SHERPA

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

Guidelines for the Ethical Use 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).

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.[1] 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*.[2] In particular, these guidelines are built closely on the EU's High-Level Expert Group on Artificial Intelligence (AI HLEG). Our aim has been to build on their fundamental values, but we seek to go further in producing guidelines that are more operational and directly useful in development practices.

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

These guidelines begin by briefly describing the different types of requirements, starting with the top values (Section 2). Next, we describe how the ethical analyses can be mapped 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.

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

[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.

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

2. High-Level Requirements

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

In this Section we will briefly describe these high-level requirements to provide an insight into the fundamental principles and values behind the specific requirements. Readers who are familiar with the AI HLEG will notice that our high-level requirements are based directly on its high-level 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: 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[4] 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."[5] 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.

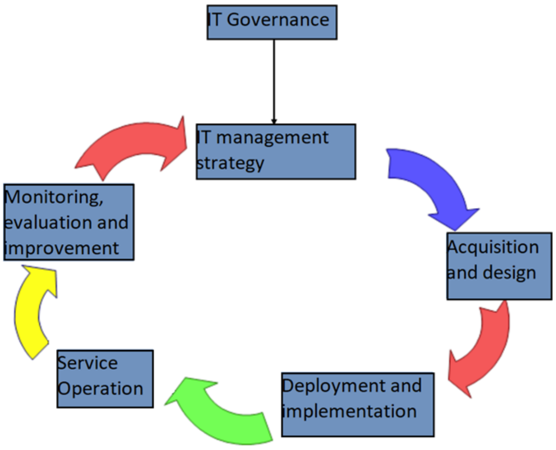
At the strategic level, there is no need to address which specific guidelines should be adopted for which IT-related purpose, although an engaged board can opt to issue more specific directives. To the extent that an organization has adopted broader ethics guidelines, corporate values, or corporate social 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.

[4] 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

[5] 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,[6] 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.

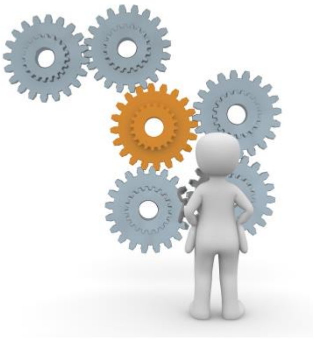
We will now consider how to apply ethical considerations to the use of AI and big data systems for each of these five processes. We will do so with special reference to the COBIT model, since it already defines various points at which it recommends the inclusion of ethics considerations (see Table 3,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

[6] 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"..[7]
>
> ***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.

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

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 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.

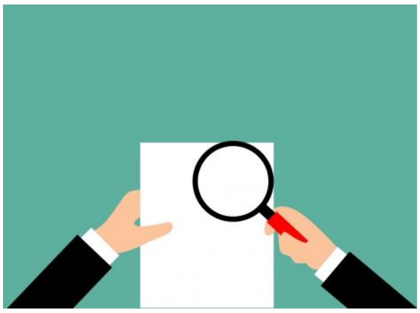
3.2.4 Service Operation

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..[8] 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.

[8] 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.

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.

2 Negative Liberty

Requirement 11: Fundamental rights.

In all phases, assess and ensure that:

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

3 Human Dignity

Requirement 12: Respect for Human Dignity.

In all phases (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.

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.

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).

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.

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.

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.

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 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).

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.

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).

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:

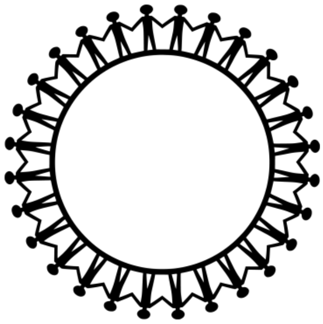
- a culture is established and encouraged in which open and structured communication is provided to stakeholders, in line with their requirements.
- information to stakeholders, end-users, and other affected persons about the system’s capabilities and limitations, is communicated in a clear, understandable, 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.

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.

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.

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:

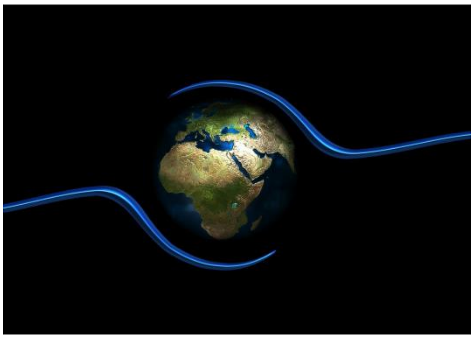
- an adequate inclusion of diverse viewpoints during the use of the system;
- that deployment is based on an acknowledgement that different cultures may respond differently, have different thought processes and patterns, and express themselves differently.

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.

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.

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).

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 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.

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.

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.

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 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.

5 Human Oversight

Requirement 71: *Avoiding automation bias.*
In the acquisition and design, deployment and implementation, and monitoring phases, assess and ensure:

- an appropriate level of human control for the system's use, 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 for work processes (e.g., by ensuring that safeguards are embedded before purchasing or deploying the system).

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,[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 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

[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.

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

The aim of most AI and big data systems in the domain of medicine is to make a transition from population-based healthcare to personalised medicine programs, by using the various data sources, data collecting devices, and data analytics to make medical recommendations using each patient's data records. This is becoming possible as medical records contain data including demographic information, information from laboratory tests, imaging and diagnostics data, as well as clinical notes and prior interventions.[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;

[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.

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

5.6 Covert and deceptive AI and big data systems

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

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

Requirements:

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

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

AI systems may interact with humans using spoken or written natural language, and may use an on-screen appearance of an animated person, or avatar. Without an avatar, they may still take on an identity as if they were a person (e.g., Alexa, Siri). These systems are called conversational agents. AI may also be embedded in robots that resemble humans in their appearance and movements. The recognition and expression of emotions may result in better interaction with human users, but also raises ethical issues. The recognition and processing of human emotions may infringe on human autonomy, freedom and privacy. The expression of emotions by machines may lead to unwanted attitudes and beliefs in humans, who may be deceived or manipulated and develop unwanted attachments.

Requirements:

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

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

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

When this level of tracking, monitoring and messaging is performed for political purposes, it contains risks of political manipulation of voters through psychologically exploitative microtargeting and distribution of fake news as part of misinformation campaigns.[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]

[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.

Requirements:

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

5.9 AI and big data systems in defence

The deployment of AI and big data systems in defence contexts occurs in a wide range of applications. These include: conventional military defence (e.g. development of military AI), counter-nuclear proliferation, counter-chemical/biological WMD, counter-terrorism, and cybersecurity as well as counter-intelligence. These applications have data sources that range from human actors, geospatial tools (e.g. mapping and satellite data), measurement and signature sensing tools (i.e. for identifying distinctive features of emitters), as well as online data.[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;

[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.

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

5.10 Ethically aware AI and big data systems

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

Firstly, ethically aware AI may be considered problematic due to the nature of ethics. Ethics is not an algorithmic exercise of applying systematically ranked moral principles to situations.[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;

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

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