Where the aims of a research project include drawing conclusions from a sample that will be relevant for other populations, evidence of the extent and nature of diversity in the sample population will be an important consideration. This will require recognition that traditional inclusion criteria have often excluded children, women, and women of child-bearing age, prisoners, undocumented immigrants and people with physical, intellectual or emotional disabilities (Mertens and Ginsberg 2009). In some circumstances, diversity will not be essential, as in case control studies, or where the focus is on problems, like thalassemia or sickle cell anaemia, that only affect specific population groups. Diversity may also be a consideration in determining whether a research team is appropriately constituted for the task, particularly where it is necessary to engage with minority communities or other groups who may be underrepresented within the researcher workforce.

Dual Use (PRO-RES 2019)
Research that, based on current understanding, can reasonably be anticipated to provide knowledge, information, products, or technologies that could be used to produce harm, often in the form of a threat to public health and safety, plants, animals, the natural environment, or national security. Also used to describe research that can have both civilian and military benefit.

Embed Explainable AI [XAI]
An attempt to address problems with explainability in AI, to better understand systems' underlying mechanisms and find solutions.

Encryption (ICT4LT 2019)
A system of coding that helps prevent access to private information on computer networks.

End-user (ICT4LT 2019)

The final user of a piece of software or hardware, i.e. the individual person for whom the product is created, as distinct from the people who create and produce the product.

Environmental Health (NEHA 2019)
The science and practice of preventing human injury and illness and promoting well-being by identifying and evaluating environmental sources and hazardous agents, and limiting exposures to hazardous physical, chemical, and biological agents in air, water, soil, food, and other environmental media or settings that may adversely affect human health.

Equity (PRO-RES 2019)
Arrangements or distributions that are proportional to the contribution or desert and, in these senses, fair. An equitable arrangement may also treat individuals equally, where their contributions or deserts are equal, but equity is more often equated with fairness than equality.

Ethics Committee [EC] (EUTAPI 2019)
A body made up of a range of individuals to assess ethical risks and/or performance of certain situations or events.

Explainability (Gall 2019)
The extent to which the internal mechanics of a machine or deep learning system can be explained in human terms. (See Interpretability).

False negative [FN] (Google Developers 2019)
A case in which a model mistakenly predicts the negative class. For example, the model inferred that an email message was not spam (the negative class), but that email message was spam.

False positive [FP] (Google Developers 2019)
A case in which a model mistakenly predicts the positive class. For example, the model inferred that an email message was spam (the positive class), but that email message was not spam.

Fundamental Human Rights
(See Human Rights)

General Data Protection Regulation [GDPR] (EUPATI 2019)
Regulation which replaced the Data Protection Directive 95/46/EC. It was designed to harmonise data protection laws across Europe, to protect and empower all EU citizens' data protection, and to reshape the way organisations across the region approach data privacy.

Hacker (ICT4LT 2019)
A person who tries to gain access to information stored on other people's computers.

Harm (PRO-RES 2019)
Joel Feinberg (1984) defined harm as the 'defeating of an interest', where the interests of an individual are defined as 'the range of things in which that individual has a stake'.

Human Rights (IEEE 2017)
"Human rights are rights inherent to all human beings, regardless of race, sex, nationality, ethnicity, language, religion, or any other status. Human rights include the right to life and liberty, freedom from slavery and torture, freedom of opinion and expression, the right to work and education, and many more. Everyone is entitled to these rights, without discrimination" (UN, 1948).

Input (ICT4LT 2019)
Anything that goes into a computer to be processed and/or stored.

Internet of Things [IoT] (Lean Methods Group 2019)
The network of physical objects or "things" embedded with electronics, software, sensors, and connectivity to enable them to exchange data with the manufacturer, operator, and/or other connected devices. Each thing is uniquely identifiable through its embedded computing system but can interoperate within the existing Internet infrastructure.

Interpretability (Gall 2019)
The ability to discern the mechanics of a machine or deep learning system without necessarily knowing why. (See Explainability). Deep Learning models are often non-interpretable because their different layers can be hard to decipher.

K-means Clustering (Data Science Glossary 2019)
A data mining algorithm which clusters, classifies, or groups objects, based on their attributes or features, into K number of groups (clusters).

K-nearest Neighbours [KNN] (Data Science Glossary 2019)
A data mining algorithm which clusters, classifies, or groups objects, based on their similarity to nearby neighbors restricted by K as the number of neighbours or the proximity of those neighbours.

Machine Learning [ML] (Council of Europe 2019)
The process by which a computer system develops algorithms from data. Machine learning makes it possible to construct a mathematical model from data, including variables which may not be known in advance.

Metadata (Council of Europe 2019)
Data used to define, contextualize, or characterize other data.

Moral Agent (IEEE 2017)
An agent able to define and implement their meaning of good and evil.

Natural Language Processing [NLP] (AI Trends 2019)
The automated processing of human languages. NLP typically involves machine interpretation of text or speech recognition.

Neural Networks (Council of Europe 2019)

Family of machine learning originally inspired by the functioning of biological neurons and which, subsequently, came close to statistical methods. The so-called formal neuron is designed as an automaton with a transfer function that transforms its inputs into outputs according to precise logical, arithmetic, and symbolic rules. Assembled in a network, these formal neurons can quickly operate classifications and gradually learn to improve them.

Obligations (PRO-RES 2019)
Explicit or implicit requirements for the conduct of persons and organizations established by ethics principles, codes or guidelines, professional regulators, by specific conditions drawn up as part of the approval of a research design or by general law.

Open data (Council of Europe 2019)
Publicly available structured datasets.

Operating Procedures (PRO-RES 2019)
Detailed specifications of processes to be followed. (See Governance).

Organisational Users
Members of organisations that will use SIS products and implement them in practice. They may purchase SIS from third-party developers, or they may create them in-house.

Outlier (EUTAPI 2019)
An observation point that is distant from other observations, e.g. some data points will be further away from the sample mean than what is deemed reasonable. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.

Personal Data (CEN 2017)
Information relating to an identified or identifiable natural person.

Personal Data Processing (Council of Europe 2019)
Any operation or set of operations using automated processes and applied to personal data such as collection, recording, organisation, structuring, storage, adaptation or modification, retrieval, consultation, use, communication by transmission, dissemination or any other form of making available, linking or interconnection, limitation, erasure or destruction.

Privacy (PRO-RES 2019)
“The protection of: (i) control over information about oneself, (ii) control over access to oneself, both physical and mental, and (iii) control over one’s ability to make important decisions about family and lifestyle in order to be self-expressive and to develop varied relationships” (Hughes et al 2010).

Privacy-by-design (Search Encrypt 2019)
An approach taken when creating new technologies and systems whereby privacy is incorporated into tech and systems, by default.

Pseudonymisation (Council of Europe 2019)
Personal data which may not be attributed to a specific data subject without recourse to additional information, provided that the required additional information is kept separately and securely, subject to technical and organisational measures.

Redress
The ability to seek justice for felt injustices, or to set right what one felt was wrong. In the context of AI and big data systems, the ability to seek recourse of actions as a result of decisions being made about oneself by AI and big data systems.

Reliability (PRO-RES 2019)
The demonstration that repeated use of the same method produces the same, or consistent, findings.

Reproducibility (Resnik 2015)
The ability of an independent researcher to achieve the same results of an experiment, test, or study, under the same conditions.

Repurpose (ICT4LT 2019)
To reuse content or data in a way different from that which was originally intended.

Respect (PRO-RES 2019)
Treating an agent with dignity, having proper regard for the feelings, wishes or rights of another person, group, or institution.

Responsibility (IEEE 2017)
The condition or fact of being answerable or accountable for something within one's power, control, or management. It involves the duty of fulfilling an obligation; the quality of being reliable or trustworthy; the state or fact of being accountable for actions; liability for some action.

Responsible Research and Innovation [RRI] (CEN 2017)
Transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the acceptability, sustainability and societal desirability of the innovation process and its marketable products, in order to allow a proper embedding of scientific and technological advances in society.

Risk (PRO-RES 2019)
Risk is the likelihood that research will result in damage, injury, liability, loss, or any other negative occurrence to some group or individual. The risk may fall on the wider society, on some specific section of society, or on an individual.

Robustness
The capacity of a computer or information system to adapt and recover during errors or incorrect inputs. It refers to how technologically capacity to fail gracefully, with as minimal harmful repercussions and effects as possible.

Safety (PRO-RES 2019)
Safety is the effort to mitigate or manage risk, wherever it may arise. It involves measures to protect the health, wellbeing, and rights of stakeholders, including wider society. The interests of these groups may be in conflict. There will often be a trade-off between risk and safety. Some aspects of safety may be the subject of legal or regulatory obligations on those who employ researchers, regardless of the choices or preferences of individual researchers.

Security (PRO-RES 2019)
Freedom from, or protection against, potential harm (or other unwanted coercive change) from external forces. Its beneficiaries may be individuals and social groups, institutions, or whole societies. Security mostly refers to protection from hostile forces, but it has a wide range of other senses: for example, as the absence of harm (e.g. freedom from want); as the presence of an essential good (e.g. food security); as resilience against potential damage or harm (e.g. secure foundations); as secrecy (e.g. a secure telephone line); as containment (e.g. a secure room or cell); and as a state of mind (e.g. emotional security). The term may also be used to refer to acts and systems intended to provide security, as in the case of 'data security'.

Sexism (PRO-RES 2019)
Actions based on the belief or unexamined assumption that people recognized as members of one sex or gender are less intelligent, able, skilful, etc. than people recognized as members of the other sex or gender. Sexism most commonly refers to actions that assume women's capacities and abilities are inferior to those of men. It may also be extended to cover the assumptions made about those individuals who identify themselves as 'non-binary', declining to be recognized as members of any specific sex or gender.

Singularity
A hypothetical future scenario where AI becomes more intelligent than human beings, who no longer have control over the AI.

Smart Information Systems
The combination and use of artificial intelligence to analyse, describe, and predict information from big data. It focuses on narrow AI and its specific application to big data sets.

Social Cohesion
The level of bonding, solidarity, and unity felt within society.

Societal Wellbeing
The overall state, mood and level of happiness and comfort within society.

Stakeholder(s)
An individual or group who are directly or indirectly affected (either positively or negatively) by a particular action, organization, or group.

Stigmatisation (PRO-RES 2019)
The process of marking social disapproval of some individual, group or institution that is not considered to comply with the expectations of those around it.

Strong AI (AI Trends 2019)
An area of AI development that is working toward the goal of making AI systems that are as flexible, useful and skilled as the human mind.

Superintelligence (IEEE 2017)
The capacity to apprehend what is beyond the normal range of human intelligence or understanding.

Sustainability (CEN 2017)
Principle of responsible care and use of economic, social, institutional, and environmental resources so that those resources are preserved for future generations.

Traceability (IEEE 1998)
The degree to which a relationship can be established between two or more products of the development process, especially products having a predecessor-successor or master-subordinate relationship to one another.

Trade-off(s)
A compromise between two desirable or undesirable outcomes.

Training set (Google Developers 2019)
The subset of the dataset used to train a model.

Transhumanism
The idea that human beings have the potential and ability to transcend current physical and mental limitations through technology.

Transparency (PRO-RES 2019)
A lack of hidden agendas and conditions associated with some action, accompanied by the open availability of all the information required for collaboration, cooperation, and collective decision making. Agreements, dealings, practices, and transactions are open to all for verification. The implication of transparency is that every action should be scrupulous enough to bear public scrutiny. This includes clarity about the rules and reasons behind regulatory measures.

Trust (IEEE 2017)
Firm belief in the reliability, truth, or ability of someone or something; to believe or accept a statement, story, etc. without seeking verification or evidence for it.

Value (PRO-RES 2019)
In ethical thinking, value goes beyond the economic calculations that are often used to produce a single metric in order to evaluate the likely costs and benefits of actions for individuals and organizations, whether as the one acting or being acted upon.

Value, Instrumental (Resnik 2015)

Something that is valuable for the sake of achieving something else, e.g. a visit to the dentist is valuable for dental health.

Value, Intrinsic (Resnik 2015)
Something that is valuable for its own sake, e.g. happiness, human life.

Vested interests (PRO-RES 2019)
Seeking to promote a theory or misrepresent a finding for personal reasons, such as personal advancement or protection of a friend/family member. Applies both to individuals and to public and private groups and organisations.

Vulnerability (PRO-RES 2019)
A limited capacity to protect one's own interests or security from harm, exploitation or other wrongdoing. It is not a fixed property of individuals or institutions but depends upon the context and resources, material or cognitive, that are available to support the person, group or organization.

Weak AI (AI Trends 2019)
Also known as narrow AI, weak AI refers to a non-sentient computer system that operates within a predetermined range of skills and usually focuses on a singular task or small set of tasks.

Wellbeing (IEEE 2017)
With reference to a person or community: the state of being healthy, happy, or prosperous; physical, psychological, or moral welfare.

Whistleblowing (PRO-RES 2019)
The exposure of information or actions within an organization, public or private, that may be considered illegal, unethical, or abusive. Where an organization lacks transparency, whistleblowing may be an important means of revealing misconduct. As such, it often receives special legal protection against potential reprisals. Sometimes incentives may be offered to encourage whistleblowing.

Annexes

Annex 1: Survey

1 Introduction

Some of the methods used to help draft our AI and big data ethical guidelines were stakeholder engagement, feedback, and input into what makes good ethical guidelines. Prior to a 2-day workshop in Brussels on July 3rd and 4th, we sent out a short survey to participants before drafting our guidelines. The survey was sent out to 28 participants, of which 18 completed it. The survey consisted of multiple-choice questions and the option to discuss their responses further. The questions were aimed to draw out the viewpoints of a wide range of stakeholders about what they see as valuable guidelines, what should be incorporated within them and how they should relate to policy. The responses to the survey were incorporated into our guidelines and laid the groundwork for the 2-day workshop.

Before the survey was sent out, the University of Twente partners identified a range of questions for the survey participants through several brainstorming sessions in April 2019. This list of questions was narrowed down to the 11 most important ones and were distributed to the stakeholders via the SHERPA website. Respondents were given a fortnight to complete the survey, which was then evaluated in early June. The following report will outline the types of data retrieved from this survey, the insights and recommendations provided by the participants, and how these have in turn helped shape our guidelines. As one participant remarked: “As this is a growing field, it is necessary to gather many opinions/interpretations before developing strict guidelines”.

In this Annex, we present the questions used in the survey (section 2) and our analysis thereof (section 3).

2 Questions from the Survey

The rationale behind the survey was to identify what should be included or excluded from AI and big data ethical guidelines. The survey aimed at deriving both quantitative and qualitative data from the stakeholders' responses. Quantitative data was retrieved through the closed-answered questions and by calculating the division of responses amongst participants. Whereas, qualitative data was retrieved through open questions, where the participants were given the option to expand and discuss the reasons behind their answers and to provide additional feedback and insights.

The questions begin with introductory feedback about the industry that the participant belongs to, if their organisation has ethical guidelines, and which guidelines are used in their workplace (questions 1 - 3). This is followed by three questions focusing on the style, context, and the approach taken with AI and big data guidelines (questions 4 - 6). These questions were to help outline the most effective and appropriate way to frame our guidelines with what is most preferable within different industries and organisations.

Guidelines are often criticised as having 'no teeth' when it comes to implementation, so questions 7 and 8 focused on whether guidelines should be supported by legislation and current standardisation protocols in the field. The remaining questions (9 - 11) concentrated on what participants viewed as the most pressing ethical issues and principles that should be discussed within AI and big data guidelines. Overall, the 11 survey questions (see Table 1, next page) provided us with insights to begin drafting our guidelines prior to the July workshops.

No.	Questions
1	In which industry/area do you work?
2	Are ethical guidelines already in place in your line of work?
3	Which guidelines are currently in place?
4	Should ethical guidelines, for the combination of big data technology and AI, be: General and short; elaborate and detailed; or somewhere in between? Please explain your answer ...
5	Should they be: Directional with little room for interpretation; leave room for interpretation and different value systems; or somewhere in between? Please explain your answer ...
6	Should they be: The same for all industries; tailored to the values and needs of different organisations; or somewhere in between? Please explain your answer ...
7	Effective ethical guidelines for the combinations of big data technology and AI: Should be supported by legislation; can still be effective without the support of legislation; or somewhere in between? Please explain your answer ...
8	Should these ethical guidelines be supported by ISO or CEN standards? Yes; No; or Unsure Please explain your answer ...
9	Please tick all of the issues you think should be covered in a set of ethical guidelines for the combinations of big data technology and AI: Privacy; Security; Transparency/Explainability; Bias; Inequality; Safety/Health; Singularity Issues; Informed Consent; Autonomy/Freedom; Human Rights; Discrimination; Democracy; Trust; Responsibility/Accountability; Moral Reasoning for AI; Sustainability; Illegal and/or misuse; Employment; Dual use and defense applications; Justice; Ownership of Data; Power Asymmetries'; Fake News; Surveillance; Digital Divide; Economic Issues; Special issues in different application domains (e.g., agriculture, finance and insurance, government, education, healthcare); and special issues for different products and techniques, e.g., visal image processing, language processing, mobile data, deep learning, embedded AI, etc.
10	Are there additional issues that should be included?
11	What other suggestions do you have for ensuring effective ethical guidelines for these technologies?

Table 1 [Annex 1]: Survey Questions

3 Analysis

Out of the 18 stakeholders who took the survey, their backgrounds consisted of: industry (5); academic/university (4); government (4); and 'other' (5). The 'other' section consisted of individuals working in healthcare, business services, a business support organisation, independent national agency for research ethics, and an independent ethics consultant. The eclectic mix of participants ensured that the survey was not inherently biased from the start by confining itself too narrowly to one particular group of stakeholders. The survey was completely anonymous, so we did not request participants' names, associations, age, gender, or other personal information.

3.1 Guidelines being used already

Only 50% of respondents had ethical guidelines in place at work. This was somewhat surprising, as we expected this figure to be higher because of the background of individuals taking the survey - individuals interested in ethics and who were about to attend a 2-day ethics guidelines workshop. This low number, for such potentially interested respondents, may be indicative that the number of organisations incorporating ethical guidelines may be substantially lower in reality.

The types of ethical guidelines also varied depending on the area that the respondents worked. For example, 40% of the industry respondents had ethical guidelines in place, which were often the business' internal guidelines. 50% of the government respondents had ethical guidelines, which consisted of general civil service codes and data ethics frameworks. The remaining respondents used internal research ethics committees (RECs) for ethical concerns or issues within their research, or followed their national sets of ethical guidelines.

3.2 Should good AI and big data guidelines be general, detailed, or in-between?

In response to the question about whether good SIS guidelines should be 'general and short', 'elaborate and detailed', or 'somewhere in between', 11% thought that SIS guidelines should be one of the extremes, whereas 77% said that they should be somewhere in between. Those that stated the guidelines should be general and short was because they are easier to remember, and that extensive guidelines should be for specific cases or apply the general guidelines. The respondents that called for them to be elaborate and detailed emphasised the complexity of SIS required detailed guidelines. Otherwise, they would not capture the complexity and nuances implicit within the topic, thus becoming weak, vague, and of no real practical use.

A majority of respondents stated that both extremes were not sufficient approaches to adopt, but also highlighted the same benefits and issues as those favouring the extremes. If the guidelines are too short and general, they will be left open to interpretation, will not provide any guidance on how to implement in practice, and will have no real justification of why they should be adopted in the first place. If they are too short, they may look like 'ethics-washing' or will be too vague for organisations to implement accountable procedures. Whereas, most respondents also stated that if they are too long and inflexible, they will similarly be disregarded by organisations in practice. The best approach is to have a short concise outline of ethical guidelines that can be simply and precisely explained, but with further detailed and rigorous explanations following thereafter. There is a need for the main principles and norms to be

explained coherently and concisely, but with much greater elaboration and discussion of how to implement these guidelines in practice.

For the structure of the guidelines, it was reiterated that a short executive summary, providing a snapshot of the main guidelines and process to derive them, should be followed by short and concise guidelines, and subsequently, exploration and elaboration upon these. The guidelines should be sufficient in length to allow for a 'first glance' to convey the main principles to adopt for ethical SIS development and use, which will allow the reader to follow-up on what is meant by each principle in the guideline. For the guidelines to be effective for industry and policymakers, they must not be 'too theoretical', and must be written for a wider audience. They also need to be detailed enough that management can implement the guidelines and identify non-compliance in the workplace. Some of the practical suggestions to achieve this were: a checklist; a regular evaluation report; problem logging report/system in place; and an initial training-in session for staff to accommodate them with the guidelines.

3.3 Directional, open for interpretation or somewhere in-between?

In response to the question about whether good SIS guidelines should be 'directional with little room for interpretation', 'leave room for interpretation and different value systems', or 'somewhere in between', there was a close split between the latter two (44% vs. 49%), with only 5% of respondents preferring the most restrictive type of guidelines. The reason put forward for having the guidelines as directional, leaving little room for interpretation, is because if they are not well-defined, it leaves them open to interpretation and potential abuse. However, we also need to ensure to not incorporate a 'one-size-fits-all' approach, as there will be different applications, domains, and types of SIS that require different concerns and emphasis. Issues relating to emerging technologies may change over time, and guidelines need to be flexible enough to adapt.

Several respondents commented that we need value systems that have a degree of flexibility. While the core principles within guidelines should be clear, users of the guidelines need to be able to apply them in a wide array of circumstances. Multiple respondents remarked that guidelines need to incorporate a wide audience, should contain overarching principles and that it was important to clearly and practically spell out what needs to be done in practice. If there is no way to envision how they can be incorporated in different industries, then there is no way that they can have any kind of regulatory force. Some claimed that there should even be 'red lines', whereby there should be no room for compromise or misinterpretation on certain matters.

3.4 Should guidelines be tailored to different industries; the same for all Industries; or somewhere in-between

In response to the question about whether good SIS guidelines should be the same for all industries (11%), tailored to the values and needs of different organisations (33%), or somewhere in between (56%), there was an uneven split. Those that replied that guidelines should be the same for all industries stated that there should be fundamental values and principles that cover most, if not all, industries. However, they should be nuanced enough to be interpreted and adapted to different contexts and applications.

The respondents that indicated guidelines should be tailored to the values and needs of different industries, all emphasised the same point - because of the divergence in industries, their needs, and

values; their guidelines needs will vary, as well. The medical industry is different to the transportation industry, which will be different from the insurance industry. While there is a need to have general guidelines that can cover most industries, there is still a need to have industry-specific guidelines. There will be different use cases, concerns, and ways to tackle these issues. If we attempt to establish guidelines for all industries, there is the risk that the guidelines will lose context.

Many of the respondents acknowledged that perhaps there could be 'tiered' guidelines, whereby, there were broad overarching guidelines that could be applicable in most industries, but that specific guidelines could be established to build upon these and make them more context-specific for different industries. While universal guidelines are effective for establishing normative standards, there is a need for these to be further elaborated upon by sectoral bodies at different levels. An initially broad set of guidelines may be effective for establishing some common ground amongst industries, but there is a need for specialisation and niches within different industries. There should be a common template, but one that allows room for adaptation for different organisations and industries.

3.5 Should guidelines be supported by legislation or can they be effective without it?

The respondents stated that there is a difference between hard and soft legislation, and that ethical guidelines do not necessarily imply the former. Ethical guidelines do not serve the same function as legislation, but more as tools or reference points for researchers and SIS professionals. They are normative, but not necessarily legal. Guidelines are meant to represent some of the most ethically important concerns about a particular topic or area of application, but are not intended as regulatory. Often, legislation provides the bare minimum ethical standards, whereas, ethical guidelines go a step further to provide self-regulation within industry.

While there is often an abundance of legislation to support ethical guidelines in other areas, they are only minimally being implemented in the area of AI and Big Data. In the words of one of the respondents: "AI should be regulated". Ethical guidelines should be seen as the pre-emptive shift for decision-makers to implement regulation, as law is generally reactive and lags behind technological and moral development. There needs to be a closer relationship between ethical guidelines and legislation so as to not dilute the importance of the former. Ethical guidelines need "teeth" in order to be effective. If there is no legislation, sanctions, and disapprobation for misconduct, then many organisations will simply overlook the guidelines.

However, others pointed out that much of the legislation that is currently in place will sufficiently cover many of the significant harms caused by AI and Big Data analytics. It is important to find where there are gaps to ensure policy is created to amend these concerns. Existing AI laws do not adequately cover many of the issues that are/will be identified in ethical guidelines. Thus, they can work effectively to articulate shared values, which will eventually be incorporated into legislation.

3.6 Should guidelines be supported by ISO or CEN Standards?

50% of people said that SIS guidelines should be supported by ISO or CEN standards, 4% said they should not, while 46% said that they were unsure. Most of the respondents stated that standardisation was important to guide organisations and ensure that they were following best practices. If supported by standardisation, ethical guidelines may provide extra strength, while also allowing standardisation to build

upon the broader values expressed in the guidelines. Most indicated that they would support ISO and CEN standardisation of ethical guidelines for many of the same reasons expressed with legislation in Section 1.2.5.

3.7 Ethical Issues to Include in the guidelines and additional feedback

We drafted a list of 28 potential ethical issues and principles for respondents to identify what should be included in our SIS ethical guidelines. The list is ordered by the most, to the least, important issues for the respondents. The table is not meant to emphasise what we believe are the most important issues in the debate, but it represents what key stakeholders believe are the most pressing issues now and should be incorporated within AI and big data ethical guidelines (see Table 2, at the end of the Annex).

In relation to additional comments and feedback that should be implemented into our guidelines, the stakeholders indicated that the target audience was important. If the guidelines are directed at Data Protection Officers, Policymakers, ICT companies, or individual developers, then the language, direction, focus and style may need to be changed accordingly. The guidelines should also consider ways to incorporate many of the current initiatives already taking place, rather than trying to reinvent the wheel. The guidelines should be closely linked to EU policy, such as the GDPR, and should incorporate commitments to policies and standardisation in business practices. One such example of this is to adopt the approaches such as the precautionary principle, or current guidelines in the area, such as relating to transparency, such as the one from the Open Ethics Initiative's three domains: "1. Algorithmic accountability and Source Code (Open Source Code, Proprietary Source Code); 2. Decision Space and its restriction (Restricted Decision, Unrestricted Decision); 3. Data Source (Open Data, Proprietary Data, Limited Access Data)".

Issue	No. Considered Important
Transparency/Explainability	17
Bias	17
Responsibility/Accountability	17
Privacy	16
Ownership of Data	15
Security	14
Safety/Health	14
Human Rights	14
Trust	14
Power Asymmetries	13
Informed Consent	12
Dual use and defence applications	12
Inequality	11
Autonomy/Freedom	11
Discrimination	11
Sustainability	9
Illegal and/or misuse	9
Special issues in different application domains (e.g., agriculture, finance and insurance, government, education, healthcare)	9
Moral Reasoning for AI	8
Surveillance	8
Special issues for different products and techniques (e.g., visual image processing, language processing, mobile data, deep learning, embedded AI).	8
Democracy	6
Employment	6
Fake News	6
Justice	5
Digital Divide	5
Singularity Issues	4
Economic Issues	4

Table 2 [Annex 1]

Annex 2: Analysis of Guidelines

1. Introduction

This is a write-up of the first comparative analysis of 25 existing guidelines on AI and big data. It is attached in this report to give a background to the first steps in producing these guidelines.

As mentioned earlier, we surveyed over 70 documents, which was then reduced to 25 by applying eight selection criteria (see table 1, below). In this annex, we will present a brief overview of the first analysis of these 25 guidelines.

Criteria
Language: The document should be in English, or have an official translation in English.
Date: The document should be from 2012 or later, because of the pace of developments in AI.
Ethics focus: The document, or at least a large part of it, should have a clear ethical focus.
AI or Big Data focus: The document should have a focus on AI and/or Big Data.
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).
Guidance: The document should provide clear guidelines, norms or proposals for behaviour.
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.
Recognition and endorsement: The document is widely known, cited and/or used, and/or endorsed by important industry sectors, multinationals, organisations or governments.

Table 1 [Annex 2] Selection Criteria for Ethical Guidelines on AI and Big Data Systems

The guidelines were analysed in the following ways. First a listing with the general information such as the author, exact name etc. was made. Second a grid (Grid 1) containing further information such as focus, target group, implementation etc. was developed. Thirdly, to gain more insight in the moral values used in the guidelines a second grid (Grid 2) was created. It contains the values that were mentioned (and not always deeply elaborated) in the various guidelines. Four, a grid (Grid 3) with the most important values was made and put into relationship with the institutions that consider them important. In the last step, a list of values was compiled that contains the value and its transformation into the respective guidelines developed by the various institutions. Here only the values that were elaborated more deeply were taken.

2. General information

Below we list and enumerate the 25 guidelines that match our selection criteria.

The general information on the guidelines include the following points:

1. Author (organisation)
2. Name of guideline
3. Year of publication
4. Weblink

Guideline 1 (abbreviated as AI4people)
Organisation: AI4People (initiative of Atomium - European Institute for Science, Media and Democracy)
Title: AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations
Year of publication: 2018
Weblink: https://www.eismd.eu/wp-content/uploads/2019/03/AI4People's-Ethical-Framework-for-a-Good-AI-Society.pdfinformation

Guideline 2 (abbreviated as UNI Global Union)
Organisation: UNI Global Union
Name of guideline: Top 10 Principles for Ethical Artificial Intelligence
Year of publication: 2017
Title and weblink: Guidelines for the use of SIS, especially by organisations, http://www.thefutureworldofwork.org/docs/10-principles-for-ethical-artificial-intelligence/

Guideline 3 (abbreviated as Public Voice)
Organisation: The Public Voice
Name of guideline: Universal Guidelines for Artificial Intelligence
Year of Publication: 2018
Title and weblink:
https://thepublicvoice.org/ai-universal-guidelines/

Guideline 4 (abbreviated as Algo.rules)
Organisation: Bertelsmann Stiftung and iRights.Lab
Name of Guideline: Algo.Rules - Rules for the Design of Algorithmic Systems
Year of publication: 2019
Title and weblink: https://algorules.org/en/home/

Guideline 5 (abbreviated as SIIA)
Organisation: Software & Information Industry Association (SIIA)
Name of Guideline: Ethical Principles for Artificial Intelligence and Data Analytics
Year of publication: 2017
Title and weblink: https://www.siia.net/Press/SIIA-Releases-Ethical-Principles-for-Artificial-Intelligence-and-Data-Analytics-with-Support-from-the-Future-of-Privacy-Forum-and-the-Information-Accountability-Foundation

Guideline 6 (abbreviated as CERNA)
Organisation: Allistene - Digital Sciences and Technologies Alliance. CERNA - Committee for the Study of Research Ethics in Digital Sciences and Technologies
Name of Guideline: Research Ethics in Machine Learning
Year of publication: 2018
Title and weblink: https://hal.archives-ouvertes.fr/hal-01724307

Guideline 7 (abbreviated as EGE)
Organisation: European Group on Ethics in Science and New Technologies (EGE)
Name of Guideline: Statement on artificial intelligence, robotics and 'autonomous' systems
Year of publication: 2018
Title and weblink: https://doi.org/10.2777/531856

Guideline 8 (abbreviated as OECD)
Organisation: OECD Legal instruments
Name of Guideline: Recommendation of the Council on Artificial Intelligence
Year of publication: 2019
Title and weblink: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Guideline 9 (abbreviated as IEEE)
Organisation: IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
Name of Guideline: Ethically Aligned Design. Version 2 - For Public Discussion
Year of publication: 2018
Title and weblink: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Guideline 10 (abbreviated as ADMA)
Organisation: Association for Data-driven Marketing and Advertisting (ADMA)
Name of guideline: Best practice guideline: Big Data
Year of publication: 2013
Title and weblink: http://datascienceassn.org/sites/default/files/Big%20Data%20Best%20Practice%20Guideline.pdf

Guideline 11 (abbreviated as IPCO)
Organisation: Information and Privacy Commissioner of Ontario.
Name of guideline: Big Data Guidelines
Year of publication: 2017
Title and weblink: https://www.ipc.on.ca/wp-content/uploads/2017/05/bigdata-guidelines.pdf

Guideline 12 (abbreviated as CoE-2017)
Organisation: Council of Europe. Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data
Name of guideline: Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data
Year of publication: 2017
Title and weblink: https://rm.coe.int/16806ebe7a

Guideline 13 (abbreviated as UNDG-2017)
Organisation: United Nations Development Group (UNDG)
Name of guideline: Data Privacy, Ethics and Protection. Guidance Note on Big Data for Achievement of the 2030 Agenda
Year of publication: 2017
Title and weblink: https://undg.org/wp-content/uploads/2017/11/UNDG_BigData_final_web.pdf
Additional information) This Guidance Note is not a legal document. It provides only a minimum basis for self-regulation, and therefore may be expanded and elaborated on by the implementing organizations

Guideline 14 (abbreviated as AI HLEG)
Organisation: European Commission. High-Level Expert Group on Artificial Intelligence (AI HLEG)
Name of guideline: Ethics Guidelines for Trustworthy AI
Year of publication: 2019
Title and weblink: https://ec.europa.eu/futurium/en/ai-alliance-consultation
Additional information) This document was written by the High-Level Expert Group on AI (AI HLEG). The members of the AI HLEG named in this document support the overall framework for Trustworthy AI put forward in these Guidelines, although they do not necessarily agree with every single statement in the document.

Guideline 15 (abbreviated as Barcelona Declaration)
Organisation: Outcome of a B·Debate on 8 March 2017, organized by Biocat, supported by the Obra Social la Caixa, Barcelona.
Name of guideline: Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe
Year of publication: 2017
Title and weblink: https://www.iiia.csic.es/barcelonadeclaration

Guideline 16 (abbreviated as Data Ethics Framework)
Organisation: UK Government. Department for Digital, Culture, Media & Sport
Name of guideline: Data Ethics Framework
Year of publication: 2018
Title and weblink: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/737137/Data_Ethics_Framework.pdf

Guideline 17 (abbreviated as Montreal Declaration)
Organisation: Université de Montréal. Forum on the Socially Responsible Development of AI
Name of guideline: Montreal Declaration for a Responsible Development of Artificial Intelligence
Year of publication: 2017
Title and weblink: https://www.montrealdeclaration-responsibleai.com/the-declaration

Guideline 18 (abbreviated as GAI)
Organisation: Data & Society, Latonero Mark
Name of guideline: Governing Artificial Intelligence. Upholding Human Rights & Dignity
Year of publication: 2018

Title and weblink: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

Guideline 19 (abbreviated as ICDPPC-2018)
Organisation: International Conference of Data Protection and Privacy Commissioners (ICDPPC)
Name of guideline: Declaration on Ethics and Data Protection in Artificial Intelligence
Year of publication: 2018
Title and weblink: https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf

Guideline 20 (abbreviated as ICO-2017)
Organisation: Information Commissioner's Office (ICO)
Name of guideline: Big Data, Artificial Intelligence, Machine Learning and Data Protection
Year of publication: 2017
Title and weblink: https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

Guideline 21 (abbreviated as DHSC-2019)
Organisation: UK Government. Department of Health & Social Care
Name of guideline: Code of conduct for data-driven health and care technology
Year of publication: 2019
Title and weblink: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology

Guideline 22 (abbreviated as Toronto Declaration-2018)
Organisation: Amnesty International / Access Now
Name of guideline: The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems
Year of publication: 2018
Title and weblink: https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf

Guideline 23
Organisation: Google (abbreviated as **Google Perspectives**)
Name of guideline: Perspectives on Issues in AI Governance
Year of publication: 2018
Title and weblink: https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf

Guideline 24 (abbreviated as IMB-2018)
Organisation: IBM
Name of guideline: Everyday Ethics for Artificial Intelligence. A Practical Guide for Designers & Developers
Year of publication: 2018
Title and weblink: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

Guideline 25 (abbreviated as Telefonica-2018)
Organisation: Telefónica
Name of guideline: AI Principles of Telefónica
Year of publication: 2018
Title and weblink: https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles

3. Overview of ethical guidelines

A first grid to give an overview of the following criteria was developed and the guidelines examined accordingly.

Guideline Name: The name of the guideline document or its abbreviation is given.

Focus: What is the focus of the guideline document? The options are: AI (Artificial Intelligence) or BD (Big Data) or BOTH.

Content/ target group: The guidelines are developed for the people that: a) Development and design of AI and/or BIG DATA, or b) Use AI and/or BIG DATA, or C) General (no specific focus on development or use).

User of guideline: The guidelines are developed for a) Computer scientists, or b) Policy makers, or c) End users of AI/Big Data, or d) Organisational users, or e) Society as a whole

No of pages of the guidelines was given**.**

Specific product/techniques and/or application: Any specific AI / big data product/technique mentioned? Any specific application domain (e.g. defence, law enforcement, media, healthcare etc.)? If no, leave blank. If yes, which ones, and how extensively in number of words?

Implementation plan: Does the document mention a specific plan for the implementation of the suggestions that it promotes? If the answer is no, a blank was left.

Further instruments: Are there any other instruments mentioned in the document? A blank means no.

Education / training: Does the document suggest that people (jn general or some groups of people such as students) should be educated / trained on the use of AI and/or big data? A blank means no.

Certification: Does the document mention the issuing of a certification for any reason? A blank means no.

Overview of Grid:

Guideline name (abbreviation)

Focus of the guideline

- AI
- Big data
- Both

Content/target group

- Development and design (a)
- Use (b)
- General (not specified) (c)

Particular class of users (for which the guidelines are developed)

- Computer scientists (a)
- Policy makers (b)
- End users (c)
- Organisations (d)
- Society as a whole €

Number of pages

Specificity and aplication

- Specific AI or big data products or techniques
- Specific application domains (e.g. defence, law enforcement, media, healthcare etc.)

Implementation plan

- Documents mentions a plan for using

Further instruments

Education/training about SIS (jn general or some groups of people such as students)

Certification

Guideline	Focus	Content / target group	User of guideline	No. of pages	Specific product and/or field application	Implementation plan	Further instruments	Education and teaching about SIS	Certification
1. AI4People	AI	general	general	30 pages	general	Some suggestions given (multistakeholder approach, public consultation etc.)	Ethics committee with internal auditing powers; EU oversight agency; "post-release" monitoring systems; European observatory for AI; public consultation mechanisms; self-regulatory codes of conduct	AI literacy programs for consumers; creation of educational curricula and public awareness activities (schools, business, university, management)	Codes of conduct: 'ethical AI' through trust-labels
2. UNI Global	AI (and machine leaning in particular)	Design and use	Unions, shop stewards and workers, AI designers and management	10 pages	Focus on consequences on work	Global convention needed; establishment of multi-stakeholder Decent Work and Ethical AI governance bodies on global and regional levels			
3. Public Voice	AI	Design and use	General For ethical standards For national and international law For designers	4	General				
4. Algo.Rules	Algorithmic systems	Design and use (for use of decision makers only)	People who research the development and use of algorithmic systems; • People who collect, collate, categorize,	6	Algorithmic systems	yes			

Guideline	Focus	Content / target group	User of guideline	No. of pages	Specific product and/or field application	Implementation plan	Further instruments	Education and teaching about SIS	Certification
			bundle and curate data; People who make decisions regarding the use and objective of an algorithmic system; • People in management at institutions or businesses that use or develop algorithmic systems; • People who program algorithmic systems; • People who test, develop and adapt algorithmic systems; • People who design how users are presented with the results of an algorithmic process ; • People who use algorithmic systems in their daily work to make decisions or determine how they will be applied.						
SIIA	AI and Big data	Responsible use of data	Policymakers, organizations developing and using data and models,	24	Data practices; issue of data and analytical models that might have a	Some information	Independent review board; public oversight system		

Guideline	Focus	Content / target group	User of guideline	No. of pages	Specific product and/or field application	Implementation plan	Further instruments	Education and teaching about SIS	Certification
			activists, scholars, ethicists, and civil societies		disparate impact; application field: general				
CERNA	Machine Learning	Design	IT researchers, developers, and designers	49 pages	Machine learning: personalised recommendation, chatbots, autonomous vehicles, robots that interact with people (taken as examples of ML)	No	No	No	No
EGE	AI, Robotics, Autonomous systems	design, production, use and governance	general	24 pages	General (Examples explicitly cited: autonomous driving, weapon systems, bots)	Harmonised European (and global) approach, wide-ranging and systematic public engagement and deliberation on the ethics implementation. of AI, platform for joining together the diverse global initiatives, common, internationally recognised ethical and legal framework for the design, production, use and governance of AI	No	Facilitate training in STEM and digital subjects	No
OECD	AI	General	Additional recommendation for policy makers	8 pages	No	Yes	Yes	Yes	no
IEEE	AI	General (slightly focusses on design)	General (anyone involved in the research, design, manufacture, or messaging around intelligent and	266 pages Very detailed, Very well structured	general Defence (autonomous weapons systems)	Yes (e.g. Multistakeholder approach, teach and involve the public); interdisciplinary approach, bringing	Independent internationally coordinated ethics expert bodies; Review boards;	Ethical aspects should be taught from school age to university;	Development of universal/ international standards (e.g. a 'safety mindset'

Guideline	Focus	Content / target group	User of guideline	No. of pages	Specific product and/or field application	Implementation plan	Further instruments	Education and teaching about SIS	Certification
			autonomous systems, including universities, organizations, governments, and corporations making these technologies a reality for society)		Affective Systems Mixed realities	together humanities, social sciences, science, engineering, and other discipline	Codes of ethics and conduct for AWS	Design of an academically oriented handbook/ educational material based on this report; Educate the public on societal impacts of AI;	for designers); review boards

Guideline	Focus	Content / target group	User of guide line	No. of pages	Specific product and/or field application	Implementation plan	Further instruments	Education and teaching about SIS	Certification
Best practice guideline : Big Data (ADMA)	BD	Development and design, Use	End Users, Organisations	39	Specific product:creation of a Single Customer View (SCV) Spec. Technique: Responsible big data business practices - check list Application: Marketing	How to approach a Big Data project p16 (words) Responsible Big Data business practices – check list p36	Australian Privacy Principles (APP)		
Big Data Guidelines	BD	Use	Policy makers	22	Application: , public health and the provision of health care,(20 words) detection of fraud (15 words)		The 1980 report of the Ontario Williams Commission (Public Government for Private People: The Report of the Commission on Freedom of Information and Individual Privacy) Canadian Charter of Rights and Freedoms		
Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data	BD	Use	Policy Makers	6	Application: policy maiking strategies to the parties of the convention		Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data	Yes, parties […] consider info and digital literacy as essential educ. skill.	
Data Privacy, Ethics and Protection. Guidance Note on Big Data for Achievement of the 2030 Agenda	BD	Use	Policy Makers	15		Guidance Note be implemented through more detailed operational guidelines that account for the implementation of UNDG member organizations'	Quadrennial Comprehensive Policy Review of the operational activities for development of the UN system, Transforming our World: The 2030		

							Agenda for Sustainable Development, UN Guidelines for the Regulation of Computerized Personal Data Files		
Ethics Guidelines for Trustworthy AI	AI	Development and design Use	Society as a whole	39	Specific product: Trustworthy ai Specific technique: Trustworthy AI assessment list, Technical and non-technical methods to realise Trustworthy AI	Technical and non-technical methods to realise Trustworthy AI	European Group on Ethics in Science and New Technologies ("EGE")	Yes. Trustworthy AI encourages the informed participation of all stakeholders.	Yes. To organisations that can attest to the broader public that an AI system is transparent, accountable and fair
Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe	AI	Use	Computer scientists End Users Organisations	9	Specific product: code of conduct for AI practitioners in Europe	Call for EU funding agencies and companies to invest in the AI development particularly in the creation of a complete ecosystem with a network of high end research labs with sufficient structural (as opposed to project-based) funding, diffusion of AI techniques to form a significant number of AI practitioners, and proper conditions and stimuli for successful AI entrepreneurship. Of particular importance is the development of open resources, such as corpora, ontologies and software frameworks, which should be available as a common infrastructure on which specific applications get built.			Yes. The European Parliament has recently decided to create an agency for robotics and AI which could potentially take up this task.

Data Ethics Framework	BD	Use	Computer scientists Policy Makers Organisations	38	Specific product: Data Ethics Workbook		Government Social Research (GSR), guidelines for using social media data responsibly in research. The GSR have also published a social media ethics grid UK Statistics Authority Quality Assurance of Administrative Data framework		
Montreal Declaration for a Responsible Development of Artificial Intelligence	AI	Development and design	Computer scientists, Policy Makers, End Users, Organisations	21	Specific product: The Montreal Declaration				
Governing Artificial Intelligence. Upholding Human Rights & Dignity	AI	Development and design Use	Computer scientists , Policy Makers, End Users, Organisations		Application: Business: (700 words) Civil Society (200) Governments (400)	How a human rights approach could be practically implemented through policy, practice, and organizational change. Article sets some initial recommendations	Lists plethora of useful instruments can be found in notes section.	Yes, through academia	

Declaration on Ethics and Data Protection in Artificial Intelligence	AI	Development and design Use	Society as a whole	6	Specific product: calls for common governance principles on artificial intelligence, establishes working group on Ethics and Data Protection in Artificial Intelligence			Yes, through investing in awareness raising, education, research and training	
Big Data, Artificial Intelligence, Machine Learning and Data Protection	BD	Use	Society as a whole	113			Damiani, Ernesto et al. Big data threat landscape and good practice guide. ENISA, Forrester Consulting. Big data needs agile information and integration governance. Forrester Research Inc, August 2013, Information Commissioner's Office. Anonymisation: managing data protection risk code of practice. ICO, November 2012, D'Acquisito, Giuseppe et al. Privacy by design in big data. An overview of privacy enhancing technologies in the era of big data analytics. ENISA, December 2015.		Yes, ICO idea on setting up a privacy seals scheme for data protection and mentions GDPR encouragement of "establishment of data protection certification mechanisms
Code of conduct for data-driven health and care technology	BD, AI	Use	End Users	20	Application: Health and care system Specific product: NHS code of conduct		The Data Ethics Framework, published by the Department for Digital, Culture, Media		

			Organisations				and Sport (mentioned), Digital Assessment Questionnaire (DAQ), Data Sharing and Privacy Toolkit, data quality and maturity index, Data Security and Protection Toolkit, The Department for Digital, Culture, Media and Sport has published a code of practice for consumer 'internet of things' (IoT) security		
The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems	AI	Development and design	Computer scientists Policy Makers	16	Specific product: Toronto Declaration Application: Machine learning systems		FAT/ML Principles for Accountable Algorithms and a Social Impact Statement for Algorithms,IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design, The Asilomar AI Principles, developed by the Future of Life Institute	Yes , public bodies must carry out training in human rights and data analysis for officials involved in the procurement, development, use and review of machine learning tools.	
Perspectives on Issues in AI Governance	AI	Development	Society as a whole	30		In regards to explainability: create a scale illustrating different levels of			Yes, Governments could

		and design Use				explanations. This scale could be used as a yardstick for setting minimum acceptable standards in different industry sectors and application contexts			collaborate to establish safety certification marks
Everyday Ethics for Artificial Intelligence. A Practical Guide for Designers & Developers	AI	Development and design	Computer scientists	33			IEEE course called "Artificial Intelligence and Ethics in Design		
AI Principles of Telefónica	AI	Development and design Use	Society as a whole	10		a methodology 'Responsible AI by Design', to be used internally			

Table 2 [Annex 2] Overview of Guidelines

4. Guidelines and values mentioned

In this grid the values that are mentioned in the different guidelines are listed. For each guideline the corresponding values / moral principles are specified.

Guideline	Values mentioned
AI4people	Beneficence Non-maleficence Autonomy Justice (distribution of resources and accountability) Explicability (incorporating intelligibility and accountability)
EGE	Human dignity (recognition of the inherent human state of being worthy of respect) Autonomy Transparency Predictability Responsibility Non maleficence Justice, equity, and solidarity Democracy Rule of law and Accountability Security, safety, bodily and mental Data protection and privacy Meaningful human contact
UNI Global	Beneficence Non maleficence Sustainability Responsibility Transparency Accountability (in the context of transparency) Fairness (unbiased) Justice (distribute benefits broadly and equally)
Public Voice	Beneficence Non maleficence (minimise risks) – above all Protection of human rights Fairness Accountability Transparency Explanation Human Determination Identification Fairness Responsibility by individuals Accuracy, reliability, and validity Safety Security
Algo.Rules	Competency Accountability/responsibility

	Beneficence Non-maleficence Security Identification principle Intelligibility Protection of individual rights
SIIA	Beneficence Sustainability Non maleficence Fairness (non-discrimination) Accountability Transparency Justice (fair share of benefits and burdens) Equality Human rights Respect for persons and law and human rights Accuracy of data Due diligence
CERNA	Trustworthiness (the systems behave as their designers claim) Equity (system treats all its users equitably) Fairness Transparency Traceability Explainability (ensured by traceability) Responsibility Compliance (to specification of system and law) Consent
OECD	Inclusive growth, sustainable development and well-being Human-centred values and fairness (freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social Justice, and internationally recognised labour rights) Transparency and explainability Robustness, security and safety Accountability
IEEE	Beneficence (superset of human rights) Non maleficence (minimise risks) Awareness of dual use Privacy and individual access control Informed consent Non-Discrimination/ Fairness Diversity Security, safety Human rights (protection of) Transparency Accountability Dignity Equality (equal access)

	Responsibility Participation, Empowerment
Best practice guideline : Big Data (ADMA)	Privacy Transparency Responsible use of data
Big Data Guidelines	Fairness Data minimisation Data quality principle Privacy Transparency No bias Non discrimination Protection of individuals in regards decisions made based on Big Data processing
Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data	Privacy, Data Protection Responsible use of data Non-discrimination, Transparency Ethics, No bias Consent Protection of individuals in regards decisions made based on Big Data processing Data minimisation
Data Privacy, Ethics and Protection. Guidance Note on Big Data for Achievement of the 2030 Agenda	Fairness Data minimisation Responsible use of data Privacy Ethics Data quality Non discrimination No bias Accountability Transparency
Ethics Guidelines for Trustworthy AI	Human centric Trustworthiness Ethics Robust AI Fairness Democracy Human dignity Freedom Accountability Privacy Responsible use of data Transparency

	Societal and environmental well-being
Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe	Non-maleficence Accountability Transparency Human centric
Data Ethics Framework	Ethics Human centric Privacy Equality/non-discrimination Responsible use of data Accountability Data minimisation Data quality No bias Transparency
Montreal Declaration for a Responsible Development of AI	Human centric Privacy Accountability Transparency Protection of citizens rights Diversity Protection form harms Protection of individuals in regards decisions made based on Big Data processing Societal and environmental well-being, diversity
Governing Artificial Intelligence. Upholding Human Rights & Dignity	No discrimination Democracy Equal accessibility Privacy Data protection Transparency Non discrimination
Declaration on Ethics and Data Protection in Artificial Intelligence	Fairness Data minimisation Societal and environmental well-being Accountability Transparency Human centric Privacy No bias, Non discrimination Ethics Citizens' rights Data quality

Big Data, Artificial Intelligence, Machine Learning and Data Protection	Fairness Transparency Consent Responsible use of data/data protection Data minimisation Data quality No bias Accountability Rights of citizens Trustworthiness Privacy Ethics
Code of conduct for data-driven health and care technology	Human centric Accountability Fairness responsible use of data transparency data quality no bias Privacy
Perspectives on Issues in AI Governance	Accountability Fairness Responsible use of data/data protection Transparency Human centric Ethics Protection of individuals in regards decisions made based on Big Data Processing/AI development
Everyday Ethics for Artificial Intelligence.	Accountability Ethics Fairness No bias Transparency Diversity Human centric Privacy
AI Principles of Telefónica	Fairness Transparency Accountability, explainability Human centric Privacy Using data in a responsible manner

Table 3 [Annex 2]

5. Values in relation to number of guidelines using it

In this grid the values that were not just mentioned but explicitly elaborated in the various guidelines are presented in the first column. The institutions doing so are in the second column. This indicates the priorities of the values used.

Beneficence	AI4people, UNI Global, Public Voice, Algo.Rules, SIIA, OECD, IEEE, AI HLEG, DATA ETHICS FRAMEWORK, MONTREAL DECLARATION
Sustainability	SIIA, EGE, IEEE, UNDG-2017
Societal and environmental well being	AI HLEG, MONTREAL DECLARATION, ICDPPC-2018
Non maleficence/doing no harm (if explicitly mentioned)	AI4 people, UNI Global, IEEE, AI HLEG, Barcelona Declaration, MONTREAL DECLARATION, Google Perspectives, IMB-2018, CERNA
Autonomy	AI4people, EGE, IEEE, AI HLEG
Human centric approach	Barcelona Declaration, ICDPPC-2018, DHSC-2019, Google Perspectives, IMB-2018
Consent (and symmetry)	CERNA, IEEE, CoE-2017, ICO-2017, DHSC-2019
Justice (equal distribution)	AI4people, UNI Global, SIIA, EGE, IEEE, AI HLEG
Explicability/Explainability	AI4people, Algo.Rules, CERNA
Protection of individuals with regards to decisions made based on big data processing and AI development	IPCO, Toronto Declaration-2018
Transparency	UNI Gobal, Public Voice, SIIA, CERNA, OECD, IEEE, ADMA, IPCO, CoE-2017, UNDG-2017, AI HLEG, Barcelona Declaration, DATA ETHICS FRAMEWORK, MONTREAL DECLARATION, GAI, ICDPPC-2018, ICO-2017, DHSC-2019, Toronto Declaration-2018, Google Perspectives, IMB-2018, Telefonica-2018
Responsibility	Public Voice, UNI Global, Algo.Rules, CERNA, EGE, MONTREAL DECLARATION, Google Perspectives
Accountability (and assessment)	ALgo.Rules, SIIA, Public Voice, EGE, OECD, IEEE, UNDG-2017, AI HLEG, Barcelona Declaration, DATA ETHICS FRAMEWORK, MONTREAL DECLARATION, ICDPPC-2018, DHSC-2019, ICO-2017, Toronto Declaration-2018, Google Perspectives, IMB-2018, Telefonica-2018

No bias, no discrimination, Fairness	AI4people, UNI Global, Public Voice, SIIA, CERNA, IPCO, CoE-2017, UNDG-2017, AI HLEG, DATA ETHICS FRAMEWORK, GAI, ICDPPC-2018, ICO-2017, DHSC-2019, Toronto Declaration-2018, Google Perspectives, IMB-2018, Telefonica-2018, IPCO, MONTREAL DECLARATION
Diversity	MONTREAL DECLARATION, Toronto Declaration-2018, IMB-2018
Safety/security	Public Voice, Algo.Rules, EGE, OECD, IEEE
Trustworthiness	AI HLEG, DATA ETHICS FRAMEWORK, ICO-2017
Data quality	IPCO, CoE-2017, UNDG-2017, AI HLEG, DATA ETHICS FRAMEWORK, ICDPPC-2018, ICO-2017, DHSC-2019
Protection of citizens' rights	AI HLEG, MONTREAL DECLARATION, ICDPPC-2018, ICO-2017
Protection of human rights	IEEE (Human Rights were just mentioned in many guidelines), CoE-2017, AI HLEG
Human dignity	EGE, AI HLEG, Google Perspectives
Democracy	EGE, GAI
Data protection and privacy	EGE, IEEE, ADMA, IPCO, CoE-2017, UNDG-2017, AI HLEG, DATA ETHICS FRAMEWORK, MONTREAL DECLARATION, GAI, ICDPPC-2018, ICO-2017, DHSC-2019, IMB-2018, Telefonica-2018, CoE-2017, Toronto Declaration-2018, Google Perspectives
Personal data minimisation	IPCO, CoE-2017, UNDG-2017, DATA ETHICS FRAMEWORK, ICDPPC-2018, ICO-2017
Ethics	IPCO, UNDG-2017, DATA ETHICS FRAMEWORK, ICDPPC-2018, ICO-2017, DHSC-2019, Google Perspectives, IMB-2018

Table 4 [Annex 2]

6. Transformation of values into guidelines/recommendations

In this section the various interpretations and operationalisations for the most important values and principles are given. We describe how a value transformed into a specific requirement in the different guidelines. For each subsection we start by presenting, for each institution, a brief definition of the value (if applicable and available) and then the most important rules listed underneath.

However, many guidelines do not give clear definitions of the values they use. The definitions are listed in case they are given followed by the recommendation into which the value is transformed. Furthermore, in many guidelines it is not clear how the values relate to the recommendation or requirement.

Lastly, in most papers ethical norms are mixed with procedural norms. Accordingly, some procedural norms are included when considered important. But recommendations on particular applications are omitted.

6.1 Beneficence

All guidelines agree on the fact that AI devices must be designed for the benefit of human beings. Some of the guidelines make further explanations on how this is to be achieved. If available, a short specification of the understanding of beneficence is given also.

AI4PEOPLE

- Promote well-being, preserve dignity, and sustain the planet.
- Assess whether current regulations are sufficiently grounded in ethics to provide a legislative framework that can keep pace with technological developments. This may include a framework of key principles that would be applicable to urgent and/or unanticipated problems.
- Develop and use AI technologies within the EU that are socially preferable (not merely acceptable) and environmentally friendly (not merely sustainable but favourable to the environment).
- Include ethical, legal and social considerations in AI research projects.
- Foster socially positive innovation.
- Cross-disciplinary and cross-sectoral cooperation and debate concerning the intersections between technology, social issues, legal studies, and ethics is needed.
- Develop agreed-upon metrics for the trustworthiness of AI products and services.

UNI Global

- Artificial intelligence must put people and planet first. AI systems must remain compatible and increase the principles of human dignity, integrity, freedom, privacy and cultural and gender diversity, as well as with fundamental human rights.
- Establish multi-stakeholder Decent Work and Ethical AI governance bodies on global and regional levels. The bodies should include AI designers, manufacturers, owners, developers, researchers, employers, lawyers, CSOs and trade unions

PUBLIC VOICE

- The Guidelines are intended to maximize the benefits of AI (for human beings).

ALGO.RULES

- Algorithmic systems have to be designed for the benefit of society.

SIIA

- Secure the well-being of people affected by data practices. Maximise possible benefits and minimize possible harms.
- There is a responsibility to produce social benefits.
- Define the benefits that will be created by a data analytics project and identify the parties that gain tangible value from the effort.

OECD

- One should have inclusive growth, sustainable development and well-being as aims.
- Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments

IEEE

Wellbeing: encompassing human satisfaction with life and the conditions of life as well as an appropriate balance between positive and negative affect.

- Prioritize wellbeing Prioritize benefits to humanity and the natural environment from the use of AI.
- AI should prioritize human well-being as an outcome in all system designs, using the best available, and widely accepted, well-being metrics as their reference point.

AI HLEG

- says that AI systems need to be human centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom. In its turn the human centric approach is based on human dignity endorsed in various human rights instruments (ECHR, European Charter on Human Rights e.t.c).
- Furthermore, Users should be able to make informed autonomous decisions regarding AI systems (Human Agency), and ai systems should operate under human oversight.

DATA ETHICS FRAMEWORK

- all data processing by public sector must have a clear user and expected public benefit. A clear user need must be identified and public benefit assessed through help of HM Treasury guidelines (Green book, managing public money guidelines).

MONTREAL DECLARATION

- development and use of AI must permit the growth of well-being of all sentient beings. This includes improvement of living conditions, allow people to exercise their mental and physical capacities and AIS use should not contribute to increasing stress, anxiety, or a sense of being harassed by one's digital environment.

- Furthermore, ai systems should respect of people's autonomy including AIS must allow individuals to fulfil their own moral objectives and their conception of a life worth living and IS must not be developed or used to impose a particular lifestyle on individuals, whether directly or indirectly, by implementing oppressive surveillance and evaluation or incentive mechanisms e.t.c

6.2 Societal and environmental well being

AI HLEG

- AI must operate in the most environmentally friendly way and extensive use of AI should not alter our conception of social agency.
- Furthermore, extensive use of AI could also affect people's physical and mental wellbeing. The effects of these systems must therefore be carefully monitored and considered.
- Finally, the use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision-making but also electoral contexts.

MONTREAL DECLARATION

- development of AIS must be compatible with maintaining societal bonds among people and generations. This can be achieved through

a) AIS must not threaten the preservation of fulfill moral and emotional human relationships
b) AIS must be developed with the goal of collaborating with humans and foster collaboration between them c) AIS should not be implemented to replace people in duties that require quality human relationships
c) AIS development should not encourage cruel behavior toward robots designed to resemble human beings or non-human animals in appearance or behavior
d) AIS should not be implemented to replace people in duties that require quality human relationships
e) AIS should help improve risk management and foster conditions for a society with a more equitable and mutual distribution of individual and collective risks.

- Furthermore, development of AI must have as its aim the creation of just and equitable society and AIS development must produce social and economic benefits for all by reducing social inequalities and vulnerabilities.
- Finally, AIS development and use must be carried out in a manner so as to ensure stronger environmental sustainability. This includes

a) greatest energy efficiency and to mitigate greenhouse gas emissions over its entire life cycle,
b) AIS hardware, its digital infrastructure and the relevant objects on which it relies, must aim to generate the least amount of electric and electronic waste,
c) AIS hardware, its digital infrastructure and the relevant objects on which it relies, must minimize our impact on ecosystems and biodiversity at every stage of its life cycle,
d) public and private actors must support the environmentally responsible development of AIS.

ICDPPC-2018

- AI developers must take into consideration not only the impact that the use of artificial intelligence may have on the individual, but also the collective impact on groups and on society at large, and also ensure that artificial intelligence systems are developed in a way that facilitates

human development and does not obstruct or endanger it, thus recognizing the need for delineation and boundaries on certain uses.

- Furthermore, capabilities of artificial intelligence systems can be used to foster an equal empowerment and enhance public engagement, for example through adaptable interfaces and accessible tools.

6.3 Sustainability

Since beneficence is often used in connection with sustainability, the latter is listed at this point.

SIIA

- Favour data analytic projects that effectively predict future behaviour and generate beneficial insights over a reasonable period of time.

EGE

- AI technology must be in line with the human responsibility to ensure the basic preconditions for life on our planet and ensure the priority of environmental protection and sustainability.

IEEE

- A/IS should contribute to achieving the UN Sustainable Development Goals.

UNDG-2017

- data access, analysis or other use must be consistent with the United Nations Charter and in furtherance of the Sustainable Development Goals, pursuing lawful and legitimate and fair use.

6.4 Non-maleficence

All guidelines from EUREC sections agree on the principle of doing no harm. All guidelines mention or caution against the many potentially negative consequences of AI technologies that have to be avoided. In most guidelines this is not made explicit and/or just mentioned. The principle of non-maleficence is listed here only if it is mentioned directly since any ethical guideline can be subsummized under this principle.

AI4 PEOPLE

- The principle is mentioned as an extra principle and is to be interpreted with particular regards to privacy, security and “capability caution”.

UNI GLOBAL

- Ban arm races: Lethal autonomous weapons, including cyber warfare, should be banned.

IEEE

- Mitigate risks and negative impacts, including misuse, as A evolve as socio-technical systems. In particular by ensuring AI are accountable and transparent.

AI HLEG

- A crucial component of achieving Trustworthy AI is technical robustness, which is closely linked to the principle of prevention of harm. This includes

a) preventative approach to risks,
b) Resilience to attack and security
c) Fallback plan and general safety
d) accuracy and
e) reliability and reproducibility

BARCELONA DECLARATION

- application of knowledge-based AI requires the availability of human expertise and sufficient resources to analyse and model it.
- The application of data-driven AI requires enough high quality data and careful choices of which algorithms and parameters are appropriate in each case.
- These application prerequisites need to be investigated and spelled out in much more detail so that those responsible for applying AI can exercise the necessary prudence.
- Furthermore, all artificial systems that are used in our society have to undergo tests to determine their reliability and security.

MONTREAL DECLARATION

- Every person involved in ai development must anticipate the adverse effects of AIS use. It suggests to

a) develop mechanisms that consider the potential for the double use — beneficial and harmful —of AI research and AIS development (whether public or private) in order to limit harmful uses,
b) When the misuse of an AIS endangers public health or safety and has a high probability of occurrence, it is prudent to restrict open access and public dissemination to its algorithm,
C) AIS must meet strict reliability, security, and integrity requirements and be subjected to tests that do not put people's lives in danger, harm their quality of life, or negatively impact their reputation or psychological integrity.

GOOGLE PERSPECTIVES

- Safety concerns are one of the main reasons people give for seeking to ensure there is a "human in the loop" in AI implementations. This is based on the perception that having a person overseeing an AI system's recommendations will provide a fail-safe mechanism to protect against mistakes. Unfortunately in many instances this is a fallacy.
- In practice, it is seldom scalable to have a person checking every recommendation from an AI's system, so oversight ends up being limited to just those that the system is less sure about (i.e., that fall below a probability threshold).
- Thus fundamental mistakes about which the AI system is confident will be missed.

6.5 Autonomy

AI4PEOPLE

Autonomy: The power to decide (and whether to decide). Autonomy is also understood in the sense of "meta-autonomy", or a "decide-to- delegate".

- Strike a balance between the decision-making power we retain for ourselves and that which we delegate to artificial agents.
- Assess which tasks and decision-making functionalities should not be delegated to AI systems.
- Autonomy of humans should be promoted, but also the autonomy of machines should be restricted and made intrinsically reversible.

EGE

Autonomy: capacity of human persons to legislate for themselves, to formulate, think and choose norms, rules and laws for themselves to follow. It translates into responsibility and control over and knowledge about AI. Autonomy involves transparency and predictability as well.

- Control over AI (that systems do not impair freedom of human beings)
- Transparency (prerequisite to intervene or terminate AI if they would consider this morally required)
- Predictability (prerequisite to intervene or terminate AI if they would consider this morally required

IEEE

- Definitions of autonomy need to be clearly drawn, both in terms of AI and human autonomy.
- The critical difference between human autonomy and autonomous systems involves questions of free will, predetermination, and being (ontology).

AI HLEG

- an AI context, freedom of the individual for instance requires mitigation of (in)direct illegitimate coercion, threats to mental autonomy and mental health, unjustified surveillance, deception and unfair manipulation. In fact, freedom of the individual means a commitment to enabling individuals to wield even higher control over their lives, including (among other rights) protection of the freedom to conduct a business, the freedom of the arts and science, freedom of expression, the right to private life and privacy, and freedom of assembly and association.

6.6 Human centric approach

BARCELONA DECLARATION

- it is necessary to have clear rules constraining the behaviour of autonomous AI systems, so that developers can embed them in their applications.
- It is also necessary to clarify who is responsible for failure - as is indeed the case with all products. Furthermore, human role including intelligence and expertise should be preserved and not replaced by AI.

ICDPPC-2018

- overall human control on such systems must be existent through developer's provision of adequate information on the purpose and effects of artificial intelligence systems.

DHSC-2019

- adoption or innovation to be built; firstly, it is necessary to outline its specific use and context in which it will be used. Uses can be a) clinical, such as understanding co-morbidities b) practical, such as access to technologies, or time to spend interacting with a service c) emotional, such as needs for reassurance as well as diagnosis or treatment.
- Furthermore, any innovation must have a clear object which would add value to people and the health and care system.

GOOGLE PERSPECTIVES

- "Human in the loop" is shorthand for systems which include people at one or more points in the decision-making process of an otherwise automated system. The challenge is in determining whether and where in the process people should play a role, and what precisely that role should entail. Selecting the most prudent combination comes down to a holistic assessment of how best to ensure that an acceptable decision is made, given the circumstances.
- Similarly, the degree of choice and control that users have has an impact on the ethics of fully automated processes.
- Delegating tasks and decisions to a machine is not bad, even in high stakes settings, so long as people have meaningful choice about doing so and can revise their decision.
- Furthermore, even with advanced AI systems able to design learning architectures or generate new ideas, the choice of which to pursue should still be overseen by human collaborators, not least to ensure choices fall within an organization's legal and financial constraints.
- Finally, article outlines considerations for successful human-ai collaboration and example of Youtube case study on this matter.

IMB-2018

- it is designers and developers' responsibility to keep users empowered with control over their interactions.

6.7 Consent (and symmetry)

CERNA

- The use of interconnected machine learning systems raises an imperative of consent in the light of the impact that the learning capacities of these systems can have on individuals and groups.

Consent must be possible in the following situations.

- Researchers must include the possibility for systems to be used with or without their learning capacity.
- From the project design phase onwards, researchers must consult with people or groups identified as potentially likely to be influenced by it (the must consent to it).
- Researchers should be aware that learning capacity and the networking of such capacities can lead to new problems that affect the consent of both user and society.

IEEE

- Individuals need to achieve and retain parity regarding their personal information.
- The asymmetric power of institutions (including public interest) over individuals should not force use of personal data when alternatives such as personal guardians, personal agents, law-

enforcement-restricted registries, and other designs that are not dependent on loss of agency are available.

- When loss of agency is required by technical expedience, transparency needs to be stressed in order to mitigate these asymmetric power relationships.
- Algorithmic guardian platforms should be developed for individuals to curate and share their personal data.
- The same AI that parses and analyzes data should also help individuals understand how personal information can be used. AI can prove granular-level consent in real time.
- Where the subject does not have a direct relationship with the system, consent should be dynamic and must not rely entirely on initial terms of service or other instruction provided by the data collector to someone other than the subject. AI should be designed to interpret the data preferences, verbal or otherwise, of all users signalling limitations on collection and use, discussed further below.
- Tools, settings, or consumer education are increasingly available and should be utilized to develop, apply, and enforce consumer consent.
- In the same way that companies are doing privacy impact assessments for how individual data is used, companies need to create *employee data impact assessments* to deal with the specific nuances of corporate specific situations. It should be clear that no data is collected without the consent of the employee.
- Researchers or developers of AI have to take into account the issue of vulnerable people, and try to work out an AI that alleviates their helpless situation to prevent possible damage caused by misuse of their personal data.
- Build an AI advisory commission, composed of elder advocacy and mental health self- advocacy groups, to help developers produce a level of tools and comprehension metrics to manifest meaningful and pragmatic consent applications.

CoE-2017

- the free, specific, informed and unambiguous consent shall be based on the information provided to the data subject according to the principle of transparency of data processing and, where applicable, processors shall provide easy and user-friendly technical ways for data subjects to react to data processing incompatible with the initial purposes and withdraw their consent.
- If the consent was given in a situation where a clear imbalance of power existed which could affect the consent, controller has to prove that it did not existed.

ICO-2017

- If an organisation is relying on people's consent as the condition for processing their personal data, then that consent must be a freely given, specific, and informed indication that they agree to the processing.
- Furthermore, the data controller must be able to demonstrate that the consent was given, and the data subject must be able to withdraw that consent.
- However, article provides a recent report by the European Union Agency for Network and Information Security (ENISA) which called for more technical innovation in obtaining consent.

DHSC-2019

- if planning to use identifiable patient data in the development and/or testing of the technology, ensure that there is appropriate consent to access the data or some other legal basis, such as approval under section 251 of the NHS Act 2006 and Health Service (Control of Patient Information) Regulations 2002.
- Moreover, since May 2018 the national data opt-out allows people to opt out of their confidential patient information being used for purposes beyond their individual care and treatment. By 2020 any health and care organisation that processes and/or disseminates data that originates with the health and adult social care system in England is required to be in compliance with the national data opt-out policy (Anonymised data not included).

6.8 Justice (equal distribution)

AI4PEOPLE

Justice: Promoting prosperity and preserving solidarity (distribution of ressources)

- Use AI to correct past wrongs such as eliminating unfair discrimination;
- Ensure that the use of AI creates benefits that are shared (or at least shareable)
- Prevent the creation of new harms, such as the undermining of existing social structures.

UNI Global

Justice: equal distribution

- AI technologies should benefit and empower as many people as possible. The economic prosperity created by AI should be distributed broadly and equally, to benefit all of humanity.
- Develop global as well as national policies aimed at bridging the economic, technological and social digital divide

SIIA

Justice: equitable distribution

- Distribute the benefits and burdens of data practices equitably.
- Aim for an equitable distribution of the benefits of data practices and avoid data practices that disproportionately disadvantage vulnerable groups.
- Individuals have rights based on justice to a fair share of the benefits and burdens of social life.

EGE

Justice: fair distribution

- AI should contribute to global justice and equal access to the benefits and advantages that AI, robotics and 'autonomous' systems can bring.
- Discriminatory biases in data sets used to train and run AI systems should be prevented or detected, reported and neutralised at the earliest stage possible.
- Concerted global effort towards equal access to 'autonomous' technologies and fair distribution of benefits and equal opportunities across and within societies are needed.
- Formulate new models of fair distribution and benefit sharing.
- Ensure accessibility to core AI technologies.
- Facilitate training in STEM.

IEEE

Equal availability

- Equitable distribution of the benefits of AI technology worldwide should be prioritized.
- Most of the recommendations of how to achieve this are procedural in character and of a political or social nature. Therefore, they are not mentioned here the focus being on ethical guidelines.

AI HLEG

- trustworthy ai can foster equality through increase citizens' health and well-being in ways that foster equality in the distribution of economic, social and political opportunity**.**

6.9 Explicability/Explainability

AI4PEOPLE

Explicability: enables the principles of beneficence, non-maleficence, justice and autonomy, and includes intelligibility and (in the ethical sense) accountability

- Intelligibility: an answer to the question "how does it work?"
- Accountability: an answer to the question: "who is responsible for the way it works".
- Develop a framework to enhance the explicability of AI systems that make socially significant decisions. Central to this framework is the ability for individuals to obtain a factual, direct, and clear explanation of the decision-making process, especially in the event of unwanted consequences. This is likely to require the development of frameworks specific to different industries, and professional associations should be involved in this process, alongside experts in science, business, law, and ethics.
- Develop appropriate legal procedures and improve the IT infrastructure of the justice system to permit the scrutiny of algorithmic decisions in court. Include the creation of a framework for AI explainability specific to the legal system.
- Assess the capacity of existing institutions, such as national civil courts, to redress the mistakes made or harms inflicted by AI systems. This assessment should evaluate the presence of sustainable, majority-agreed foundations for liability from the design stage onwards, in order to reduce negligence and conflicts.
- Develop a redress process or mechanism to remedy or compensate for a wrong or grievance caused by AI.

ALGO.RULES

Explainability: understood in the sense of intelligibility.

- The function and potential effects of an algorithmic system must be understood.
- The decision-making processes within an algorithmic system must always be comprehensible. Information about the data and models on which the system is based, its architecture and potential effects must be published in easily understood terms.

CERNA

Explainability: the system's operation can be explained. It is ensured by traceability.

- Researchers should be mindful of non-interpretability or lack of explainability in the actions of a machine learning system. The compromise between performance and explainability should be assessed according to the context of use and should be set out in the documentation addressed to the trainer and the user.

- When seeking to enhance the explainability of a machine learning system, researchers should be careful to describe the limitations of their explanation heuristics and to ensure that the interpretations of their results are exempt from bias.
- Researchers should ensure that the system's results are interpretable and explainable to the human users concerned by such results. Researchers should contribute to the necessary modification in job descriptions of professionals who use the results of machine learning in the interaction with humans. Researchers should develop expert agents for explanation and verification of the behaviour of learning systems.

6.10 Protection of individuals with regards to decisions made based on big data processing and AI development

IPCO

- when using profiling as part of a big data project, you should verify the results of any decisions based solely on profiling in cases where the decisions significantly affect individuals and ensure that individuals are given the opportunity and sufficient support to challenge or respond to such decisions. The results should be verified in a manner that is independent of the predictive model or profile used.

TORONTO DECLARATION-2018

- It is critical that states provide meaningful opportunities for effective remediation and redress of harms where they do occur. This may include a) accessible and effective appeal and judicial review, b) clarify which bodies or individuals are legally responsible for decisions made through the use of such systems, c) Provide effective remedies to victims of discriminatory harms linked to machine learning systems used by public or private bodies, including reparation that, where appropriate, can involve compensation, sanctions against those responsible, and guarantees of non-repetition. This may be possible using existing laws and regulations or may require developing new ones.

6.11 Transparency

UNI Global

A transparent artificial intelligence system is one in which it is possible to discover how, and why, the system made a decision, or in the case of a robot, acted the way it did.

- AI will need to be transparent and accountable to an accident investigator, so the internal process that led to the accident can be understood.
- Workers must have the right to demand transparency in the decisions and outcomes of AI systems as well as the underlying algorithms. This includes the right to appeal decisions made by AI algorithms, and having it reviewed by a human being
- A device must be present that can record information about said system in the form of an "ethical black box" that not only contains relevant data to ensure transparency and accountability of a system, but also includes clear data and information on the ethical considerations built into said system. This includes codes of ethics for the development, application and use of AI so that throughout their entire operational process, AI systems remain compatible and increase the

principles of human dignity, integrity, freedom, privacy and cultural and gender diversity, as well as with fundamental human rights.

PUBLIC VOICE

- Individual must have the Right to know the basis of an adverse determination. This includes access to the factors, the logic, and techniques that produced the outcome of AI.
- The aim of this principle is to enable independent accountability for automated decisions.

SIIA

Transparency means disclosing the purposes of data practices, why data collection is required to fulfil the purposes, and how data will be used.

- Clearly communication risk assessment and harm minimization related to data practices.
- Organizations should disclose what data they collect, the purposes for which it is used, and which analytic techniques and models are used to process data and produce an outcome.
- Organizations should provide explanations of how advanced modelling techniques produce their results.
- Organizations should publicly describe the model governance programs they have in place.

CERNA

Transparency: a system is transparent if its operation is not hidden. It depends on traceability:

- Traceability: the availability of sufficiently detailed information on its actions (stored in a log) for those actions to be subsequently backtracked
- Researchers must ensure that machine learning is traceable, and provide protocols for that purpose. The traces are themselves data, and as such also demand ethical handling.
- Traceability ensures explainability. See explainability.

OECD

Transparency and explainability: responsible disclosure regarding AI systems

- AI Actors should commit to transparency and responsible disclosure regarding AI systems.
- To this end AI actors should provide meaningful information:
    - General understanding
    - Create awareness of interaction with machines
    - Understand outcome of AI systems
    - Make challenge of outcome of AI systems possible (in case of being adversely affected): easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision

IEEE

Transparency can be understood as traceability, as non-deception and honest design and as intelligibility. A lack of transparency increases the risk and magnitude of harm and also the difficulty of ensuring accountability.

- AI systems, and especially those with embedded norms, must have a high level of transparency, from traceability in the implementation process, mathematical verifiability of its reasoning, to honesty in appearance-based signals, and intelligibility of the system's operation and decisions.

- Based on the cultural context, application, and use of AI, people and institutions need clarity around the manufacture and deployment of these systems to establish responsibility and accountability and avoid potential harm.
- Additionally, manufacturers of these systems must be able to provide programmatic-level accountability proving why a system operates in certain ways to address legal issues of culpability, if necessary, apportion culpability among several responsible designers, manufacturers, owners, and/or operators, to avoid confusion or fear within the general public.
- If an AI causes harm it must always be possible to discover the root cause, by assuring traceability for said harm.
- Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed, and levels of compliance determined. For designers, such standards will provide a guide for self-assessing transparency during development and suggest mechanisms for improving transparency.

First, anticipating the process of evaluation already during the implementation phase requires defining criteria and metrics for such evaluation

Second, a systematic risk analysis and management approach can be useful

Third, additional mitigation mechanisms must be made available.

- Software engineers should be required to document all of their systems and related data flows, their performance, limitations, and risks. Ethical values that have been prominent in the engineering processes should also be explicitly presented as well as empirical evidence of compliance and methodology used, standards providing oversight of the manufacturing process of intelligent and autonomous technologies need to be created.
- An independent, internationally coordinated body should be formed to oversee whether such products actually meet ethical criteria.
- Engineers must acknowledge and assess the ethical risks involved with black-box software and implement mitigation strategies.
- Technologists should be able to characterize what their algorithms or systems are going to do via transparent and traceable standards.
- Technologists should seek indirect means of validating results and detecting harms.
- Software engineers should employ “black-box” (opaque) software services or components only with extraordinary caution and ethical care

ADMA

- transparency allows consumers to make informed choices about sharing their personal information and thus constituting a cornerstone element in any Big Data processing.

IPCO

- promoting openness and transparency, a description of each big data project should be published on the host institution’s website to enable individuals to become informed about how their personal information is being processed.
- Furthermore, to protect the privacy of individuals, you should consider treating personal information that is publicly available the same as non-public personal information when conducting big data projects.

- Finally, ensure that individuals who are the subject of profiling are informed of additional information regarding the nature of the predictive model or profile being used.

CoE-2017

- controllers should identify the potential impact on individuals of the different uses of data and inform data subjects about this impact.
- Furthermore, the results of the assessment process (risk assessment) should be made publicly available, without prejudice to secrecy safeguarded by law.
- In the presence of such secrecy, controllers provide any confidential information in a separate annex to the assessment report. This annex shall not be public, but may be accessed by the supervisory authorities.

UNDG-2017

- transparency is generally encouraged when the benefits of transparency are higher than the risks and possible harms. Except in cases where there is a legitimate reason not to do so, at minimum, the existence, nature, anticipated period of retention and purpose of data use as well as the algorithms used for processing data should be publicly disclosed and described in a clear and non-technical language suitable for a general audience.
- Furthermore, whenever possible, the data should be made open, unless the risks of making the data open outweigh the benefits or there are other legitimate bases not to do so. Disclosure of personal information, even if derived from public data, should be avoided or otherwise carefully assessed.

AI HLEG

- transparency is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models.
- Transparency must involve a) Traceability (data labelling as well as the algorithms used, should be documented) b) explainability (ability to explain both the technical processes of an AI system and the related human decisions) c) Communication (AI systems should not represent themselves as humans to users).

BARCELONA DECLARATION

- it must be clear whether an interaction originates from a human or from an AI system, and that, in the case of an artificial system, those responsible for it can be traced and identified. This solution could possibly be implemented by a system of water marking and become mandatory in Europe.

DATA ETHICS FRAMEWORK

- being open about your work is critical to helping to make better use of data across government. When discussing your work openly, be transparent about the tools, data, algorithms and the user need. This includes a) Documenting your work clearly is an essential part of working in the open and being accountable b)Your technology choices should support coding in the open where possible c) Discussing your work openly at events, blogging and documenting work clearly on Github. Furthermore, a) If data is non-sensitive and non-

personal, you should make it open and assign it a digital object b) Developed data science tools should be made available for scrutiny wherever possible.

- Finally, It is essential that government policy be based on interpretable evidence in order to provide accountability for a policy outcome. You should also plan how you will explain your work to others, ensuring your approach can be held.

MONTREAL DECLARATION

- ai code of algorithms must be accessible by relevant public authorities and stakeholders for verification and control purposes.
- Furthermore, the discovery of AIS operating errors, unexpected or undesirable effects, security breaches, and data leaks must imperatively be reported to the relevant public authorities, stakeholders, and those affected by the situation.
- Moreover, in accordance with the transparency requirement for public decisions, the code for decision-making algorithms used by public authorities must be accessible to all, with the exception of algorithms that present a high risk of serious danger if misused.
- It is also necessary for people to know that decisions that affects them derives from ai and any user of a service employing chatbots should be able to easily identify whether they are interacting with an AIS or a real person.
- Artificial intelligence research should remain open and accessible to all.
- Finally, results of security, reliability, integrity test must be open to the public and any errors and flaws discovered in AIS and SAAD should be publicly shared.

GAI

- mentions NYC law that aims to help ensure that algorithms used by city agencies are transparent, fair, and valid by setting up a task force to make recommendations on algorithmic regulation, transparency, and bias. While these rules apply only to New York and do not appeal to human rights directly, this move to regulate AI may become a model for other cities.

ICDPPC-2018

- transparency should be promoted including: a) the development of innovative ways of communication, taking into account the different levels of transparency and information required for each relevant audience b)investing in public and private scientific research on explainable artificial intelligence c)making organizations practices more transparent (auditability of systems, algorithmic transparency) d)ensuring that individuals are always informed appropriately when they are interacting directly with an artificial intelligence system or when they provide personal data to be processed by such systems e)providing adequate information on the purpose and effects of artificial intelligence systems.

ICO-2017

- the complexity of big data analytics can mean that the processing is opaque to citizens and consumers whose data is being used and thus contribute to the lack of trust.
- However, Data protection Act introduced a specific transparency requirement, in the form of a ‘fair processing notice’, or more simply a privacy notice which should help organisations comply with transparency requirement.

DHSC-2019

- Individuals have the right to be informed about the collection and use of their personal data. This is a key transparency requirement under the Data Protection Act 2018.
- A privacy notice should identify who the data controller is, with contact details for its data protection officer. It should also explain the purposes for which personal data is collected and used, how the data is used and disclosed, how long it is kept and the controller's legal basis for processing.

TORONTO DECLARATION-2018

- States must secure maximum transparency. This can be done through
  a) Publicly disclose where machine learning systems are used and how are operating in the public sphere,
  b) Enable independent analysis and oversight by using systems that are auditable,
  c) Avoid using 'black box systems' that cannot be subjected to meaningful standards of accountability and transparency, and refrain from using these systems at all in high-risk contexts.

GOOGLE PERSPECTIVES

- generally, Google seeks to share the research to contribute to growing the wider AI ecosystem. However, we do not make it available without first reviewing the potential risks for abuse.
- Although each review is content specific, key factors that we consider in making this judgment include: a) Risk and scale of benefit vs downside b) Nature and uniqueness c) Mitigation options.

IMB-2018

- developers and designers must ensure an AI system's level of transparency is clear.
- Users should stay generally informed on the AI's intent even when they can't access a breakdown of the AI's process.

TELEFONICA-2018

- organisation must
  a) inform users of the data that we use and its purpose,
  b) take sufficient measures to ensure the understanding of the AI's decisions,
  c) tell our users when they are interacting with an AI system.

6.12 Responsibility

PUBLIC VOICE

- Institutions must be responsible for decisions made by an AI system: The identity of an AI system and the institution responsible must be made clear.
- Principle of Identification: The institution responsible for an AI system must be made known to the public.
- Individuals and not machines are responsible for automated decision-making, if a system fails a human assessment of the outcome must be made.

- Termination Obligation is the ultimate statement of accountability for an AI system. The obligation presumes that systems must remain within human control. If that is no longer possible, the system should be terminated

UNI GLOBAL

- AI must be responsible, safe and useful, where machines maintain the legal status of tools, and legal persons retain control over, and responsibility for, these machines at all times.
- Ban the Attribution of Responsibility to Robots: legal responsibility for a robot should be attributed to a person. Robots are not responsible parties under the law
- Design and operate AI systems to comply with existing law, including privacy.
- Workers should have the right to access, manage and control the data AI systems generate, given said systems' power to analyse and utilize that data.
- Workers must also have the 'right of explanation' when AI systems are used in human-resource procedures, such as recruitment, promotion or dismissal.

ALGO.RULES

- A natural or legal person must always be held responsible for the effects involved with the use of an algorithmic system.
- Accountability must be clearly assigned. The accountable person must be aware of the responsibilities associated with their tasks. This also applies to responsibilities that are shared by several people or organizations. The allocation of responsibility must be fully documented and transparent for internal and external parties. Responsibility may not be transferred to the algorithmic system itself, users or people who are affected by the system.
- The allocation of responsibility must be fully documented and transparent.
- If an algorithmic system results in a questionable decision or a decision that affects an individual's rights, it must be possible to request an explanation and file a complaint.

CERNA

Responsibility (both legal and moral aspects)

In order to be able to attribute liability in the event of a dysfunction, it must be possible to distinguish two agents: the system's designer, and it user.

- The designer is responsible in the case of poor design and the user in case of inadequate use.
- Researchers should develop and implement methods of monitoring, whether automatic or supervised by a human or another machine. Monitoring should apply to the data, to the operation of the machine, and to its chain of decision-making, with the goal of facilitating the attribution of responsibility for both normal and dysfunctional performance of the system.
- When documenting a machine learning system, researchers should give a sincere, honest, and complete description of any limits of which they are aware, pertaining to how much a decision or action by the system is attributable either to the source code or to the learning process.

EGE

Moral responsibility is an integral part of the conception of the person. Moral responsibility, in whatever sense, cannot be allocated or shifted to 'autonomous' technology.

- Meaningful Human Control is essential for moral responsibility.

- Autonomous' systems should only be developed and used in ways that serve the global social and environmental good.
- Align with a plurality of fundamental human values and rights.
- Risk awareness and a precautionary approach is needed.
- Avoid unacceptable risks.
- Do not compromise human freedom and autonomy by illegitimately and surreptitiously reducing options for and knowledge of citizens.
- Augment access to knowledge and access to opportunities for individuals.
- Authentic concern for research ethics among researchers is needed.
- Social accountability of developers and researchers is needed.
- Academic cooperation to protect fundamental rights and values is needed.

MONTREAL DECLARATION

- responsibility over decisions should be attributed to humans even if based on ai - recommendations.
- Furthermore, the decision to kill must always be made by human beings, and responsibility for this decision must not be transferred to an AIS. People who authorize AIS to commit a crime or an offence, or demonstrate negligence by allowing AIS to commit them, are responsible for this crime or offence.
- However, when damage or harm has been inflicted by an AIS, and the AIS is proven to be reliable and to have been used as intended, it is not reasonable to place blame on the people involved in its development or use.

GOOGLE PERSPECTIVES

- organizations should remain responsible for the decisions they make and the manner in which they act on them and sets reasons why it is not appropriate for moral or legal responsibility to be shifted to a machine a) It is unnecessary (There will always be a natural person or corporation liable within existing laws and legal frameworks), b) It is impractical (Even if it was possible to come up with a workable definition of robots or AI that warrant legal personhood (which is far from a given), it would be impossible to hold such entities accountable for violations of their obligations), c) It is immoral (Responsibility is an intrinsically human property), d) It is open to abuse (It would make it easier for bad actors to shield themselves from liability for illegal activities performed by machines they had created).

6.13 Accountability and assessment

ALGO.RULES

Assessment

- The objectives and expected impact of the use of an algorithmic system must be documented and assessed prior to implementation.
- Assessment: The effects of an algorithmic system must be reviewed on a regular basis. An algorithmic system must be subject to active monitoring in order to determine whether the targeted objectives are actually achieved, and the use of the system does not violate existing legislation.

Accountability

- If an algorithmic system results in a questionable decision or a decision that affects an individual's rights, it must be possible to request an explanation and file a complaint.

SIIA

Assessment

- A company has to document ethical evaluations and make them available in accordance with balancing risks and benefits (not possible without transparency)

Accountability

- Give weight to the considered judgments of people affected by data practices.
- Do not withhold information about data practices, when there are no compelling reasons to do so.

PUBLIC VOICE

- Assessment: An AI system should be deployed only after an adequate evaluation of its purpose and objectives, its benefits, as well as its risks.

EGE

Rule of law and accountability

- Rule of law, access to justice and the right to redress and a fair trial is needed.
- Protections against risks stemming from 'autonomous' systems that could infringe human rights, such as safety and privacy is needed.
- Fair and clear allocation of responsibilities and efficient mechanisms of binding law is needed.
- Clarify with whom liabilities lie for damages caused by undesired behaviour of 'autonomous' systems.
- Harm mitigating systems are needed.

OECD

Accountability

- AI actors should be accountable for the proper functioning of AI systems.

IEEE

- Legislatures/courts should clarify issues of responsibility, culpability, liability, and accountability for AI where possible during development and deployment (so that manufacturers and users understand their rights and obligations).
- Systems for registration and record-keeping should be created so that it is always possible to find out who is legally responsible for a particular AI. Manufacturers/operators/ owners of AI should register key, high-level parameters.

UNDG-2017

- appropriate governance and accountability mechanisms should be established to monitor compliance with relevant law, including privacy laws and the highest standards of confidentiality, moral and ethical conduct with regard to data use (including this Guidance Note).

AI HLEG

- mechanisms must be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use. Those include a)

Auditability b) Minimisation and reporting of negative impacts c)Trade-off and d) Redress. Furthermore, Organisations should set up governance frameworks, both internal and external, ensuring accountability for the ethical dimensions of decisions associated with the development, deployment and use of AI systems.

BARCELONA DECLARATION

- when an AI system makes a decision, humans affected by these decisions should be able to get an explanation why the decision is made in terms of language they can understand and they should be able to challenge the decision with reasoned arguments.

DATA ETHICS FRAMEWORK

- An important aspect of complying with data protection law, is being able to demonstrate what measures you are taking to ensure this (see Article 5(2) of the GDPR (the accountability principle) and Article 30 on keeping records of processing activities). Your organisation and information assurance teams will be responsible for this at a high level including ensuring policies and training are in place. However, it is essential to show how you are doing this at an individual level, through thorough documentation of things like Data Protection Impact Assessments.

MONTREAL DECLARATION

- the decisions made by AIS affecting a person's life, quality of life, or reputation should always be justifiable in a language that is understood by the people who use them or who are subjected to the consequences of their use.

ICDPPC-2018

- accountability should be encouraged by means of audit, continuous monitoring and impact assessment of artificial intelligence systems, and periodic review of oversight mechanisms, fostering collective joint responsibility and establishing demonstrable governance processes for all relevant actors.

DHSC-2019

- it must be explained to a member of the public why the data used was needed and how it is meeting the user need.
- The Data Ethics Framework, published by the Department for Digital, Culture, Media and Sport, sets out clear principles for how data should be used in the public sector.

ICO-2017

- under the GDPR its importance is elevated by introducing an explicit accountability principle that requires organisations to demonstrate compliance with all the other principles in the regulation.
- A further accountability provision under the GDPR is the requirement to appoint a data protection officer (DPO).
- Detecting discriminatory decisions in hindsight will not be sufficient to comply with the accountability provisions of the GDPR. Big data analysts will need to find ways to build discrimination detection into their machine learning systems to prevent such decisions being made in the first place. Except from GDPR requirements on accountability other elements of accountability are discussed including 'algorithmic accountability". This means being able to check

that the algorithms used and developed by machine learning systems are actually doing what we think they're doing and aren't producing discriminatory, erroneous or unjustified results. Finally, generally, Big data organisations need to exercise caution before relying on machine learning decisions that cannot be rationalised in human understandable terms.

TORONTO DECLARATION-2018

- States must be accountable. This can be done through the following steps
  a) Publicly disclose where machine learning systems are used and how are operating in the public sphere,
  b) Enable independent analysis and oversight by using systems that are auditable,
  c) Avoid using 'black box systems' that cannot be subjected to meaningful standards of accountability and transparency, and refrain from using these systems at all in high-risk contexts.

GOOGLE PERSPECTIVES

- reasonable explanation on why AIS behave in a certain way is a necessary element in dealing with AIS.
- However, a sensible compromise is needed that balances the benefits of using complex AI systems against the practical constraints that different standards of explainability would impose.
- It provides examples on what should be considered while giving reasonable explanation. This includes: a) who is asking and what do they seek? b) What does it relate to? c) When and where is it being delivered? D) How feasible is it to explain, technically and financially? Furthermore, article gives example of some of the hallmarks of a good explanation for lay users which among others include: Is the explanation clear, specific, relatable and actionable? And does the explanation take appropriate account of sensitivities? However, sometimes it is difficult to provide extensive information due to commercial or other restrictions, in such case, article provides alternative methods. Those include: a) Flagging facilities b) Adversarial testing c) Auditing d) Avenues for contesting an outcome.

IMB-2018

- every person involved in the creation of AI at any step is accountable for considering the system's impact in the world, as are the companies invested in its development.
- Developers must
  a) Make company policies clear and accessible
  b) Understand where the responsibility of the company/software ends
  c) Keep detailed records of your design processes and decision making,

d) Adhere to your company's business conduct guidelines.

- Also, understand national and international laws, regulations, and guidelines.
- Furthermore, decision making process should be explainable. This includes

a) A user should be able to ask why an AI is doing what it's doing on an ongoing basis,
b) Decision making processes must be reviewable,
c) When an AI is assisting users with making any highly sensitive decisions, the AI must be able to provide them with a sufficient explanation of recommendations, the data used, and the reasoning behind the recommendations,
d) Teams should have and maintain access to a record of an AI's decision processes and be amenable to verification of those decision processes.

TELEFONICA-2018

- telefonica ensures that ai always respects human rights and committed to the UN Sustainable Development Goals and contributes to preventing improper use of technology.

6.14 No bias, no discrimination (fairness)

AI4PEOPLE

- Develop auditing mechanisms for AI systems to identify unwanted consequences, such as unfair bias, and (for instance, in cooperation with the insurance sector) a solidarity mechanism to deal with severe risks in AI-intensive sectors.

UNI Global

No bias

- Control system for negative or harmful human-bias are needed.
- Make sure to identify bias and not to propagate it by the system.

PUBLIC VOICE

No bias

- Institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions. Normative consequences must be addressed as well.

SIIA

No bias and discrimination

- Avoid discrimination based on characteristics such as gender, race, genetics or age.
- Do not arbitrarily target groups based on attributes such as race, gender, religion, or ethnicity.
- Avoid data that cause substantial injury and cannot be avoided and have no countervailing benefits.
- Practice disparate impact assessment
    - Responsibility to take steps to determine whether their analytics systems have discriminatory effects
    - Establish and maintain effective and comprehensive data and model governance programs, including policies and procedures to assess the ethical implications of data analysis
    - Build governance, controls, and ethical assessment into the process of developing, revising, and updating models,
    - Make an assessment and documentation of all data assets.
    - Make an assessment of all analytical techniques.

CERNA

No bias and discrimination

- The designer and the trainer should pay attention to the training data and the conditions of data capture throughout the operation of the system. In order to check the absence of bias, they must rely on measurement tools that have yet to be developed.
- When selecting data, trainers of machine learning systems must ensure that those data reflect the diversity of the groups of users of those systems.

- The trainers (who may also be the designers or users) should pay attention to protected variables, e.g., variables that may permit social discrimination. These variables, such as ethnicity, sex or age, must not be used or be regenerated based on correlations. Personal data must also be protected as required by existing legislation.
- Researchers must ensure that no human bias is automatically expressed in a decision by learning systems in which human intervention is a part of the specification.

More on no bias

IPCO

- in order to prevent biased decision making based on processing of Big Data, collection of samples must accurately represent the targeting population and being properly randomised avoiding the results to be biased that is excluding certain members of the population.
- Furthermore, when analysing an integrated data set as part of a big data project, you should ensure that it is representative of the target population to the extent necessary to fulfil the purposes of your big data project.

CoE-2017

- in order to avoid bias controllers should adopt adequate by -design solutions.
- This would make it possible to assess the potential bias of the use of different parameters in analysing data and provide evidence to minimise the use of information and mitigate the potential negative outcomes identified in the risk-assessment.

UNDG-2017

- in order to avoid bias, all data must be of an utmost high quality.

AI HLEG

- AI must use inclusive data representing all groups in order to avoid biased outputs.

DATA ETHICS FRAMEWORK

- processor should be aware of the types of bias that can exist in the data you are using by reviewing how the data was collected including a) measurement bias (selection of data or samples in a way that does not represent the true parameters)
  b) social bias (information is based on historic decisions and actions by humans, or was shaped by laws no longer in force)
  c) social media (data should be properly investigated to identify any representation or selection bias)
  d)Practitioner bias (Data practitioners and others involved in a project may inadvertently introduce their own confirmation bias into the design of projects, analyses, or interpreting outputs)
  e)Survey methodology (Surveys must be carefully designed and used to ensure they cover your target population).

GAI

- refers to the Amnesty International initiative on AI and human rights which states that AI is built by humans and it will be shaped by human values.
- If we build AI systems that are a mirror to our current societies, they will be riddled with the historical biases and inequalities of our societies**.**

ICDPPC-2018

- investment in research into technical ways to identify, address and mitigate biases should be conducted, and elaboration on specific guidance and principles in addressing biases and discrimination should be done as well.

ICO-2017

- machine learning itself may contain hidden bias. A common phrase used in the discussion of machine learning is “garbage in garbage out”. Essentially, if the input data contains errors and inaccuracies, so will the output data. While supervised machine learning in particular often involves a pre-processing stage to improve the quality of the input data, the human-labelling of a training dataset can create a further opportunity for inaccuracies or bias to creep in.

DHSC-2019

- data users should be aware of potential biases in the data used for training algorithms – consider the representativeness of the database used for training the algorithm.
- If the data provided for the AI to learn is limited to certain demographic categories or disease areas, this could potentially limit the applicability of the AI in practice as its ability to accurately predict could be different in other ethnic groups.

TORONTO DECLARATION-2018

- the people behind the technology bring their own biases, and are likely to have limited input from diverse groups in terms of race, culture, gender, and socio-economic backgrounds.

GOOGLE PERSPECTIVES

- Google takes our responsibilities in this arena extremely seriously, not least in developing tools to tackle unfair bias. Those include a)Facets (two downloadable visualization tools to aid understanding and analysis of machine learning datasets) b)What If Tool (TensorFlow plugin offering an interactive visual interface for exploring model results, without the need for writing any further code) c)Model and Data Cards (accompany each released model and provide details of the model’s intended purpose, how it performs in tests) d) Training With Fairness Constraints (TensorFlow algorithms to train AI systems that satisfy standard desired statistical fairness goals).
- Furthermore, AI could be applied to analyze connections between input data and output predictions to surface any underlying biases that are embedded in existing processes. If these biases were determined to be unmerited, then decision-making practices could be tweaked in an effort to limit their effect.

IMB-2018 says that

a) Real-time analysis of AI brings to light both intentional and unintentional biases must be conducted,

b) Design and develop without intentional biases and schedule team reviews to avoid unintentional biases. Unintentional biases can include stereotyping, confirmation bias, and sunk cost bias.

c) Instil a feedback mechanism or open dialogue with users to raise awareness of user-identified biases or issues.

TELEFONICA-2018

- AI must be used in a fair manner, particularly it must be ensured that applications do not lead to biased results and unfair and discriminatory impacts and there are no discriminatory elements when AI learns and algorithms decide.

More on no discrimination

IPCO

- When analyzing an integrated data set as part of a big data project, you should be aware of the potential for variables to correlate with protected personal characteristics and ensure that your analysis does not result in any such variables being used as proxies for prohibited discrimination.

CoE-2017

- The use of Big Data should preserve the autonomy of human intervention in the decision-making process including a) decision based on Big Data analysis should take into account all circumstances and not be merely de-contextualised information, b) decisions affecting individuals upon their request should be justified by human decision maker and if reasonable arguments against relying on results of Big Data processing are forwarded, no decision should be taken, c) any indication of presumable discrimination, would oblige the controllers and processors to demonstrate lack of such discrimination, and d) any decision can be challenged before a competent authority.

UNDG-2017

- data quality must be assessed for biases to avoid any adverse effects, where practically possible, including giving rise to unlawful and arbitrary discrimination.

AI HLEG

- in an AI context, equality entails that the system's operations cannot generate unfairly biased outputs (e.g. the data used to train AI systems should be as inclusive as possible, representing different population groups). This also requires adequate respect for potentially vulnerable persons and groups, such as workers, women, persons with disabilities, ethnic minorities, children, consumers or others at risk of exclusion.

DATA ETHICS FRAMEWORK

- analysis or automated decision making must not result in outcomes that lead to discrimination as defined in the Equality Act 2010.

MONTREAL DECLARATION

- Ai should be developed in such way as not to promote discrimination and eliminate relationships of domination between groups of people based on their wealth, power and knowledge.

GAI

- ai systems are inherently discriminatory providing the example of reports of how discriminatory algorithms are already deployed in the justice system, wherein judges use these tools for sentencing that purport to predict the likelihood a criminal defendant will reoffend.
- Another example is the Allegheny Family Screening Tool (AFST), a predictive risk model deployed by the County Office of Children, Youth, and Families to forecast child abuse and neglect. The model has inherent flaws: it only contains information about families who use public services, making it more effective at targeting poor residents. In its turn such discriminatory effects can lead to harms in other human rights areas, such as education, housing, family, and work. Generally, without safeguards, even AI systems built for mundane bureaucratic functions can be repurposed to enact discriminatory policies of control.

ICDPPC-2018

- respect of international legal instruments on human rights and non-discrimination should be ensured.

TORONTO DECLARATION-2018

- all actors, public and private, must prevent and mitigate against discrimination risks in the design, development and application of machine learning technologies.
- They must also ensure that there are mechanisms allowing for access to effective remedy in place before deployment and throughout a systems lifecycle.
- Furthermore, states must not engage in, or support discriminatory or otherwise rights-violating actions or practices when designing or implementing machine learning systems in a public context or through public-private partnerships. To the contrary, States have positive obligations to protect against discrimination by private sector actors and promote equality and other rights, including through binding laws. States must conduct regular impact assessments throughout entire life circle of AIS and Taking appropriate measures to mitigate risks identified through impact assessments such as pre-release trials and subjecting systems to trials and audits.

TELEFONICA-2018

- it must be ensured that applications do not lead to biased results and unfair and discriminatory impacts and there are no discriminatory elements when AI learns and algorithms decide.

More on fairness

IPCO

- Data processing must be fair that is conducted by an institution with a legal authority and by approval of research ethics board (REB) or similar body.

ICDPPC-2018

- design development and use of AIS must be in accordance with fairness principle including a) ensuring that the use of artificial intelligence systems remains consistent with their original purposes b) considering collective impact on groups and on society at large c) are developed in a way that facilitates human development and does not obstruct or endanger it.

ICO-2017

- a key question for organisations using personal data for big data analytics is whether the processing is fair. Fairness involves several elements. Transparency – what information people have about the processing – is essential. But assessing fairness also involves looking at the effects of the processing on individuals, and their expectations as to how their data will be used. Effects of the processing may include profiling which perpetuates discrimination, for example on the basis of race and thus unacceptable.
- As for the expectations, it is necessary to consider whether a particular use of personal data be within the reasonable expectations of the people concerned.
- Furthermore, data processing must be legitimate and necessary for the purpose that is more than just potentially interesting.

GOOGLE PERSPECTIVES

- Unfair stereotypes and negative associations embedded in algorithmic systems (deliberately or accidentally) can cause or amplify serious and lasting harm. However, even for situations that seem simple, people can disagree about what is fair, and it may be unclear what optimal approach should dictate policy, especially in a global setting.
- However, if well implemented, an algorithmic approach can help to boost the consistency of decision-making, especially compared to the alternative of individuals judging according to their own internal (and thus likely varying) definitions of fairness.

IMB-2018 says that

a) Real-time analysis of AI brings to light both intentional and unintentional biases must be conducted,

b) Design and develop without intentional biases and schedule team reviews to avoid unintentional biases. Unintentional biases can include stereotyping, confirmation bias, and sunk cost bias.

c) Instil a feedback mechanism or open dialogue with users to raise awareness of user-identified biases or issues.

6.15 Diversity

MONTREAL DECLARATION

- development of Ai must be consistent with maintaining social and cultural diversity including:

a) AIS development and use must not lead to the homogenization of society through the standardization of behaviours and opinions,

b) AIS development and deployment must take into consideration the multitude of expressions of social and cultural diversity present in the society,

c) AI development environments, whether in research or industry, must be inclusive and reflect the diversity of the individuals and groups of the society,

d) AIS must avoid using acquired data to lock individuals into a user profile, fix their personal identity, or confine them to a filtering bubble,

e) AIS must not be developed or used with the aim of limiting the free expression,

f) for each service category, the AIS offering must be diversified to prevent de facto monopolies from forming and undermining individual freedoms.

TORONTO DECLARATION-2018

- this declaration underlines that inclusion, diversity and equity are key components of protecting and upholding the right to equality and non-discrimination. All must be considered in the development and deployment of machine learning systems in order to prevent discrimination,

particularly against marginalised groups. Inclusion, diversity and equity entails the active participation of, and meaningful consultation with, a diverse community, including end users, during the design and application of machine learning systems, to help ensure that systems are created and used in ways that respect rights – particularly the rights of marginalised groups who are vulnerable to discrimination.

IMB-2018

- developers should consider hiring diverse teams to help represent a wider variation of experiences to minimize bias.
- Embrace team members of different ages, ethnicities, genders, educational disciplines, and cultural perspectives.

6.16 Safety/Security

PUBLIC VOICE

- Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices, and implement safety controls.
- Institutions must secure AI systems against cybersecurity threats.
- The reliability and robustness of an algorithmic system as well as its underlying data with respect to attacks, access and manipulation must be guaranteed.
- Security must be built into the architecture of the algorithmic system (security by design). The system must be tested in a protected environment prior to implementation.
- Security precautions must be documented.

ALGO.RULES

In the sense of manageability and control

- An algorithmic system must be manageable throughout the lifetime of its use. Control must always be maintained.
- The security of an algorithmic system must be tested before and during its implementation.
- Safety is guaranteed by competency: Those who develop, operate and/or make decisions regarding the use of algorithmic systems must have the necessary expertise and appropriate-to-scale understanding of how the technology functions and its potential effects.

EGE

Security is to be understood in three way: safety, bodily and mental integrity: 1. external safety for their environment and users, 2. reliability and internal robustness, e.g. against hacking, 3. emotional safety with respect to human-machine interaction

- Ensure that 'autonomous' systems do not infringe on the human right to bodily and mental integrity and a safe and secure environment.

PUBLIC VOICE

Accuracy, Reliability and Validity

- High quality of data must be ensured.

OECD

Robustness, security and safety

- AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.
- AI actors should ensure traceability to guarantee safety.
- AI actors should implement mechanisms and safeguards, such as capacity for determination.

IEEE

Safety (and beneficence)

- AI must be verifiably safe and secure throughout their operational lifetime.
- To best honour human rights, society must assure the safety and security of AI so that they are designed and operated in a way that benefits humans.

Technical

- AI techniques should be robust.
- It is important to contribute to research on concrete problems in AI safety.
- Work to ensure that AI are transparent, i.e., that their internal reasoning processes can be understood by human operators.
- Work to build safe and secure infrastructure and environments for development, testing, and deployment of powerful AI.
- Work to ensure that AI "fail gracefully" (e.g., shutdown safely or go into some other known-safe mode) in the face of adversarial inputs.
- Teams working on developing AGI systems should be prepared to put significantly more effort into AI safety research as capabilities grow.
- Teams working on developing AGI systems should cultivate a "safety mindset" like a "security mindset," vigilant of ways they can cause harm and invest in preventing those.
- When designing an advanced AI, researchers and developers should pay the upfront costs to ensure, to the extent possible, that their systems are "safe-by-design.

General

- Organizations working on sufficiently advanced AI should set up review boards to consider the implications of risk-bearing proposed experiments and development.
- Technologists should work to minimize the extent to which beneficial outcomes from the system hinge on the virtuousness of the operators.
- Adopt the stance that superintelligence should be developed only for the benefit of all of humanity.
- De-stigmatize and remove other soft and hard barriers to AI researchers and developers working on safety, ethics, and beneficence, as well as being open regarding that work.

6.17 Trustworthiness

AI HLEG

- trustworthiness is a prerequisite for people and societies to develop, deploy and use AI systems. Trustworthy AI must be a) legal b) ethical c) robust

DATA ETHICS FRAMEWORK

- in order to make best use of data we must ensure we have robust and consistent practices. This involves

a) working in multidisciplinary teams
b) getting help from experts outside your team
c) ensuring accountability of algorithms
d) avoiding outputs of analysis which could result in unfair decision making
e) designing for reproducibility
f) testing your model under a range of conditions
g) defining acceptable model performance: false negatives and false positives.

ICO-2017

- Senior managers in big data organisations need to know whether they can trust what the data is apparently telling them. This can involve looking at, for example, the sources of the data, how accurate it is, whether is sufficiently up to date, how securely it is kept and whether there are restrictions on how it can be used.

6.18 Data quality

IPCO

- personal information should be accurate, complete and kept up-to-date to the extent necessary to fulfill the purposes of its use. However, the level of accuracy it requires is not absolute, but rather depends on the proposed use of the information.

CoE-2017

- to minimise the presence of redundant or marginal data, avoid potential hidden data biases and the risk of discrimination or negative impact on the rights and fundamental freedoms of data subjects, in both the collection and analysis stages controllers must adopt adequate (must be tested through simulations) by-design solutions.

UNDG-2017

- data-related activities should be designed, carried out, reported and documented with an **adequate level of quality** and transparency subject to a periodic assessment.
- Automatic processing of data, including the use of algorithms, without human intervention and domain expertise should be avoided.

AI HLEG

- when data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given data set.
- In addition, the integrity of the data must be ensured. Feeding malicious data into an AI system may change its behaviour, particularly with self-learning systems.

- Processes and data sets used must be tested and documented at each step such as planning, training, testing and deployment. This should also apply to AI systems that were not developed in-house but acquired elsewhere.

DATA ETHICS FRAMEWORK

- when designing a new use of data, you must understand the impact of data provenance on accuracy, reliability and representativeness. If there are errors in the way data is collected or interpreted, they should be reported to policy or operational staff.
- The UK Statistics Authority Quality Assurance of Administrative Data framework provides useful resources to help understand the data that you are using, how it was collected and any likely quality impacts. Furthermore, improvement of documentation of the metadata, should be conducted if possible.

ICDPPC-2018

- reasonable steps must be taken by developers in order to ensure the personal data and information used in automated decision making is accurate, up-to-date and as complete as possible.

ICO-2017

- the fourth principle of the DPA requires that personal data is accurate and, where necessary, kept up to date. This is obviously good practice in terms of information management but it is also linked to the rights of the individual: people have a right to have inaccurate data corrected.

DHSC-2019

- data must be accurate and complete, this must be achieved through 2 stage approach which includes: a) Algorithms should be trained to understand the levels of data quality first b) then achieve their objective by using the variables given.
- This 2-stage approach should be built in so that high fluxes in data quality are handled appropriately.
- Furthermore, assessment of data quality should not be a one-off check – continuous anomaly detection should be in place to provide alerts to changes in a data source. Good data linkage will avoid reducing data quality.

6.19 Protection of citizens' rights

AI HLEG

- AI systems offer substantial potential to improve the scale and efficiency of government in the provision of public goods and services to society. At the same time, citizens' rights could also be negatively impacted by AI systems and should be.
- Safeguarded what is needed is to identify or measure a risk (e.g., on democracy, the rule of law and distributive justice, or on the human mind itself.) Adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk.

MONTREAL DECLARATION

- for public AIS that have a significant impact on the life of citizens, citizens should have the opportunity and skills to deliberate on the social parameters of these AIS, their objectives, and the limits of their use.

ICDPPC-2018

- Empowerment of every individual should be promoted, and the exercise of individuals' rights should be encouraged, as well as the creation of opportunities for public engagement including individuals' right not to be subject to a decision based solely on automated processing if it significantly affects them and, where not applicable, guaranteeing individuals' right to challenge such decision.

ICO-2017

- People have the right to be told what personal data about them is being processed, the purposes for which it is being processed and who it may be disclosed to. In order to make compliance with this obligation easier it is suggested that if an organisation's move to big data means that disparate data stores are brought together, this may make it easier to find all the data on an individual.
- Furthermore, citizens have other rights such as the right of prevention of processing likely to cause damage or distress; the prevention of direct marketing; the right not to be subject to purely automated decision making; and the rectification of inaccurate data. GDPR further extends the rights in regards profiling where people have the right not to be subject to a decision based solely on automated processing, including profiling, if it "significantly affects" them, such as automated decisions made on online credit applications or e-recruitment.

6.20 Protection of human rights

IEEE

- AI should be designed and operated in a way that both respects and fulfils human rights, freedoms, human dignity, and cultural diversity.
- For the foreseeable future, AI should not be granted rights and privileges equal to human rights: AI should always be subordinate to human judgment and control.
- Governance frameworks, including standards and regulatory bodies, should be established to oversee processes assuring that the use of AI does not infringe upon human rights, freedoms, dignity, and privacy, and of traceability to contribute to the building of public trust in AI.
- A way to translate existing and forthcoming legal obligations into informed policy and technical considerations is needed. Such a method should allow for differing cultural norms as well as legal and regulatory frameworks.

CoE-2017

- controllers and processors should adequately take into account the likely impact of the intended Big Data processing and its broader ethical and social implications to safeguard human right and fundamental freedoms, and ensure the respect for compliance with data protection obligations as set forth by Convention 108.

- Even though ethical values may differ the common guiding ethical values can be found in international charters of human rights and fundamental freedoms, such as the European Convention on Human Rights.
- If there is a possible impact an ad-hoc ethics committee must be established or rely on the existing ones in order to identify the particular value and adequately safeguard it.

AI HLEG

- trustworthy ai must be most and foremost ethical which in its turn based on the fundamental rights enshrined in the EU Treaties, the EU Charter and international human rights law. Respect for fundamental rights, within a framework of democracy and the rule of law, provides the most promising foundations for identifying abstract ethical principles and values, which can be operationalised in the context of AI.
- The 4 imperative ethical principles appearing in the article are respect for human autonomy, prevention of harm, fairness and explicability.
- Furthermore, any potential tension between them must be acknowledged and addressed.

6.21 Human Dignity

EGE

Human dignity: recognition of the inherent human state of being worthy of respect. It comprises the capacity of human persons to legislate for themselves, to formulate, think and choose norms, rules and laws for themselves to follow. It translates into responsibility and control over and knowledge about AI.

- There must be limits to determinations and classifications concerning persons, made on the basis of algorithms and 'autonomous' systems.
- There must be limits to the ways in which people can be led to believe that they are dealing with human beings while in fact they are dealing with algorithms and smart machines.

AI HLEG

- AI systems should hence be developed in a manner that respects, serves and protects humans' physical and mental integrity, personal and cultural sense of identity, and satisfaction of their essential needs.

GOOGLE PERSPECTIVES

- there have been arguments that allowing certain kinds of life-determining medical decisions to be made solely by machines may fail to respect the right to human dignity.

6.22 Democracy

EGE

Democracy (the human right to self-determination through the means of democracy. Of key importance to our democratic political systems are value pluralism and diversity.

- Key decisions on the regulation of AI development and application should be the result of democratic debate and public engagement.
- Everyone needs understand risks and opportunities (therefore right to receive education or access information on new technologies and their ethical implications needed)

- Democratic values need not be jeopardised, subverted or equalised by new technologies that inhibit or influence political decision making an
- Right to receive and impart information without interference.

GAI

- bots have manipulated the political process through social media and affected the democratic process. An example is given of Russian government operatives affecting US elections in 2016. The rights around political participation are referenced, for example, in the right to self-determination and the right to equal participation in political and public affairs in the ICCPR. Therefore the co-opted use of automated system by a bad faith actor creates human rights liability that demands redress.

6.23 Data protection and privacy

EGE

- Do not collect and spread data or be run on sets of data for whose use and dissemination no informed consent has been given.
- There is the right to be free from technologies that influence personal development and opinions.
- There is the right to establish and develop relationships with other human beings and the right to meaningful human contact (not only with machines).
- There is the right to be free from surveillance.
- There is the right to not be profiled, measured, analysed, coached or nudged.

IEEE

Personal data and access control

- People have the right to define access and provide informed consent with respect to the use of their personal data.

Digital persona

- The ethics of creating secret and proprietary AI from people's personally identifiable information (PII) need to be considered.
- Individuals should have access to trusted identity verification services to validate, prove, and support the context-specific use of their identity.

Agency and Control

- Individuals should have access to means that allow them to exercise control over use of personal data at the time the data is used.
- Personal data access and consent should be managed by the individual using systems that provide notification and an opportunity for consent at the time the data is used, versus outside actors being able to access personal data outside of an individual's awareness or control.

Transparency and access

- Service providers should ensure that personal data management tools are easy to find and use within their service interface.

- A system to assess privacy impacts related to AI needs to be developed, along with best practice recommendations, especially as automated decision systems spread into industries that are not traditionally data-rich.

Privacy and safety (procedural norms)

- Frameworks such as Privacy by Design can guide the process of identifying appropriate system and software requirements in early stages of design.
- Best practices such as Privacy Impact Assessments will assist with identification of data misuse cases at early stages of system/ software design.
- Improving digital literacy of citizens should be a high priority for the government and other organizations.
- Governments should enforce transparency related to data collection, data ownership, data stewardship, and data usage and disclosure.
- Organizations should be held accountable for data misuse, financial loss, and harm to the reputation of the data object if data is mishandled. This requires that organizations have appropriate policies and agreements in place, that terms and conditions of the agreements are clearly communicated with the data object and that data misuse cases and legitimate use cases are well-defined in advance.

More on Privacy

ADMA

- in order to preserve privacy and avoid extensive governmental control, organisations processing Big Data has to be a) transparent b) put customer first and c) consider the brand not just the law.

IPCO

- privacy element must be accessed by a research ethics board or other relevant body which will take into account among others whether the potential benefits to be derived from the project outweigh the foreseeable risks to the individuals whose personal information is being collected and whether the personal information that is to be collected is reasonably limited. Furthermore, personal information in policy analysis cannot be re-used for administrative purposes since it would compromise the integrity of the project and erode public trust in the process. Therefore de-identification of personal information must be used.
- To be successful, a big data project need only retain personal information for the duration of the project.

CoE-2017

- in order to secure the highest privacy for personal data, anonymisation procedures should be adopted involving Technical as well as with legal or contractual obligations to prevent possible reidentification of the persons concerned.
- Furthermore, controllers should constantly review the risk of re-assessment.

UNDG-2017

- in order to provide a higher standard of privacy protection before proceeding to the Big Data processing risk assessment should be conducted.

AI HLEG

- to allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them.

DATA ETHICS FRAMEWORK

- data processing must comply with data protection rules of GDPR and Data Protection Act 2018 (DPA 2018).
- Furthermore, before proceeding to Data processing personal data must be de-identified (anonymised) to the greatest degree possible. If data is anonymised to the greatest degree possible, it is likely to be out of scope of data protection law as it is no longer considered personal data. However if data is Pseudonymised, is subject to the same laws as fully identifiable personal data. Recital 26 of the GDPR provides further information.

MONTREAL DECLARATION

- AIS should not be intruding into privacy and intimacy affairs of humans. This includes protection of personal spaces a) from the intrusion of AIS and data acquisition and archiving systems (DAAS). b) The intimacy of thoughts and emotions must be strictly protected from AIS and DAAS c) right to digital disconnection in their private lives d) extensive control over information regarding their preferences e) DAAS must guarantee data confidentiality f) exercise extensive control over their personal data g) Individuals should be free to donate their personal data to research organizations h) The integrity of one's personal identity must be guaranteed.

GAI

- beyond various methodological shortcomings, the research demonstrated how a disregard for privacy rights increases the risks of algorithmic surveillance, where data that is collected and analyzed threatens to reveal personal information about users. This can put individuals and groups at risk, particularly those living under regimes that would use such information to repress and discriminate.
- Therefore, if developers treat privacy as a human right and not an ethical preference it would strengthen privacy considerations. Furthermore in the context of business it provides an example of letter of complaint from a number of non-profits and investment groups to Facebook largest institutional shareholders after political firm Cambridge Analytica surreptitiously gained access to the private data of tens of millions of Facebook users to influence their voting behaviour.

ICDPPC-2018

- privacy by default and privacy by design principles must be applied in development of AI including: a) implementing technical and organizational measures and procedures – proportional to the type of system that is developed – to ensure that data subjects' privacy and personal data are respected, both when determining the means of the processing and at the moment of data processing, B) assessing and documenting the expected impacts on individuals and society at the beginning of an artificial intelligence project and for relevant developments during its entire life cycle, and C) identifying specific requirements for ethical and fair use of the systems and for respecting human rights as part of the development and operations of any artificial intelligence system.

ICO-2017

- if personal data can be fully anonymised, it is no longer personal data and thus does not fall under data protection regulation. However, it may not be possible to establish with absolute certainty that an individual cannot be identified from a particular dataset, taken together with other data that may exist elsewhere. The issue is not about eliminating the risk of re-identification altogether, but whether it can be mitigated so it is no longer significant.
- Organisations using anonymised data need to be able to show they have robustly assessed the risk of re-identification, and have adopted solutions proportionate to the risk. This may involve a range of technical measures, such as data masking, pseudonymisation and aggregation, as well as legal and organisational safeguards.
- Furthermore, to establish whether the processing is fair, it is particularly important to assess, before processing begins, to what extent it is likely to affect the individuals whose data is being used and to identify possible mitigation measures. The tool for such an analysis is a privacy impact assessment (PIA).
- It is also advised to implement privacy by design approach, which includes anonymisation, pseudonymisation, as well as 'differential privacy' adopted by Apple and Google.
- Moreover, privacy by design solutions also includes access controls, audit logs and encryption.
- Finally, it has been suggested that one way to increase an individual's control over the use of their data is through what are usually called personal data stores, or sometimes personal information management services. These are third-party services that hold peoples' personal data on their behalf and make it available to organisations as and when the individuals wish to do so.

DHSC-2019

- a core element of at-scale adoption and uptake is to ensure that security and data protection methodology have been incorporated. NHS Digital has launched a new Data Security and Protection Toolkit to replace the previous Information Governance Toolkit to ensure that patient information is kept safe.

IMB-2018

- it must be allowed by users to deny service or data by having the AI ask for permission before an interaction or providing the option during an interaction.
- Privacy settings and permissions should be clear, findable, and adjustable.

TELEFONICA-2018

- in order to ensure utmost protection for privacy rights telefonica uses privacy by design methodology which includes technical and organisational protection measures.

More on data protection

CoE-2017

- the fundamental right to protection of personal data, must be secured through sufficient degree of control from the data subject which includes awareness and real freedom of choice.

GAI

- The European Union has demonstrated an interest in regulating technology companies with an appeal to rights-based principles. The EU's General Data Protection Regulation (GDPR) establishes new protections for European citizens' rights around data protection and privacy, which impacts any organization collecting European residents' data.

More on responsible use of data

ADMA

- organisations processing Big Data should process data in a responsible manner including imposition of adequate security measures including technical and organisational measures as well as maintaining high professional standards among the employees. Moreover a careful selection of third parties to whom data might be passed must be conducted (due diligence).

IPCO

- protective measures must be used including personal information in policy analysis cannot be re-used for administrative purposes since it would compromise the integrity of the project and erode public trust in the process. Therefore de-identification of personal information must be used.
- Finally to be successful, a big data project need only retain personal information for the duration of the project.

CoE-2017

- controllers should implement preventive policies and risk assessment including examination of the likely impact of the intended data processing on the rights and fundamental freedoms of data subjects especially in relation to sensitive data, in order to avoid potential negative outcome such as infringement of right to non-discrimination.
- Furthermore Pseudonymisation measures must be adopted.
- Finally, public and private entities should carefully consider their open data policies concerning personal data since open data might be used to extract inferences about individuals and groups.

UNDG-2017

- a risk assessment should be conducted before dealing with Big Data processing. Furthermore, a higher degree of care should be applied in relation to sensitive data.
- Furthermore, Data security must be achieved during all Big Data processing including technical and organisational safeguards, de-identification procedures (pseudonymization),
- However, Data security measures should be assessed in light of the risks, harms and benefits of data use and would not compromise the utility of the data for the intended purpose.
- Moreover, when dealing with third-party collaborators, it is advisable to conduct due diligence process and entering in legally binding agreements outlining parameters for data access and handling.

AI HLEG

- in any given organisation that handles individuals' data (whether someone is a user of the system or not), data protocols governing data access should be put in place. These protocols should outline who can access data and under which circumstances. Only duly qualified personnel with the competence and need to access individual's data should be allowed to do so.

DATA ETHICS FRAMEWORK

- when accessing or sharing personal data, you must follow the Information Commissioner's Code of Practice for Data Sharing which should be read alongside the ICO's Guide to GDPR. This Code of Practice is due to be updated to align with the new Data Protection Act 2018. When accessing and sharing data under powers in Part 5 of the Digital Economy Act 2017, you must follow the relevant Codes of Practice.
- When re-using published and unpublished information relating to public tasks, you must follow the re-use of Public Sector Information Regulations 2015. Furthermore, data protection by design and by default is a legal requirement under the GDPR. It means taking a holistic approach to embed data protection from design through to application of any use of personal data. Furthermore, organisations have a responsibility to keep both personal data and non-personal data secure, complying with GDPR and carriage of risk assessment procedures. It is also necessary to follow guidelines of The Government Social Research (GSR) in regards social media data. Finally, appropriate long-term processes must be put in place to monitor policies developed using data analysis.

ICO-2017

- data processing must have a purpose limitation which involves a two stage test.

A) first, the purpose for which the data is collected must be specified and lawful (the GDPR adds 'explicit) b)second, if the data is further processed for any other purpose, it must not be incompatible with the original purpose.

A key factor in deciding whether a new purpose is incompatible with the original purpose is whether it is fair. In particular, this means considering how the new purpose affects the privacy of the individuals concerned and whether it is within their reasonable expectations that their data could be used in this way.

- Furthermore, the guideline talks about security issues and that in addition to the security issues associated with any IT system, specific security threats can arise from the nature of big data processing. These need to be addressed by data controllers as part of their risk assessment in order to meet the requirement, in both the DPA and the GDPR, to put in place appropriate security measures to protect personal data.

DHSC-2019

- proportionality of the data must be explained. This applies to both data use in the research or testing period and after the digital health product goes live and is used as part of standard care. Wherever possible it is preferable to use anonymised data in testing rather than identifiable patient data. Furthermore, risk assessment strategies must be applied.
- Finally, if conducting research an approval from the Health Research Authority (HRA) to carry out this research must be secured.
- However, if the patient data planning to be used has been anonymised in line with the ICO's code of conduct on anonymisation and meets the requirements of the common law duty of confidentiality, ethical review from the HRA will not be needed. However, HRA approval may still be required.
- A data protection by design and by default' must be adopted. From a practical perspective, the important documents underpinning this are data flow maps, data protection impact assessments (DPIA) and privacy notices.

- Data sharing agreements or contracts between data controllers and are strongly recommended.

TORONTO DECLARATION-2018

- states must secure effective oversight over public and private bodies dealing with AIS, including a) Ensure that machine learning-supported decisions meet international accepted standards for due process, b) Create mechanisms for independent oversight, including by judicial authorities c) Ensure that public bodies carry out training in human rights and data analysis for officials involved d)Proactively adopt diverse hiring practices and engage in consultations to assure diverse perspectives so that those involved in the design, implementation, and review of machine learning represent a range of backgrounds and identities.
- When dealing with private sector any state authority procuring machine learning technologies from the private sector should maintain relevant oversight and control over the use of the system, and require the third party to carry out human rights due diligence to identify, prevent and mitigate against discrimination and other human rights harms, and publicly account for their efforts in this regard. As for the private sector, human rights due diligence procedure must be adopted in order to avoid human rights violations. This process involves a) Identification of potential discriminatory outcomes B) Take effective action to prevent and mitigate discrimination and track Responses c) Be transparent about efforts to identify, prevent and mitigate against discrimination in machine learning systems.

GOOGLE PERSPECTIVES

- It's hard to build systems that provide both the necessary restrictions for security, as well as the necessary flexibility to generate creative solutions or adapt to unusual inputs.
- However, this article provides us with Google's approach to automating the control of data center cooling. This includes: a) Continuous monitoring b) Automatic failover (to a neutral state if the AI control system does violate the safety constraints) c)Smoother transfer (during fallovers to prevent sudden changes to the system.) d)Two-layer verification (of the AI actions before implementation.) e) Constant communication (between the cloud-based AI and the physical infrastructure.)f) Uncertainty estimation (to ensure we only implement high confidence actions) g) Rules and heuristics (as backup if we need to exit AI control mode) h) Human override (is always available to override AI actions as necessary).

IMB-2018

- Users' data should be protected.
- This includes a) Forbid use of another company's data without permission when creating a new AI service b) Recognize and adhere to applicable national and international rights laws when designing for an AI's acceptable user data access permissions c)Employ security practices including encryption, access control methodologies, and proprietary consent management modules.

TELEFONICA-2018

- telefonica confirms the veracity of the logic and the data used by the suppliers.

6.24 Personal Data Minimisation

IPCO

- Collection of personal data for Big data projects must be limited to what is directly relevant and necessary to achieving a specified purpose.

CoE-2017

- Personal data shall be processed for specified and legitimate purposes and not used in a way incompatible with those purposes.
- Personal data should not be further processed in a way that the data subject might consider unexpected, inappropriate or otherwise objectionable.
- Exposing data subjects to different risks or greater risks than those contemplated by the initial purposes could be considered as a case of further processing of data in an unexpected manner.

UNDG-2017

- any personal information must be compatible or otherwise relevant, and not excessive in relation to the purposes for which it was obtained.
- The purpose of data use must be precisely defined and articulated before the data use, and cannot be changed unless there is a legitimate basis.

DATA ETHICS FRAMEWORK

- You must use the minimum data necessary to achieve your desired outcome in accordance with article 5(1) (c) of the GDPR. However, when deciding if a particular data use is ethical, we need to think beyond legal compliance only. Use data that is proportionate to the user need.

ICDPPC-2018

- AIS developers and designers must consider reasonable expectations by ensuring that the use of artificial intelligence systems remains consistent with their original purposes, and that the data are used in a way that is not incompatible with the original purpose of their collection.

ICO-2017

- Organisations should minimise the amount of data they collect and process, and the length of time they keep the data. Organisations therefore need to be able a) to articulate at the outset why they need to collect and process particular datasets, and b) adopt good information governance and in particular enforce appropriate retention schedules.

6.25 Ethics

Explanation: Even though all the principles are about ethics, in this analysis we put 'Ethics' as a stand-alone principle because there are guidelines which refer to ethics in a way which does not fit into any other Principle either because they are general or for other reasons.

IPCO

- it must be approved by a research ethics board or other relevant body which will take into account among others whether the potential benefits to be derived from the project outweigh the foreseeable risks to the individuals whose personal information is being collected and whether the personal information that is to be collected is reasonably limited,

- When using profiling as part of a big data project, you should consult with the public and civil society organizations regarding the appropriateness and impact of the proposed use of profiling and provide them with an opportunity to comment on the effects the profiling may have on society and individuals' lives.

UNDG-2017

- before proceeding to the processing Big Data, a risk of harms assessment should be conducted by an independent team of experts with a representative of a group affected.
- Assessment should be based on proportionality principles where risks would not outweigh positive impacts. Key factors of assessments being: (i) the likelihood of occurrence of harms, (ii) potential magnitude of harms and (iii) potential severity of harms.
- Consideration shall be given to multiple factors (social, religious, geographical e.t.c) and any impact shall taken into account whether legally visible or not and whether known or unknown at the time of data use.

DATA ETHICS FRAMEWORK

- data project, policy, service or procured software should be assessed against the 7 data ethics principles a) Start with clear user need and public benefit b) Be aware of relevant legislation and codes of practices c) Use data that is proportionate to the user need d) Understand the limitations of the data e) Use robust practices and work within your skillset f) Make your work transparent and be accountable g) Embed data use responsibly.

ICDPPC-2018 promotes a general "ethics by design" which includes responsible design and development of AIS.

ICO-2017

- several commentators who are concerned about the privacy impact of big data have advocated the need for an ethical approach that supports and goes beyond compliance with legal requirements. There is a role in this for councils or boards of ethics, both within organisations and at a national level.
- In addition to ethical frameworks, there is also a role for developing common standards for big data analytics. Organisations such as the BSI, the International Telecommunications Union, and the International organisation for Standardization have been working towards a set of big data standards to help establish best practice and reduce risk for organisations involved in big data processing.

DHSC-2019

- Ethical examination of data use specific to this use-case must be undertaken. This can be done through a) Achieving transparency of algorithms that have a higher potential for harm or unintended decision-making, b) Work collaboratively with partners, specify the context for the algorithm, specify potential alternative contexts and be transparent on whether the model is based on active, supervised or unsupervised learning. Show in a clear and transparent specification e.g the functionality of the algorithm.

GOOGLE PERSPECTIVES

- for ethical reasons we would suggest that people should always be meaningfully involved in making legal judgments of criminality, or in making certain life-altering decisions about medical treatment.
- It would also be useful to have broad guidance as to what human involvement should look like — for example, an evaluation of common approaches to enabling human input and control, with commentary on which are acceptable or optimal, supplemented by hypothetical examples from different contexts. Article provides some initial thought on it.

IMB-2018

- today's AI systems do not have these types of experiences to draw upon, so it is the job of designers and developers to collaborate with each other in order to ensure consideration of existing values.
- Care is required to ensure sensitivity to a wide range of cultural norms and values.
- Regard should be given to a) Consider the culture that establishes the value systems you're designing within, b) Work with design researchers to understand and reflect your users' values, c) Consider mapping out your understanding of your users' values and aligning the AI's actions accordingly with an Ethics Canvas (https://www.ethicscanvas.org/).

7 Analysis summation

Below a summation of the analysis of the guidelines, divided in two sections (one to nine and ten and onwards).

7.1 Summation of analysis of the guidelines one to nine

Focus: it is striking that most of the guidelines focus on AI; in the analysed guidelines there was none on big data in particular. Two guidelines (CERNA and UniGlobal) focus on machine learning. One guideline (Algo.rules) concentrates on algorithmic systems.

Content: most of the guidelines are made for the use and the design of AI and are thus general in nature. Only one guidelines (CERNA) focuses on the design of AI. The guideline from IEEE has a slight emphasis on the design as well but is overall general in nature. One (SIIA) mentions that the focus is on the responsible use of data but includes in the target group organizations developing and using data and models as well.

Intended users of guidelines: most guidelines appeal to a broad target group and are not specified for a particular class of users. Only the guideline developed by CERNA is addressed to IT researchers, developers, and designers specifically.

Level of detail: the guidelines analysed vary widely in length and detail. Some are very brief (with only a maximum of 6 pages), some are more detailed and the one from IEEE comprises 246 pages and is very elaborate covering many topics in detail and making tailored suggestion to specific problems.

Specific product or field of application: most of the guidelines do not focus on a particular field of application or product. The guideline from SIIA has a focus on data practices with issues of data and analytical models that might have a disparate impact as a main topic.

Other criteria: only some of the guidelines make suggestions for what is needed in terms of implementation and further instruments. Some guidelines suggest multi-stakeholder approaches, some emphasise the necessity of an international or even global approach. Two guidelines demand independent ethics committees and review boards. On top of that an EU oversight agency or EU observatory is claimed for, public consultation mechanisms as well as a public oversight system. On top of that a code of ethics and conduct is mentioned as desirable.

Teaching is explicitly mentioned in three guidelines: AI4people suggests “AI literacy programs” for consumers and the creation of educational curricula and public awareness activities at schools, businesses, universities and in management. EGE suggests to facilitate training in STEM and digital subjects whereas IEEE wants the ethical aspects to be taught from school age to university.

Values: Grid 3 shows all the values were mentioned in the guidelines. They are clustered here according to similarities in content (other ways to cluster them are of course possible). They are:

- Beneficence, sustainability, non-maleficence.
- Justice (distribution of resources and social Justice), equity, equality, fairness, no bias, no-discrimination, diversity.
- Explicability, intelligibility, transparency, predictability, responsibility, traceability, accountability, security, safety, robustness, accuracy of data, reliability, validity.

- Human determination of systems, competency, identification principle, due diligence, trustworthiness (the systems behave as their designers claim), compliance (to specification of system and law).
- Data protection, privacy, individual access control, informed consent.
- Protection of human rights, protection of individual rights, democracy, autonomy, freedom, human dignity, human centred values, participation, solidarity, meaningful human contact.
- Awareness of dual use.

The values that were not only mentioned but elaborated further in the guidelines were then analysed due to the question in which recommendations they translate. These values are: beneficence, sustainability, societal and environmental wellbeing, non-maleficence/doing no harm, autonomy, human centric approach, consent (and symmetry), justice, explicability/explainability, transparency, responsibility, accountability (and assessment), no bias, no discrimination, fairness, diversity, safety/security, trustworthiness, data quality, protection of citizens' rights, protection of human rights, human dignity, democracy, data protection and privacy, personal data minimisation. Some of these values comprise other values as was further specified in the guidelines.

It was analysed as well how often the chosen values were elaborated in the guidelines (cf. grid 3). The following list shows the importance of the values due to the number of guidelines that mention them in a decreasing order (the ones at the beginning are mentioned most often, the ones at the end least often.)

Beneficence, non-maleficence/doing no harm, transparency, accountability (and assessment), responsibility, no bias/Fairness, safety/security, justice (in the sense of equal distribution), explicability/explainability, sustainability, data protection and privacy, autonomy, consent (and symmetry), protection of human rights, human dignity, democracy.

One comment on that: There was an exception made on the values of beneficence and non-maleficence. They were not always explicitly elaborated. But it can be said that they are important in all guidelines since they can be considered as meta principles or overarching principles.

In many cases the translation of recommendation to values was not unambiguous because this was not made explicit in the guidelines. In many cases an allocation of one recommendation to different values is possible. As well not only ethical norms are given but procedural norms as well.

7.2 Summation of analysis of the guidelines ten and onwards

Focus: 6 guidelines focus on Big Data and 9 guidelines focus on AI. Only 1 guideline refers to both.
Content: 7 guidelines are made for the use of AI. 3 guidelines focus on the development and design of AI. 6 guidelines focus on both topics.
User of guidelines: most guidelines appeal to a broad target group and are not specified for a particular class of users. Three guidelines are addressed to policy makers, 2 to end users and organisations and 1 to computer scientists. The rest are addressing a mixture of groups of users.
5 guidelines make reference to Education and teaching about SIS. For example 'Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data' support that parties […] to consider info and digital literacy as an essential educational skill. 'Ethics Guidelines for Trustworthy AI' support that 'Trustworthy AI encourages the informed participation of all stakeholders.'

References

AI Trends, “AI Glossary,” *AI Trends*, n.d. https://www.aitrends.com/ai-glossary/

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

Alexiou, Athanasios, Georgia Theocharopoulou, and Panayiotis Vlamos, "Ethical issues in neuroinformatics", in Papadopoulos, Harris, Andreas S. Andreou, Lazaros Iliadis, Ilias Maglogiannis (eds.), *Artificial Intelligence Applications and Innovations*. Springer, Berlin, Heidelberg, 2013, pp. 700-705.

Amnesty International, and Access Now, “The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems”, 2018. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf

Anagnostopoulos, Ioannis, Sherali Zeadally, and Ernesto Exposito, "Handling big data: research challenges and future directions", *The Journal of Supercomputing,* Vol. 72, No. 4, 2016, pp. 1494-1516.

Ajunwa, Ifeoma, Kate Crawford, and Jason Schultz, "Limitless worker surveillance", *California Law Review,* Vol. 105, No. 3, 2017, pp. 735-776.

Association for Data-driven Marketing and Advertising (ADMA), ‘‘Best practice guideline: Big Data’’, 2013, Sydney, http://datascienceassn.org/sites/default/files/Big%20Data%20Best%20Practice%20Guideline.pdf

AXELOS, “ITIL® Foundation, ITIL 4 Edition”, *ITIL® Foundation, ITIL 4 Edition*, TSO (The Stationery Office), n.d. https://www.tsoshop.co.uk/Business-and-Management/AXELOS-Global-Best-Practice/ITIL-4/?CLICKID=002289

B•Debate, ‘‘Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe’’, 2017. https://www.iiia.csic.es/barcelonadeclaration

Beck, Kent, Mike Beedle, Arie Van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andy Hunt, Ron Jeffries, Jon Kern, Brian Marick, R. C. Martin, Steve J. Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas, "Manifesto for agile software development", 2001.

Bellazzi, Riccardo, "Big data and biomedical informatics: a challenging opportunity", *Yearbook of medical informatics,* Vol. 23, No. 1, 2014, pp. 8-13.

Stiftung, Bertelsmann, and iRights.Lab, ‘‘Algo.Rules - Rules for the Design of Algorithmic Systems’’, 2019. https://algorules.org/en/home/

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.

Brey, Philip, “Freedom and Privacy in Ambient Intelligence”, *Ethics and Information Technology*, Vol. 7, No. 3, 2005.

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

Cave, Stephen, Rune Nyrup, Karina Vold, and Adrian Weller, “Motivations and Risks of Machine Ethics,” *Proceedings of the IEEE*, Vol. 107, No. 3, 2019, pp. 562–574.

Cerna Collectif, ‘‘Research Ethics in Machine Learning’’, 2018. https://hal.archives-ouvertes.fr/hal-01724307

CEN, “Ethics assessment for research and innovation - Part 1: Ethics committee”, CEN Workshop Agreement, June 2017. ftp://ftp.cencenelec.eu/EN/ResearchInnovation/CWA/CWA17214502.pdf

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.

Chelioudakis, Eleftherios, "Deceptive AI machines on the battlefield: Do they challenge the rules of the Law of Armed Conflict on military deception?", *Available at SSRN 3158711*, 2017.

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/

Costa, Angelo, Jaime A. Rincon, Carlos Carrascosa, Vicente Julian, and Paulo Novais, "Emotions detection on an ambient intelligent system using wearable devices", *Future Generation Computer Systems,* Vol. 92, 2019, pp. 479-489.

Costa, Fabricio F., "Big data in biomedicine", *Drug discovery today,* Vol. 19, No. 4, 2014, pp. 433-440.
Council of Europe, Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ‘‘Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data’’, January 23, 2017, Strasbourg. https://rm.coe.int/16806ebe7a
Council of Europe, “Artificial Intelligence: Glossary”, *Council of Europe*. https://www.coe.int/en/web/artificial-intelligence/glossary
Data Science Glossary, “Data Science Glossary”, *Data Science Glossary.* http://www.datascienceglossary.org/#kmeans
Dictionary.com, accessed 19th August 2019. https://www.dictionary.com/
European Group on Ethics in Science and New Technologies (EGE), ‘‘Statement on artificial intelligence, robotics and 'autonomous' systems’’, 2018, Brussels. https://doi.org/10.2777/531856
EUTAPI, “Glossary”, *European Patients’ Academy*, retrieved 22 July 2019. https://www.eupati.eu/glossary/
Floridi, Luciano and Cowls, Josh and Beltrametti, Monica and Chatila, Raja and Chazerand, Patrice and Dignum, Virginia and Luetge, Christoph and Madelin, Robert and Pagallo, Ugo and Rossi, Francesca and Schafer, Burkhard and Valcke, Peggy and Vayena, Effy, ‘‘AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’’, 2018. https://ssrn.com/abstract=3284141
Gall, Richard, “Machine Learning Explainability vs Interpretability: Two Concepts That Could Help Restore Trust in AI”, *KDnuggets*, December 2018. https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html
Garattini, Chiara, Jade Raffle, Dewi N. Aisyah, Felicity Sartain, and Zisis Kozlakidis, "Big data analytics, infectious diseases and associated ethical impacts", *Philosophy & technology,* Vol. 32, No. 1, 2019, pp. 69-85.
Gonçalves, Luís, “What Is Agile Methodology,” *Luís Gonçalves*, n.d. https://luis-goncalves.com/what-is-agile-methodology/
Goodman, Bryce, and Seth Flaxman, "European Union regulations on algorithmic decision-making and a “right to explanation”", *AI Magazine*, Vol. 38, No. 3, 2017, pp. 50-57.
Google, ‘‘Perspectives on Issues in AI Governance’’, n.d. https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf
Google Developers, “Machine Learning”, *Google Developers*, retrieved July 2, 2019. https://developers.google.com/machine-learning/glossary/#i
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.
High-Level Expert Group on Artificial Intelligence (AI HLEG), ‘‘Ethics Guidelines for Trustworthy AI’’, 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation
Hovy, Dirk, and Anders Søgaard, “Tagging Performance Correlates with Author Age”, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol. 2, Short Papers, 2015.
Hovy, Dirk, and Shannon L. Spruit, “The Social Impact of Natural Language Processing”, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* Vol. 2, Short Papers, 2016.
IBM, ‘‘Everyday Ethics for Artificial Intelligence. A Practical Guide for Designers & Developers’’, 2018. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf
ICT4LT, “Glossary of ICT Terminology”, *ICT4LT*, n.d. http://www.ict4lt.org/en/en_glossary.htm#GlossI
IEEE, *IEEE Guide for Developing System Requirements Specifications*, Institute of Electrical and Electronics Engineers, New York, NY, 1998.
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, ‘‘Ethically Aligned Design. Version 2 - For Public Discussion’’, 2018. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, “Glossary for Discussion of Ethics of Autonomous and Intelligent Systems. Version 1”, October 2017. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/eadv2_glossary.pdf

Information Commissioner's Office (ICO), ‘‘Big Data, Artificial Intelligence, Machine Learning and Data Protection’’, 2017. https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

Information Commissioner’s Office (ICO), “Big Data, Artificial Intelligence, Machine Learning and Data Protection”, 2017. https://www.ipc.on.ca/wp-content/uploads/2017/05/bigdata-guidelines.pdf

International Conference of Data Protection and Privacy Commissioners (ICDPPC), ‘‘Declaration on Ethics and Data Protection in Artificial Intelligence’’, 2018. https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf

ISACA (n.d.a), “COBIT 2019 Framework: Introduction and Methodology”, *COBIT 2019 Framework: Introduction and Methodology*, n.d. http://www.isaca.org/COBIT/Pages/COBIT-2019-Framework-Introduction-and-Methodology.aspx.;

ISACA (n.d.b), “COBIT 2019 Framework: Governance and Management Objectives”, *COBIT 2019 Framework: Governance and Management Objectives*, n.d. http://www.isaca.org/COBIT/Pages/COBIT-2019-Framework-Governance-and-Management-Objectives.aspx.

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

Langheinrich, Marc, “Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems”, *International conference on Ubiquitous Computing*. Springer, Berlin, Heidelberg, 2001, pp. 273–291.

Latonero, Mark, ‘‘Governing Artificial Intelligence: Upholding Human Rights & Dignity, Data and Society’’, 2018. https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

Lean Methods Group, “Data Analytics Glossary”, *Lean Methods Group*, n.d. https://www.leanmethods.com/resources/articles/data-analytics-glossary/

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.

Matthias, Andreas, "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and information technology,* Vol. 6, No. 3, 2004, pp. 175-183.

Mitrou, Lilian, Miltiadis Kandias, Vasilis Stavrou, and Dimitris Gritzalis, "Social media profiling: A Panopticon or Omniopticon tool?", *Proceedings of the 6th Conference of the Surveillance Studies Network*, 2014.

National Environmental Health Association (NEHA), “Definitions of Environmental Health”, *NEHA,* n.d. https://www.neha.org/about-neha/definitions-environmental-health

Nissenbaum, Helen, "Toward an approach to privacy in public: Challenges of information technology", *Ethics & Behavior,* Vol. 7, No. 3, 1997, pp. 207-219.

Organisation for Economic Co-operation and Development (OECD), “Recommendation of the Council on Artificial Intelligence”, *OECD Legal Instruments*, May 22, 2019. https://legalinstruments.oecd.org/api/print?ids=648&lang=en

PRO-RES, “Glossary of terms and concepts”, *PRO-RES*, n.d. http://prores-project.eu/glossary-of-terms-and-concepts/

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.

Resnik, David B, “Glossary of Commonly Used Terms in Research Ethics”, *National Institute of Environmental Health Sciences, National Institutes of Health*, May 2015. https://www.niehs.nih.gov/research/resources/bioethics/glossary/index.cfm.

Rouse, Margaret, “Black box”, *Search Software Quality*, 2018. https://searchsoftwarequality.techtarget.com/definition/black-box

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.

SATORI, “A reasoned proposal for shared approaches to ethics assessment in the European context”, *Deliverable 4.1*, May 19, 2017.
Search Encrypt, “7 Principles of Privacy By Design”, *Medium*, Search Encrypt, November 20, 2017. https://medium.com/searchencrypt/7-principles-of-privacy-by-design-8a0f16d1f9ce
Shearer, Colin, "The CRISP-DM model: the new blueprint for data mining", *Journal of data warehousing,* Vol. 5, No. 4, 2000, pp. 13-22.
Software & Information Industry Association (SIIA), ‘‘Ethical Principles for Artificial Intelligence and Data Analytics’’, *SIIA*, 2017. https://www.siia.net/Press/SIIA-Releases-Ethical-Principles-for-Artificial-Intelligence-and-Data-Analytics-with-Support-from-the-Future-of-Privacy-Forum-and-the-Information-Accountability-Foundation
Techopedia, “Data Ownership”, *Techopedia*, n.d. https://www.techopedia.com/definition/29059/data-ownership
Telefónica, ‘‘AI Principles of Telefónica’’, 2018. https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles
The Public Voice, ‘‘Universal Guidelines for Artificial Intelligence’’, 2018. https://thepublicvoice.org/ai-universal-guidelines/
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/
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/
UK Government, Department for Digital, Culture, Media & Sport, Data Ethics Framework, UK, 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/737137/Data_Ethics_Framework.pdf
UK Government, Department of Health & Social Care, Code of conduct for data-driven health and care technology, UK, 2019. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology
UNI Global Union, “10 Principles for Ethical Artificial Intelligence”, *The Future World of Work*, n.d. http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf
United Nations Development Group (UNDG), ‘‘Data Privacy, Ethics and Protection. Guidance Note on Big Data for Achievement of the 2030 Agenda’’, 2017. https://undg.org/wp-content/uploads/2017/11/UNDG_BigData_final_web.pdf
University of Montreal, ‘‘Montreal Declaration for a Responsible Development of Artificial Intelligence’’, 2017. https://www.montrealdeclaration-responsibleai.com/the-declaration