D4.7: An Ethical framework for the development and use of AI and robotics technologies

[WP4: Artificial Intelligence and Robotics - Ethical, Legal and Social Analysis]

Lead contributor	Prof. dr. Philip Brey, University of Twente
	p.a.e.brey@utwente.nl
Other contributors	Philip Jansen, University of Twente
	Jonne Maas, University of Twente
	Björn Lundgren, University of Twente
	Anaïs Resseguier, Trilateral Research
Reviewers	Laura Crompton
	Bernd Carsten Stahl
Commentator	Rowena Rodrigues

Due date	31 March, 2020
Delivery date	8 April 2020 (after request for small extension)
Type	Dissemination Public
Dissemination level	PU = Public

Keywords	Artificial Intelligence; robotics; ethical framework; ethics by design; research ethics; policy; education; standards

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Abstract

This report proposes a comprehensive strategy for ethical AI and robotics. That is, it proposes, at least in outline, a comprehensive set of methods and procedures for developing, deploying and using AI and robotics systems in a way that adheres to ethical principles. The strategy that we propose addresses all actors in society, particularly developers, deployers, users, regulators and educators. It proposes various methods towards more ethical development and use of AI and robotics, such as methods for incorporating ethical considerations into design and development processes, guidelines for ethical deployment and use of AI and robotics systems, standards and certification, governmental policies and regulations, and education and training programs.

Within this general strategy, we pay particular attention to methods and procedures for ethical research and innovation (R&I) in AI and robotics. We propose an approach to Ethics by Design, which is the systematic inclusion of ethical guidelines, recommendations and considerations into design and development processes. We propose both a generic approach to Ethics by Design, and specific approaches within the framework of three popular development methodologies: CRISP-DM, Agile and V-Model. We conclude this report by looking forward to the steps that still need to be taken to further develop and implement our strategy.

Document history

Version	Date	Description	Reason for change	Distribution
V0.9	01 03 2020	Final draft for external review	-	01 03 2020
V1.0	08 04 2020	Final report for submission to the EC	Reviews and comments	08 04 2020
V1.1	08 07 2020	Revision request EC	Very minor revisions	08 07 2020

Information in this report that may influence other SIENNA tasks

Linked task	Points of relevance
D5.4	The code of responsible conduct for AI and robotics will require consideration of the issues identified in this report.
D6.1	The report on adapting methods for ethical analysis of emerging technologies will require contemplation about the successes and challenges in the methodology used to write this report.
D6.3	The step-by-step guidance from ethical analysis to ethical codes and operational guidelines task will require reflection about the successes and challenges in writing this report.
D6.4	The process of obtaining buy-in for the codes from EU and international institutions will need to build on the proposals in this report.

Table of contents

Executive summary

This report contains a comprehensive strategy for ethical AI and robotics. That is, it proposes, at least in outline, a comprehensive set of methods and procedures for developing, deploying and using AI and robotics systems in a way that adheres to ethical principles.

The report contains an introductory section, in which the objectives, scope and limitations of the report are set out, two main sections in which our strategy is presented, and finally a concluding section and two annexes. The two main sections of the report are sections 2 and 3. Section 2, "A Strategy for AI and Robotics," proposes the overall strategy for promoting ethical AI and robotics. It is stated that a strategy for ethical AI and robotics should contain three components: (1) an identification of relevant actors; (2) an identification of methods that these actors can use to contribute to ethical AI & robotics, and (3) proposals of ways in which these methods can be made available to these actors, and ways to motivate them to use them. Following this proposal, the report continues to identify main classes of relevant actors who can bring about ethical AI and robotics: AI & robotics developers; AI & robotics development support organizations; organizations that deploy and use AI & robotics technology; governance and standards organizations; educational and media organizations; and civil society organizations and the general public.

Next, six types of methods for ethical AI & robotics are discussed and related to these classes of actors:

1. Methods for incorporating ethics into research and development of AI & robotics (aimed at AI & robotics developers and support organizations). These methods include research ethics guidelines and protocols for R&I in AI & robotics, ethical impact assessment methodologies for emerging AI & robotics, Ethics by Design methodologies for AI & robotics and codes of professional ethics for researchers and developers of AI & robotics technologies.

2. Methods for incorporating ethics into the deployment and use of AI & robotics (aimed at organisations that deploy and use AI & robotics technology). These methods include operational ethics guidelines and protocols for the deployment and use of AI and robotics technologies for the enhancement of organisational processes and for their deployment and use in products and service, codes of professional ethics for IT professionals and managers in user organisations, and end-user guidelines.

3. Corporate responsibility policies and cultures that support ethical development and use of AI & robotics (aimed at both developers, deployers/users and support organizations)

4. National and international guidelines, standards and certification for ethical AI & robotics (aimed at governance and standards organisations; indirectly affecting developers, deployers/users and support organizations)

5. Policy and regulation to support ethical practices in AI & robotics (aimed at governance and standards organisations; indirectly affecting developers and deployers/users)

6. Education, training and awareness raising for the ethical and social aspects of AI & robotics (aimed at educators and the media)

Section 3, "A framework for Ethics by Design", contains a detailed proposal for methods for incorporating ethical criteria into the design and development methodologies for AI and robotics. It first proposes a generic method for doing this, which is independent of particular existing methodologies for the development of AI and robotic systems. This model distinguishes three (iterative) phases in systems design: specification of objectives, specification of requirements, high-level design, the optional process of data collection and preparation, detailed design and development, and testing and evaluation. For each phase, it then specifies how ethical considerations can be made part of it. For example, in the specification of objectives phase, the proposed objectives of the system are evaluated against ethical requirements, and in the high-level design phase, the proposed design is evaluated against ethical requirements, especially ones relating to transparency, autonomy, privacy and fairness.

Subsequently, proposals are made for the integration of ethical criteria within three popular AI and robotics development methodologies: CRISP-DM, Agile and the V-Model. In two annexes to the report, moreover, detailed ethical guidelines are proposed for the incorporation of ethical criteria into Agile and the V-Model.

In a concluding section of the report, the results of the study are summarized and future work towards further implementation is discussed.

List of figures

List of tables

List of acronyms/abbreviations

Abbreviation	Explanation
AI	Artificial intelligence
EC	European Commission
HR	Human resources
R&D	Research and development
R&I	Research and innovation

Table 1: List of acronyms/abbreviations

Glossary of terms

Term	Explanation
Artificial Intelligence	The science and engineering of machines with capabilities that are considered intelligent (i.e., intelligent by the standard of *human* intelligence).
Big Data	Extremely voluminous data sets that require specialist computational methods to uncover patterns, associations and trends in them.
Data mining	The process of discovering patterns in large data sets involving database systems, statistical analysis, and intelligent methods such as machine learning.
Deep learning	An approach to machine learning that applies artificial neural networks with hidden layers and the backpropagation method, in combination with powerful computer systems and voluminous training data.
Ethics by Design	The systematic inclusion of ethical guidelines, recommendations and considerations into design and development processes.
Intelligent agent	An artificially created, autonomous entity that can perceive its

	environment by means of sensors, act upon this environment through the use of actuators, and direct its activities towards reaching goals.
Machine learning	A set of approaches within AI where statistical techniques and data are used to "teach" computer systems how to perform particular tasks, without these systems being explicitly programmed to do so.
Risk assessment	a systematic process of evaluating the potential risks that may be involved in a projected activity or undertaking.
Robotics	The field of science and engineering that deals with the design, construction, operation, and application of robots.
Robot	Electro-mechanical machines with sensors and actuators that can move, either entirely or a part of their construction, within their environment and perform intended tasks autonomously or semi-autonomously.

Table 2: Glossary of terms

1. Introduction

1.1 Background

This report has been developed within the SIENNA project, a European Horizon 2020-funded project on the ethical and human rights dimensions of emerging technologies.[1] A major focus of the SIENNA project is on the ethical and human rights aspects of AI and robotics. We have already performed extensive studies of ethical aspects of AI and robotics, the legal and human rights context for AI and robotics, existing ethical codes and guidelines for AI and robotics, the state of the art in AI and robotics and its social and economic impacts, and public awareness and acceptance of AI and robotics.[2] This is the first study in which we develop our own proposals. Based in part on our previous studies, we hereby propose an extensive ethical framework for the development and use of AI and robotics technologies.

1.2 Objectives

This report proposes a comprehensive strategy for ethical AI and robotics. That is, it proposes, at least in outline, a comprehensive set of methods and procedures for developing, deploying and using AI and robotics systems in a way that adheres to ethical principles. The strategy that we propose addresses all actors in society, particularly developers, deployers, users, regulators and educators. All have a role in bringing about ethical AI and robotics. Within this general strategy, we pay particular attention to methods and procedures for ethical research and innovation (R&I) in AI and robotics. Ethical R&I is often key for ensuring ethical standards for new technologies. In R&I, major decisions are made about what technological solutions to develop and which ones not to develop, and R&I often comes with prescriptions about deployment and usage as well. However, we will also pay attention to methods for ethical deployment and use, and to the role of organisations that market and use AI and robotics, technologies, as well as policy makers, regulators and educators, in bringing it about.

1.3 Structure of the report

The main body of the report consists of two parts after this introduction (section 1). Section 2, “A Strategy for AI and Robotics,” proposes an overall strategy for promoting ethical AI and robotics. It starts with an identification of relevant actors and six categories methods for obtaining ethical AI & robotics. It then proceeds to discuss the six categories of methods in more detail, and concludes with a section on how the methods can be developed (further) and how actors can be motivated to use them. Section 3, “A framework for Ethics by Design”, contains a detailed proposal for methods for incorporating ethical criteria into the design and development methodologies for AI and robotics. It first proposes a generic method for doing this, after which it contains a detailed discussion of doing it in relation to three popular development methodologies: CRISP-DM, Agile and the V-Model. In a

[1] See https://www.sienna-project.eu/.
[2] See reports D4.1, D4.2, D4.3, D4.5 and D4.6 at https://www.sienna-project.eu/publications/

concluding section (4), the results of the study are summarized and future work towards further implementation is discussed. Finally, in two annexes, detailed ethical guidelines are proposed for the incorporation of ethical criteria into Agile and the V-Model.

The role of ethical principles

It is not an objective of this report to develop or propose general ethical principles or guidelines for AI and robotics. By now, there is already enough convergence, in our opinion, on ethical principles for AI and robotics. Over the course of 2019, in particular, many countries and international organizations proposed general ethical guidelines for AI. Notably, 2019 saw the Ethics Guidelines for Trustworthy AI of the High-Level Expert Group on Artificial Intelligence (HLEG-AI, 2019), the Recommendation of the Council on Artificial Intelligence of the OECD (2019), the guidelines for Ethically Aligned Design from the Institute of Electrical and Electronics Engineers (IEEE, 2019), and the Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence China's Ministry of Science and Technology (2019).

As several analysts have observed, there is a remarkable convergence between these recent sets of ethical guidelines. This was concluded, amongst others, in a recent study of the EU Horizon 2020—funded SHERPA project [FN], which was co-authored by some of the authors of this study (Ryan et al., 2019). The three main sets of guidelines (HLEG-AI, OECD and IEEE) display remarkable agreement in content, even though they have different formats and wordings. These documents are in essential agreement, it was found, on nine key ethical principles that include *privacy, autonomy, freedom, dignity, safety and security, justice/fairness, responsibility/accountability, well-being (individual, societal and environmental) and transparency*. In addition, none of these documents proposed major principles outside of this list. Even the Chinese guidelines converges remarkably with more "Western" guidelines: they by and large reflect these ethical principles as well.

1.4 Scope and limitations

In this report, as well as in future SIENNA proposals, we will adopt these nine key ethical principles as a starting point for ethical guidance. Specifically, given that this is a European Union funded project, we will adopt, with minor adaptations, the European version of these principles. That is, we will adopt the ethics guidelines for trustworthy AI of the HLEG-AI as our guiding set of principles, specifically its seven ethics requirements for trustworthy AI in which these nine principles are contained: *Human agency and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination and fairness; Societal and environmental well-being; and Accountability*. Because of the strong similarities between these guidelines and others used outside the European union, we expect this study to have applicability outside the European Union as well.

These kinds of general guidelines will not be sufficient to offer ethical guidance for particular products and applications, or specific contexts of use. More detailed guidelines will also be needed to address such issues, for example, ethical guidelines for unmanned aerial vehicles, or for healthcare applications of AI, or for predictive data analytics techniques. When needed, we will propose such more detailed guidelines. Our greatest concern in this report, however, is to operationalize ethical guidelines: how to make them directly usable by particular actors for particular practices. This is what much of this report will center on.

Particular attention will be paid to methods for the ethical development of AI & Robotics technologies. A large part of the report will be focused on such methods, under the heading of *Ethics by Design*. Section 3 of the main body of the report will be devoted to it, as well as the two annexes, which will develop Ethics by Design for two of the most often used development methodologies in AI and robotics.

2. A Strategy for Ethical AI and Robotics

As we argued, a set of ethical guidelines or principles is only one component of a strategy for ethical AI & robotics. It could provide some direction to activities, but only in a very general sense. Many more elements need to be in place to achieve the objective of ethical AI & robotics. Consider, for example, the development of AI & robotics technologies. Developers and other stakeholders involved, like most people, have certain ethical views and moral leanings that they respect. However, this may colour the development process. When given a list of ethical principles for AI, some developers may endorse them and make attempts to adhere to them in their activities. Such a set may point developers to actively focus on ethics during the development process. A set of principles, nevertheless, may not always be successful. Programmers could easily fail to do so due to either a lack of training in ethics, lack of knowledge of how to apply ethical principles in technology development, lack of support from management, lack of inclusion of ethics criteria in quality assessment frameworks or corporate social responsibility strategies, or other reasons. Much more is needed to make actors both motivated and competent in the incorporation of ethical considerations in their practices, and to support actors in collaborative practices towards this shared objective.

A sound strategy for ethical AI & robotics should in our view do three things:

- *Identify relevant actors*
- *Identify methods that these actors can use to contribute to ethical AI & robotics*
- *Propose ways in which these methods can be made available to these actors, and ways to motivate them to use them*

An overall strategy will be proposed in this report. Such a strategy is, in our view, a first step towards realizing ethical AI & robotics. A second step is the successful implementation of the strategy by relevant actors. Implementation will be a large part of the future focus of the SIENNA project, and future deliverables (particularly D5.4 and D6.6) will reflect this focus.

We will now proceed to identify the most relevant actor categories, and then propose relevant methods for each of them, including some shared methods that apply to different actor categories. We will end with a brief discussion of ways to make the methods available to actors and ways to motivate them.

Actors

The following actor categories are most relevant for our purposes. They have been selected on the basis of having the most influence on how AI & robotics technologies are developed, used, and perceived, and thereby on what their impacts and ethical aspects are:

1. AI & robotics developers
2. AI & robotics development support organizations
3. Organizations that deploy and use AI & robotics technology

4. Governance and standards organizations
5. Educational and media organizations
6. Civil society organizations and the general public

We will now discuss them in turn.

1. *AI & robotics developers*

Within this broad category, we can make some further distinctions. At the organizational level, developers include firms that develop AI & robotics technologies and research institutes (universities and other research performing organizations) that engage in research and innovation in AI & robotics. At the intra-organisational level, there are various units within these institutions that are involved in the planning, support and carrying out of R&I activities. At the individual level, there are also professionals in various roles (e.g., IT project manager, IT director, hardware technician, professor in robotics) that are actors in AI & robotics development.

2. *AI & robotics development support organizations*

These are organizations that provide support to the R&I activities of AI & robotics firms and research institutes. These include business and industry organisations (also known as trade organisations): organisations that support companies in a certain sector; chambers of commerce; research funding organisations; investment banks and other investors and funders; associations of universities and research institutes; science academies and associations of science academies; professional organisations for the AI & robotics fields; advisory and consultancy firms for companies and research institutes.

3. *Organizations that deploy and use AI & robotics technology*

These are private and public organisations that use AI & robotics. Its usage can be intended to improve or support various organizational functions, including operations, finance, marketing, human resources, customer service, and other. Within these organisations, one can furthermore define various units and professional roles associated with the deployment and use of AI systems within or by the organization, such as information technology managers, database administrators, and development operations engineers. Note that some organizations are simultaneously developers and users of AI & robotics systems. For example, tech companies like Apple and Google develop AI technologies, but also use them within their own organization.

4. *Governance and standards organisations*

These are organisations involved in developing, implementing or enforcing policies, standards and guidelines, specifically those regarding the development, deployment and use of AI & robotics technologies. It should be noted that organizations also make policies and guidelines for themselves. These are not our concern here. This category is intended to refer to organizations that develop or implement guidelines, policies, regulations and standards for others. This includes, first of all, national, local and supranational governments, as well as government-instituted or -supported advisory and regulatory bodies. They also include intergovernmental organisations like the United Nations, the Council of Europe, and the World Health Organization. Also included in this category are national and international standards, certification, quality assurance, accreditation and auditing organisations. Policies, standards and guidelines can also be issued by many of the AI & robotics development support organisations discussed earlier.

5. *Educational and media organisations*

Educational institutes and media organisations both have a significant role, albeit a quite different one, in shaping people's knowledge and understanding of AI & robotics, the ethical issues associated with them, and the ways in which these ethical issues can be addressed. Educational organisations, from elementary school to postgraduate education, provide the major vehicle by which individuals acquire knowledge, skills and insights regarding AI & robotics, their impacts on society, their ethical aspects, and ways to address ethical issues in their profession. Of course, not only educational organisations provide education and training. Companies may, for example, organize their own in-house trainings as well. Media organisations have a large role in generating public awareness and understanding of AI & robotics and the ethical issues raised by them and therefore should also be recognized as actors with respect to ethical AI & robotics.

6. *Civil society organisations and the general public*

Civil Society Organisations (CSOs) are non-governmental, not-for-profit organisations that represent the interests and will of citizens. They may be based on cultural, political, ethical, scientific, economic, religious or philanthropic considerations. They include civic groups, cultural, groups, consumer organisations, environmental organisations, religious organisations, political parties, trade unions, professional organisations, non-governmental policy institutes, activist groups, and several other kinds. Many CSOs want to have a role in public policy or influence the way that organizations function in which they have an interest. For some of them, the development and use of AI will be a concern, and as a result, these organisations will function as agents with respect to public policy and the actions of relevant other organisations. The general public, finally, can also perform as an actor, through its public opinions, voting patterns, consumer purchases, and use or nonuse of AI & robotics products and services.

Finally, it is worth mentioning that amongst and within these various kinds of organisations and units, there are also those that have a specific focus on ethics. These include ethics research units, ethics policy units, ethics officers, research ethics committees, integrity offices and officers, corporate social responsibility units and officers, ethics educational programmes, ethics advisory bodies, and national and international ethics committees. However, ensuring ethical standards and practices is not only the responsibility of such organisations and units; all of the listed actors have such responsibilities, although ethics organisations and units will often have a special role in ensuring the proper inclusion of ethics concerns in practices.

Methods

In the context of this report, methods are means by which actors can implement ethical guidelines and considerations. Our identification of methods for ethical AI & robotics builds on earlier proposals of the HLEG-AI (2019) and IEEE (2019). Both reports propose methods for the implementation of ethical guidelines in relation to different actors. The HLEG makes a distinction between what they call technical and non-technical methods, both of which apply to all stages of the development and use lifecycle of AI systems. Technical methods include ethics by design methods, explanation methods for transparency, methods of building system architectures for trustworthiness, extensive testing and validation, and the definition of quality of service indicators. Non-technical methods include regulation, codes of conduct, standardization, certification, accountability via governance frameworks, education and awareness to foster an ethical mindset, stakeholder participation and social dialogue, and diverse and inclusive design teams.

The IEEE (2019) report has a chapter on "methods to guide ethical research and design" for researchers, technologist, product developers and companies (pages 124-139), and a chapter on policies and regulations by governing institutions and professional organizations (pages 198-210). In its methods for ethical R&D chapter, it considers both individual and structural approaches, and distinguishes between three overall approaches: interdisciplinary education and research, corporate practices on AI & robotics, and responsibility and assessment. In its policy chapter, the IEEE advocates methods such as the founding of national policies and business regulations for SIS on human rights approaches, the introduction of support structures for the building of governmental expertise in AI and robotics, and the fostering of AI & robotics ethics training in educational programs.

The methods proposed by the HLEG-AI and IEEE are partially overlapping and in part complementary. Drawing from them, we propose six sets of methods for the ethical development and use of AI & robotics[3], for the different classes of actors that were defined earlier:

1. Methods for incorporating ethics into research and development of AI & robotics (aimed at AI & robotics developers and support organizations)
2. Methods for incorporating ethics into the deployment and use of AI & robotics (aimed at organisations that deploy and use AI & robotics technology)
3. Corporate responsibility policies and cultures that support ethical development and use of AI & robotics (aimed at both developers, deployers/users and support organizations)
4. National and international guidelines, standards and certification for ethical AI & robotics (aimed at governance and standards organisations; indirectly affecting developers, deployers/users and support organizations)
5. Education, training and awareness raising for the ethical and social aspects of AI & robotics (aimed at educators and the media)

[3] Points 1, 3-6 are taken from the SHERPA development and use guidelines (Brey, Lundgren, Macnish and Ryan, 2019). Point 2 is an added point.

6. Policy and regulation to support ethical practices in AI & robotics (aimed at governance and standards organisations; indirectly affecting developers and deployers/users)

We will refrain, for now, to propose methods for CSOs and the general public, taking into account that their role in ethical AI & robotics is often more indirect. We will now discuss these sets of methods in some more detail and relate them to the roles and responsibilities of different actors.

Methods for incorporating ethics into research and development

These are methods for making ethical considerations, principles, guidelines, analyses or reflections part of research and development processes. They apply to the first actor category identified above: AI & robotics developers. Four main classes of methods fall into this category:

1. Research ethics guidelines and protocols for R&I in AI & robotics
2. Ethical impact assessment methodologies for emerging AI & robotics
3. Ethics by design methodologies for AI & robotics
4. Codes of professional ethics for researchers and developers of AI & robotics technologies

We will now discuss them in turn.

1. *Research ethics guidelines and protocols for R&I in AI & robotics*

Research ethics guidelines and protocols for AI & robotics are ethics guidelines and procedures by which researchers, developers, research ethics committees and ethics officers can ethically assess R&I proposals and ongoing R&I practices. Such ethical assessments may or may not be accompanied with specific recommendations to proceed differently. They can, in either case, be used to improve R&I plans and practices so as to make them more ethical. As of the moment of publication of this report, few research ethics guidelines and protocols specifically for AI and robotics were in existence (see our report D4.3 Survey of REC approaches and codes for Artificial Intelligence & Robotics). While there is an abundance of general ethical guidelines for AI and robotics, few specifically focus on R&I practices and on the role of research ethics committees. We are currently working on our own proposal for research ethics guidelines and protocols for AI & robotics, and will present them in a future SIENNA report.

2. *Ethical impact assessment methodologies for emerging AI & robotics*

Ethical impact assessment methodologies are methods for assessing present and potential future impacts of emerging technologies, including specific products and applications, and identifying ethical issues associated with these impacts. EIA, in short, is an approach for assessing not only present but also potential future ethical issues in relation to a technology. EIA, in its current form, was developed within the EU FP7 SATORI project [FN]. It has also been developed into a CEN standard (CEN, 2017). EIA is not just a method for AI & robotics developers, but can also be used, amongst others, by governments in order to support technology policy, and by research funding organisations to help set priorities in research funding.

3. *Ethics by design methodologies for AI & robotics*

Ethics by design methodologies for AI & robotics are methods for incorporating ethical guidelines, recommendations and considerations into design and development processes. They fill a gap that exists in current research ethics approaches, which is that it is often not clear for developers how to implement ethical guidelines and recommendations, which are often of a quite general and abstract nature. Ethics by design methodologies identify how at different stages in the development process, ethical considerations can be included in development, by finding ways to translate and operationalize ethical guidelines into concrete design practices. Ethics by design approaches have been in existence in computer science and engineering since the early 1990s, initially under the name Value-sensitive design (Friedmann Kahn & Borning, 2006) and later also under the label of Design for Values (Van den Hoven, Vermaas and Van de Poel, 2015). In recent years, the term "ethics by design" has come into vogue. Recently, an extensive ethics by design approach for AI was published as part of the EU Horizon 2020-funded project SHERPA (Brey, Lundgren, Macnish and Ryan, 2019). As far as we can see, no other full-blown ethics by design approaches have yet been published for AI & robotics, although the IEEE is working on one. In this report, we build on the SHERPA report to present an extended approach for ethics by design that has wider applicability than the one proposed in that report.

4. *Codes of professional ethics for researchers and developers of AI & robotics technologies*

Codes of professional ethics, also called codes of conduct, are codified personal and corporate standards of behaviour that are expected in a certain profession or field. These codes are often set by professional organisations. To our knowledge, no internationally accepted codes of ethics for either artificial intelligence specialists or robotics engineers are currently in existence, and few if any national codes for these professions exist either. Wider codes of ethics, for computer scientists and electrical engineers, are in existence and cover the AI and robotics professions as well. However, these broader codes do not address the specific challenges and responsibilities of AI and robotics specialists. In this report, we do not attempt to propose codes of professional ethics for these professions. We could make some initial proposals, however, in later studies in the SIENNA project.

In the HLEG and IEEE reports, various other methods for incorporating ethics into R&D are mentioned. Some of these can however, in our opinion, be subsumed under ethics by design approaches. These include, amongst others, the development and use of explanation methods for transparency, extensive testing and validation, the definition of quality of service indicators, and better technical documentation. Others will be discussed under the heading of "corporate social responsibility cultures" below. One method merits special attention, however: interdisciplinary research, which is proposed in the IEEE report. Interdisciplinary research is, in our view, an important component of ethical AI & Robotics, if it involves collaborations that bring engineers and scientists into contact with social science and humanity scholars, including ethicists. Such research activities allow for a better incorporation of social and ethical concerns into engineering practice, and are therefore highly advisable, at different stages of the R&D continuum.

Methods for incorporating ethics into the deployment and use of AI & robotics

After the development of AI & robotics systems, services and solutions, they are deployed by organisations or individuals in order to be used.[4] The deployment and use of these technologies often require their own ethical guidelines and solutions, that are to some extent different from those that apply to their development. Ethical questions that are typically asked in relation to deployment and use include questions like: Is it ethical to deploy a system that is intended to do X / is capable of doing X / can be used to do X? How can unethical uses of the system be monitored and prevented? What is the responsibility of different actors in preventing or mitigating unethical use? What policies to prevent unethical use should be put in place and how can they be implemented effectively?

Deployment and use scenarios come in various forms, but the following are the most typical:

(1) Deploying AI or robotics technology to enhance organisational processes. An organisation acquires AI or robotics technology, and uses it within its own organisation to improve organisational processes such as manufacturing, logistics, and marketing. End-users are IT specialists or other employees in the organisation.

(2) Embedding AI and robotics technology in products and services. An organisation acquires AI or robotics technology, and incorporates it into products or services that it offers to customers. This is a different application of AI and robotics than its application in the development, manufacturing and marketing of products and services. For example, AI can be used to better design, manufacture or market automobiles that themselves do not contain AI technology. AI and robotics technologies can be embedded in products and services for different purposes:

a. To enhance the value of a product or service for customers by offering enhanced functionality or usability. E.g., by powering an online dating service with AI algorithms, or by enhancing an automobile with a self-drive mode.

b. To enhance the value of a product or service through intelligent monitoring, self-repair, communications with customer service, or data collection for future upgrades.

c. To further the interests of the organisation or of third parties, for example, by collecting data for marketing purposes or allowing for targeted messaging.

It is not always clear who is the end-user of the AI and robotics technology in these three scenarios, since the end-user of AI or robotics technology embedded in a product or service may be different from the end-user of that product or service, and there may also be multiple end-users (e.g., Uber drivers and customers using the same AI algorithms).

Taking these scenarios into consideration, the following four methods can contribute to ethical deployment and use of AI & robotics technologies:

[4] Of course, deployment and use cycles are often followed by repeated redevelopment of systems.

(1) Operational ethics guidelines and protocols for the deployment and use of AI and robotics technologies for the enhancement of organisational processes

(2) Operational ethics guidelines and protocols for the deployment and use of AI and robotics technologies in products and services

(3) Codes of professional ethics for IT managers, technical support specialists and other management, IT and engineering staff responsible for the deployment and use of the AI & robotics technologies in an organisation or its embedding in products and services

(4) End-user guidelines for ethical usage of (products and services that include) AI and robotics technologies

In Brey, Lundgren, Macnish and Ryan (2019), the previously mentioned SHERPA report, proposals were made for the first and, to some extent, the second of these methods. Building on two widely used models for the management and governance of information technology in organisations, ITIL and COBIT, as well as on the ethics requirement of the High-Level Expert Group on AI, this report proposed operational guidelines for the deployment and use of AI systems (including AI-powered robotic systems) in organisations. We will not do further work on these guidelines in this report. We also will not attempt to further develop codes of professional ethics for the different professions responsible for the deployment and use of AI & Robotics technologies. Often, codes of ethics will be in place for these professions, but they might need updates to take into account the specific demands imposed by AI & robotics technologies. We also will not attempt to develop (generic) guidelines for end-users in the context of this report.

Corporate responsibility policies and cultures

Ethical guidelines and professional ethical codes, even when fully operationalized for particular practices, will have little impact if they are not supported by organisational structures, policies and cultures of responsibility. In Brey, Lundgren, Macnish and Ryan (2019), specifically the division of the report with guidelines for the ethical deployment and use of AI (p. 53-87), an attempt was made to include these wider considerations of responsibility in organisations in the guidelines that were proposed. For instance, requirement 1 in this report, which targets the board of directors of companies, reads as follows:

> 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 (p. 61).

Requirements 2, 3 and 4, which targets IT management, are as follows:

> 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 (p. 64).
>
> 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 (p. 64-65).
>
> 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 (p. 65).

These guidelines, and several others that are proposed, serve as meta-guidelines for the proper implementation of ethics guidelines for AI & robotics in organizations. They point out that proper implementation of ethics considerations in organizations involves much more than the development and distribution of operationalized ethics guidelines, but also requires leadership from the top, adjustment of existing management strategy, definitions of roles and responsibilities, training of staff, monitoring and assurance activities, and encouragement of a common culture of responsibility. While these guidelines were developed for organisations that deploy and use AI & robotics technologies, they are also applicable to organizations that engage in AI & Robotics R&D.

National and international guidelines, standards and certification

In this report, we distinguish between *operational ethics guidelines*, which are detailed, practical guidelines developed for specific practices by specific actors, and *general ethics guidelines*, which are statements of ethical principles and general guidelines that apply to a broad range of actors and practices. While it is possible to develop operational guidelines without general guidelines, it is often beneficial to have shared general guidelines on the basis of which operational guidelines are developed. These guidelines can be supported by national governments and intergovernmental organisations. Currently the two most important sets of international guidelines for AI & robotics technologies are the Recommendation of the Council on Artificial Intelligence of the OECD (2019) and the Ethics Guidelines for Trustworthy AI of the High-Level Expert Group on Artificial Intelligence of the European Commission (HLEG-AI, 2019). These two documents currently serve as the two most important international guidance documents for ethical issues in AI & robotics.

Next to such general guidelines, which are directed at all actors, there are also ethical guidelines that are general rather than operational, but that are focused on specific actors or practices. The guidelines for Ethically Aligned Design from the Institute of Electrical and Electronics Engineers (IEEE, 2019) are a case in point. These specifically apply to design practices, and are of greatest relevance to technology developers.

Standards, developed by recognized national and international standards organisations or by particular (associations of) companies or organisations, are different from ethics guidelines in two ways. First, they apply to specific products, services, processes or methods, while ethics guidelines apply to any action, thing or event that has ethical implications. Second, they define specific norms or requirements to which the phenomenon to which the standard applies must adhere. Standards are intended to leave limited room for subjectivity and interpretation, and are intended to define intersubjective requirements that different actors can apply, identify or assess.

Standards sometimes aim to codify ethical requirements, procedures or methods. Examples are ISO 26000, which is an international standard for corporate social responsibility, CEN CWA 17145-1, which is a standard for ethics assessment by research ethics committees, and CEN CWA 17145-2, which is a standard for the method of ethical impact assessment for R&I. Standards can also *include* ethical requirements, procedures or methods, while not themselves having ethics as a focus. For example, ethics is discussed in the context of the ISO 9000 and 9001 standards for quality management.

For AI & robotics, a remarkable number of ethical standards are currently being developed by IEEE as part of its Ethically Aligned Design programme (IEEE, 2019). A total of 13 standards are in development, including standards for ethics by design, transparency of AI systems, algorithmic bias, data privacy, ethically driven robotics and automation systems, and automated facial analysis technology. ISO also has several standards in development that focus in part or in whole on ethical issues, including standards for identifying ethical and societal concerns in AI systems, bias in AI systems, trustworthiness of AI systems, quality assurance in AI and risk assessment in AI.

Certification is the process by which an external third party (typically a certifying body) verifies that an object, person or organization is in possession of certain characteristics or qualities. Amongst others, certification can be applied to persons, in professional certification, to products, to determine if it meets minimum standards, and to organizations or organizational processes, through external audits, to verify that they meet certain standards. Certification can be a means to verify and validate the quality of ethics processes and procedures in organisations. In relation to standards, in particular, certification can be a means of ensuring conformity to the requirements of the standard. IEEE is currently developing its own certification programme to certify adherence to the ethics standards it is developing. ISO does not do certification itself, but third-party certification organisations could in the future assess compliance to ISO ethics-related standards for AI.

Education, training and awareness raising

Education is a powerful method for stimulating ethical behaviour in relation to AI & robotics. In professional and academic education, specifically, education that concerns ethical and social issues in

AI & robotics would benefit future professionals, especially those in the AI & robotics field, but also those in other fields who may deploy and use these technologies in the future. Given the seriousness of ethical issues in the AI & robotics fields, a required ethics course for AI and robotics students seems advisable. Such a course could cover key ethical issues in AI & robotics, ethical guidelines and their application, responsibilities of AI and robotics professionals, and relevant standards, laws, policies, and approaches for ethical AI & robotics. Methodologies for ethics by design could be part of such a course, but for these to be used by future professionals in actual design practice, it might be better if these were to be incorporated in the standard design methodologies used in these fields.

Most professionals who develop and use AI & robotics did not have ethics education in these areas in their professional education. For them, continuing education programmes that include ethics of AI and/or robotics would be valuable. Such training programmes could even be accompanied by professional certification, for example, certification in ethics by design methodology, algorithmic bias avoidance, preparing for ethics review, or all-round ethical practice in AI or robotics. Next to external organisations setting up such training and education programmes, organisations could of course also organize their own in-house training in ethics for AI & robotics.

Turning now from educational institutions to the media, we should acknowledge that media organisations have a large role in generating public awareness and understanding of AI & robotics, including the ethical issues raised by them. These are complicated technologies that are difficult to understand for the average person. Since they are expected to have major impacts on people's lives, a proper understanding of them and the ethical issues they raise is important, and media companies are the most important type of organization who can provide such an understanding to the general public. Therefore, relevant media stories on AI & robotics and its social and ethical dimensions, whether in print, podcast, television or other formats, are important. While media organisations have a major responsibility here, AI & robotics developers also have a responsibility to communicate with the public about these issues, and governments in ensuring that sufficient information is provided.

Policy and regulation

While policy can be made by any kind of organization, our concern here is with public policy, as made by governments, as well as the laws and regulations issued by them. The key question here is: what policies, laws and regulations should governments develop, if any, to stimulate the ethical development, deployment and use of AI & robotics? Policies, laws and regulations can relate to ethical criteria in three ways: they can explicitly institute, promote or require ethics guidelines, procedures, or bodies; they can have a focus on upholding certain moral values or principles without explicitly identifying them as ethical (e.g., well-being, privacy, fairness, sustainability, civil rights); and they either explicitly or implicitly take on board ethical considerations in broader social and economic policies.

Governments are currently at a decision point for AI & robotics policy. What should they do, and how can they avoid regulating too little as well as regulating too much? Decisions that relate to ethics include the following:

- Whether or not to issue, or support the issuing of, ethical guidelines for AI & robotics
- Whether or not to put any ethical guidelines for AI & robotics into law
- Whether or not to revise existing institutional structures to better account for ethical issues or to create new governmental bodies or unites for ethical and social issues in AI & robotics
- Whether or not to mandate ethics standards, certification, education, training, ethical impact assessments or ethics by design methods in relation to ethics of AI & robotics
- Whether and how to introduce new legislation and regulations to for morally controversial AI & robotics technologies, such as automated tracking, profiling and identification technologies, behaviour and affect recognition technologies, and automated lethal weapons
- How to include ethical considerations concerning AI & robotics in policies, laws and regulations, both ones that pertain to AI & robotics specifically and more general ones that need to be updated to account for AI & robotics, such as in the areas of consumer protection, data protection, criminal law, non-discrimination provisions, civil liability and accountability
- What financial support and funding to provide, if any, for ethics research, ethics education, ethics dialogue, ethics awareness raising and other ethics initiatives in relation to AI & robotics
- How to regulate the government's own use of AI & robotics so as to ensure ethical conduct

See also the forthcoming SIENNA report D5.6, *Recommendations for the enhancement of the existing EU and international legal framework*, which will contain our proposals for new EU and international legislation and regulations to support ethical AI & robotics.

Finally, a general remark regarding these methods: it remains to be seen whether ethical AI & robotics are best served by specific ethics standards, certification, design methodologies, audits, policies and other methods, or whether it is better to integrate ethics concerns into broader standards, policies, audits, etc. This probably varies from situation to situation, but should receive proper attention as an issue to account for.

Making methods available and motivating actors

In the preceding discussion of methods, we have already made a number of suggestions regarding the responsibility of different actors for developing and making available different types of methods. Obviously, governments are the responsible party for the development governmental policies, laws and regulations, and universities are the ones that would development of ethics courses in degree programmes in AI and robotics. In other cases, it may not be immediately obvious which actor would be responsible for developing and advocating for a particular method. Which actor would be responsible for developing methods of ethical impact assessment, for example, or for developing operational ethics guidelines for the deployment and use of AI in organisations? Often, this is a matter of particular actors stepping up and taking on such responsibilities. It was not written in stone that the IEEE should embark on in an extensive programme to develop ethical guidelines, methods, standards and certification for the design and deployment of AI and robotics systems, but it nevertheless chose to do so.

On the other hand, actors may fail to step up, leaving a responsibility vacuum in society due to which important methods for ethical AI & robotics are not being developed and implemented. If this is to occur, then governments are often seen as the responsible actor to step in and enact policies, laws and regulations that help fill this vacuum. Governments, after all, have a particular responsibility for promoting the public good, protecting individual rights, and supporting fair socioeconomic conditions, and also have powers to stimulate and compel other actors to act responsibly and in the public interest.

3. A Framework for Ethics by Design

3.1 Introduction and general framework

In this part of the report, we will present a framework for Ethics by Design. The aim of this framework is to allow AI and robotics developers to include ethical requirements in a systematic and comprehensive manner in the development process. Our framework builds on an earlier framework developed within the SHERPA project (Brey, Lundgren, Macnish and Ryan, 2019). This earlier proposal is one we still stand behind, but we feel it leans too much on one particular development approach in AI, the CRISP-DM approach, and also it does not cover robotics. Here, we present a more general approach for Ethics by Design, which does not depend on a specific development approach, and we present three more specific approaches, that could be used by developers working with the specific methodologies on which they are based: CRISP-DM, Agile and the V-Model. CRISP-DM and Agile are intended for AI, and the V-Model mainly for robotics. Ethics by Design using Agile and the V-Model will be discussed in this report. For a discussion of Ethics by Design using CRISP-DM, see (Brey, Lundgren, Macnish and Ryan, 2019).

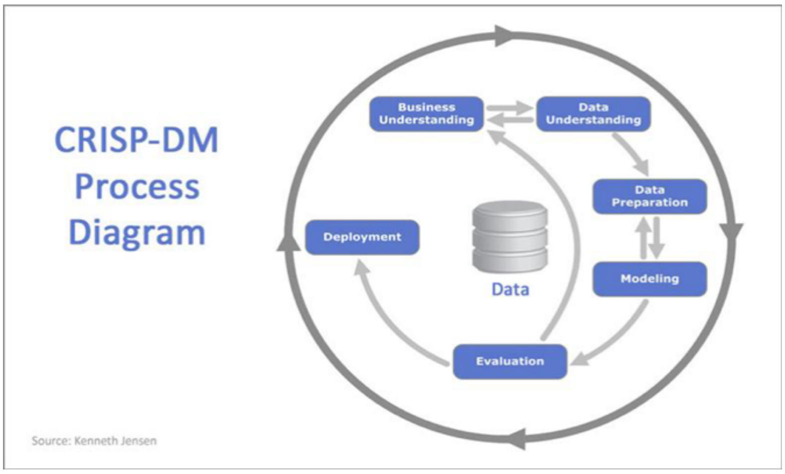


Figure 1 CRISP-DM process

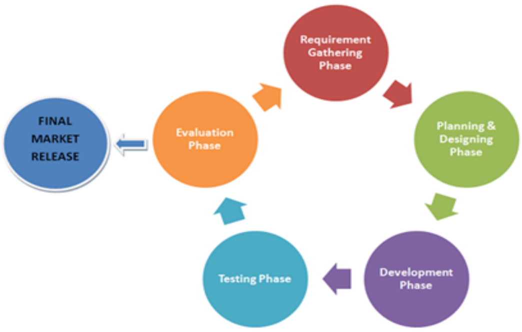


Figure 2: Agile Methodology cycle

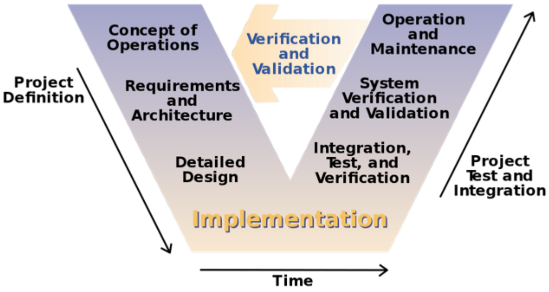

Project Definition
Concept of Operations
Requirements and Architecture
Detailed Design
Verification and Validation
Implementation
Operation and Maintenance
System Verification and Validation
Integration, Test, and Verification
Project Test and Integration
Time

Figure 3: Graphical representation of a typical V-Model approach to systems engineering

In our framework, we distinguish between high-level, intermediate level, operational, and specific operational ethics 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, such as the ones proposed by the HLEG-AI, OECD and IEEE. Intermediate-level guidelines are more specific, providing more concrete conditions that must be fulfilled. The HLEG-AI, for example, breaks its ethics guidelines down into three to five sub-requirements that are at this intermediate level. 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 specific operational guidelines for the development of AI and robotic systems.

High-level requirement

We will first briefly describe the high-level requirements we make use of, to provide an insight into the fundamental principles and values behind the specific requirements. Readers who are familiar with the SHERPA "Guidelines for the Ethical Development of AI and Big Data Systems: An Ethics by Design approach" (in Brey, Lundgren, Macnish and Ryan, 2019) will notice that our high-level requirements are identical to theirs with the exception of the removal of safety and robustness, which was removed because we feel it is better covered by standard safety and resilience engineering rather than by ethical guidelines. The SHERPA high-level requirements are in turn based directly on the High-Level Expert Group on AI's high-level requirement in their "Ethics Guidelines for Trustworthy AI", with some minor changes intended to improve their coherence and fitness for operationalization.

SIENNA High-level requirements and sub-requirements
1 Human agency, liberty and dignity: Positive liberty, negative liberty and human dignity

2 Privacy and data governance: Including respect for privacy, quality and integrity of data, access to data, data rights and ownership
3 Transparency: Including traceability, explainability and communication
4 Diversity, non-discrimination and fairness: Avoidance and reduction of bias, ensuring fairness and avoidance of discrimination, and inclusive stakeholder engagement
5 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
6 Accountability: auditability, minimization and reporting of negative impact, internal and external governance frameworks, redress, and human oversight

Table 3: SIENNA High-level requirements

General approach

The general approach to Ethics by Design that we propose here builds directly on the CRISP-DM inspired approach that was presented in Brey, Lundgren, Macnish and Ryan (2019), as well as on the Agile and V-Model inspired approaches presented later. It makes the assumption that whatever specific design approach is used for AI and robotics systems, there are some shared practices or "phases" that can be generically described and that can then be accompanied with guidelines for the incorporation of ethical considerations. We assume that development processes for AI & robotics systems involve most or all of the following practices:

1. *Specification of objectives.* This is the practice of determining what the system that is to be developed is intended for, and therefore should be capable of doing. It is often the very first step in design: formulating a general goal for the design process. It often takes place in close interaction with a customer. Examples of such goals are the goal of developing a system that is capable of recognizing faces from live video feeds with great accuracy, or of a robot that is capable of extinguishing brush fires.

2. *Specification of requirements.* This is the practice in which the kind of system that is needed is clarified, and technical and non-technical requirements and constraints are identified and formulated. This often results in a requirement list for the system to be developed. This practice could also involve an initial determination of needed and available resources, and an initial risk assessment and, as part of it, a cost-benefit analysis.

3. *High-level design.* This is the development of a high-level architecture that meets the requirements. It is sometimes preceded by the development of a conceptual model.

4. *Data collection and preparation.* For systems that involve a lot of data processing, a process will be involved of collecting, verifying, selecting, cleaning, construction, integration and formatting data.

5. *Detailed design and development.* This involves the detailed design and development of a full working system. For software development, this will involve detailed programming and coding. For hardware systems (i.e., robots), this will also involve a manufacturing phase.

6. *Testing and evaluation.* This is the process of testing and validation of a system, and its evaluation against the original objectives and requirements.

We recommend that ethical considerations are included in these six practices in the following ways:

Specification of objectives

Requirement 1: Ethical assessment of objectives

The objectives that are specified in this practice are to be evaluated against the seven ethics requirements that were described in the previous section (as well as more detailed ethics requirements that apply to particular types of systems or applications). Sometimes there can be a basic incompatibility between the objectives of a system and relevant ethics requirements. For example, the objective may be to engage in covert surveillance of people (violating principles of privacy and autonomy), or to engage in politically driven censorship of news feeds (violating principles of freedom of information and societal wellbeing (democracy)). Possible outcomes of this assessment are:

1. The objectives are compatible with the ethics requirements. Proceed to next step.

2. The business objectives are inherently incompatible with ethics requirements. The development of the system should be terminated.

3. The business objectives are incompatible with ethics requirements, but modifications of the business objectives are possible to ensure compatibility. Modify business objectives and proceed to next step.

4. It is unclear whether business objectives are compatible with ethics requirements. Cautiously proceed to the next step, and keep monitoring closely.

It is also an option at this step to ask stakeholders to be involved in the determination of objectives.

Specification of requirements

Requirement 2: Ethical assessment of resources, requirements and constraints
During the specification phase, test the inventory of resources and other requirements and constraints against ethics requirements for possible tensions. E.g., it may be found that the requirements of transparency and accountability cannot be met with available resources for the established business objectives. Make modifications to resources and to other requirements and constraints to reduce tensions with ethics requirements. Ensure that ethics requirements are included in the final list of requirements. Optionally, stakeholders can be included in this process.

Requirement 3: Expanded risk assessment
It could be considered at this stage to do an ethical risk assessment or ethical impact assessment to assess possible ethical issues or risks that might follow from the system being developed according to the objectives and requirements list. Optionally, stakeholders can be included in this process.

High-level design

Requirement 4: Ethical assessment of high-level design
To integrate ethical requirements into this process, ensure that ethical guidelines are considered, and that the design is evaluated relative to these ethical guidelines. Issues that may be particularly relevant in this design phase are those relating to transparency, autonomy, privacy and fairness.

Data collection and preparation

Requirement 5: Ethical data collection
Data collection involves the collection of initial data, its description and initial analysis, and verification of quality. To integrate ethical requirements into this process, assess how different steps in the process might support or violate ethical requirements. Make necessary changes as a result. (If appropriate changes are not possible to perform, the design objectives may need to be re-evaluated). Follow the four-way choice process established in Requirement 1. In this process, fairness (including bias, discrimination, and diversity), privacy and data quality will be particularly important.

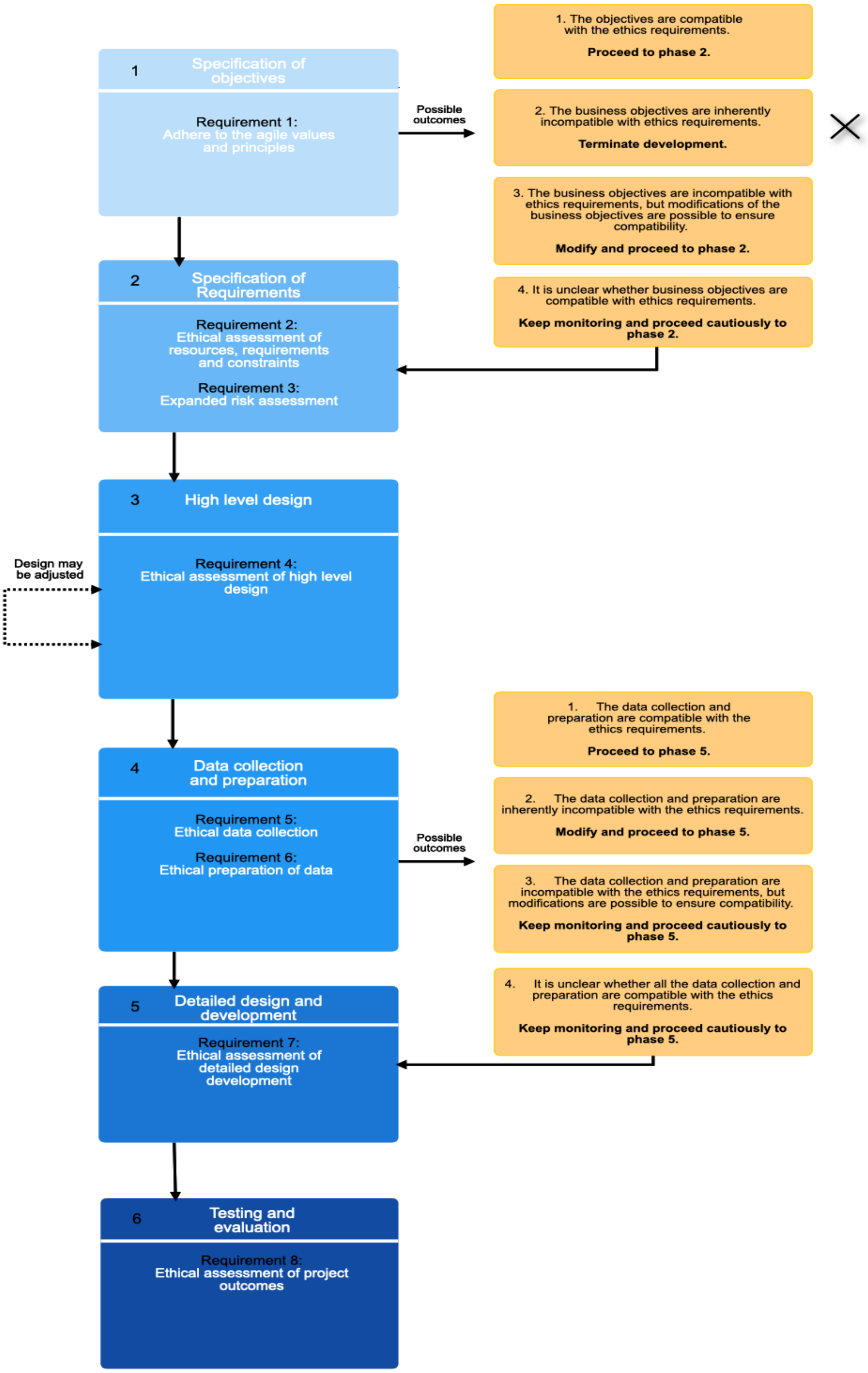

1
Specification of objectives
Requirement 1:
Adhere to the agile values and principles
Possible outcomes
1. The objectives are compatible with the ethics requirements.
Proceed to phase 2.
2. The business objectives are inherently incompatible with ethics requirements.
Terminate development.
3. The business objectives are incompatible with ethics requirements, but modifications of the business objectives are possible to ensure compatibility.
Modify and proceed to phase 2.
4. It is unclear whether business objectives are compatible with ethics requirements.
Keep monitoring and proceed cautiously to phase 2.
2
Specification of Requirements
Requirement 2:
Ethical assessment of resources, requirements and constraints
Requirement 3:
Expanded risk assessment
3
High level design
Requirement 4:
Ethical assessment of high level design
Design may be adjusted
4
Data collection and preparation
Requirement 5:
Ethical data collection
Requirement 6:
Ethical preparation of data
Possible outcomes
1. The data collection and preparation are compatible with the ethics requirements.
Proceed to phase 5.
2. The data collection and preparation are inherently incompatible with the ethics requirements.
Modify and proceed to phase 5.
3. The data collection and preparation are incompatible with the ethics requirements, but modifications are possible to ensure compatibility.
Keep monitoring and proceed cautiously to phase 5.
4. It is unclear whether all the data collection and preparation are compatible with the ethics requirements.
Keep monitoring and proceed cautiously to phase 5.
5
Detailed design and development
Requirement 7:
Ethical assessment of detailed design development
6
Testing and evaluation
Requirement 8:
Ethical assessment of project outcomes

Fig. 4

Requirement 6: Ethical preparation of data
Preparation of data involves selection of data for inclusion in the system, cleaning of data, construction of new data and data records on the basis of existing data and records, and formatting of data. assess how different steps in the process might support or violate ethical requirements. Make necessary changes as a result. (If appropriate changes are not possible to perform, the design objectives may need to be re-evaluated). Follow the four-way choice process established in Requirement 1. In this process, fairness (including bias, discrimination, and diversity), privacy and data quality will be particularly important.

Detailed design and development

Requirement 7: Ethical assessment of detailed design and development
This requirement is similar to requirement 4 for high-level design. To integrate ethical requirements into this process, ensure that ethical guidelines are considered, and that the design is evaluated relative to these ethical guidelines. Issues that may be particularly relevant in this design are those relating to transparency, autonomy, privacy and fairness.

Testing and evaluation

Requirement 8: Ethical assessment of project outcomes
As part of the testing and evaluation phase, an ethical assessment should be performed of the results. Possible outcomes are that ethical issues have been dealt with in a satisfactory way, that further development is needed, or that specific guidance for or restrictions on deployment and use need to be in place to mitigate ethical issues. It is recommended that stakeholder consultation or involvement takes place during this phase.

Figure 4 provides a flowchart for the general Ethics by Design approach with the different ethics requirements in place.

3.2 Agile

Agile software development emerged during the 1990s as a response to the traditional plan-driven ('waterfall') approach. The plan-driven approach has several phases (i.e. Requirements, Design, Implementation, Verification and Maintenance) that are not returned to once they are finished. The field of software development, however, is unpredictable and changes rapidly and it soon became clear that such a linear, sequential approach works contradictory with these quick changes. This gave rise to Agile software development. Agile incorporates the traditional plan-driven phases[5], but instead of being sequential it is a cyclical process *(Figure 5).* The term Agile thus stems from its incremental characteristic.

[5] Note that the phases are sometimes called differently (e.g., requirements, development, testing, delivery, and feedback; see https://www.smartsheet.com/understanding-agile-software-development-lifecycle-and-process-workflow). While slightly different, the phases do entail similar features and elements. In what follows, the different phases will be referred to as 'Requirement Gathering', 'Planning & Designing', Development, 'Testing', and 'Evaluation'.

Furthermore, whereas a plan-driven approach has a set goal from the beginning, Agile development only sets a preliminary goal. The flexibility of the cycles and the preliminary goal facilitates to adapt according to the client's wishes, or to potential drawbacks in the development procedure. Thanks to Agile's flexibility, correcting mistakes and adjusting to changing needs of the client is easier, making it hence less risky and more cost-effective as opposed to the plan-driven approach. Following the Agile philosophy, it is highly unlikely that a product will have been developed that is at odds with the client's wishes, as after every cycle the client is able to evaluate the product, which in addition increases transparency for the client (Gonçalves, 2019).

1. *Customer satisfaction through early and continuous delivery of valuable software.*
2. *Welcome changing requirements, even late in development.*
3. *Deliver working software frequently (weeks rather than months).*
4. *Daily cooperation between business people and developers is required.*
5. *Projects are built around motivated individuals, who should be trusted.*
6. *Face-to-face conversation is the most efficient and effective method of communication.*
7. *Working software is the primary measure of progress.*
8. *Ability to maintain a constant pace for sponsors, developers, and users through sustainable development.*
9. *Continuous attention to technical excellence and good design enhances agility.*
10. *Simplicity—the art of maximizing the amount of work not done—is essential.*
11. *The best architectures, requirements, and designs emerge from self-organizing teams.*
12. *Regularly, the team reflects on how to become more effective, and adjusts accordingly.*

Agile development is not one methodology, rather it is an umbrella term for a collection of approaches that adhere to the Agile mindset. The more known of such approaches include Scrum, XP (experience programming) and Kanban. Nevertheless, these approaches also do not have one set methodology and are differ per organization and project (Fitzgerald, Hartnett & Conboy 2006; R. Duchoba, personal communication, February 24, 2020). Thus, instead of a clear methodology, 'being Agile' implies prioritizing certain values, namely individuals and interactions, working software, customer collaboration, and responding to change over processes and tools, comprehensive documentation, contract negotiation, and following a plan (Beck et al., 2001).

Key to an Agile development process is the idea of a horizontal organization rather than a top-down structure. It is emphasized that without adhering to these four core values, Agile is most likely to fail[6]. In addition to these values, there are twelve principles that may guide Agile software development (Beck et al., 2001):

[6] See e.g., https://blog.confirm.ch/when-agile-fails-hierarchy-and-roles/

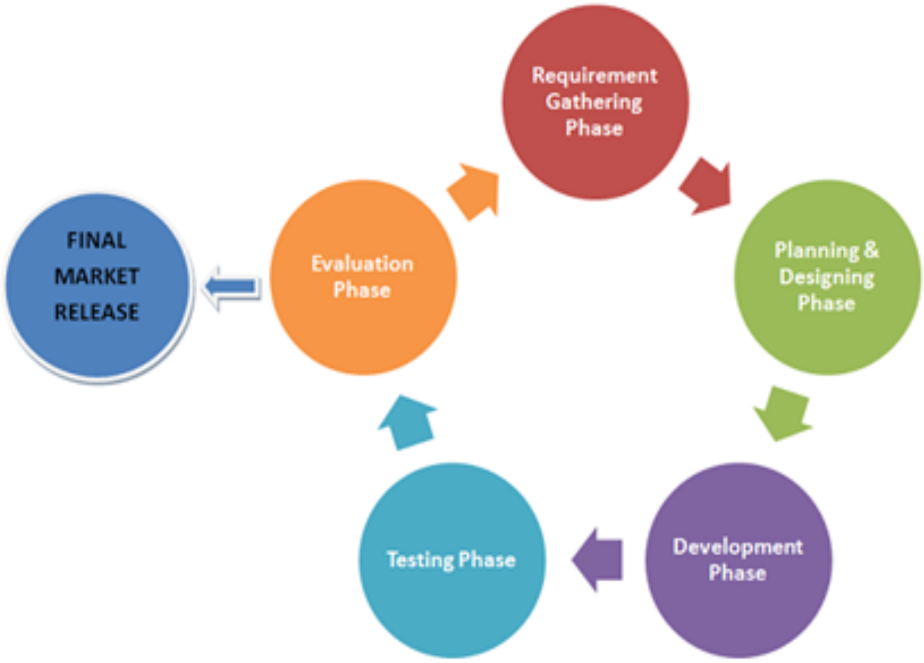


Figure 5: Agile Methodology cycle

Incorporating ethics

Before elaborating on ethical guidelines for Agile software development, three caveats should be mentioned. Firstly, it should be emphasized that the values and principles of Agile suggest that Agile is in itself already a normative framework. Incorporating ethics is therefore ultimately done by adhering to the Agile mindset, as from this mindset many other values follow (such as stakeholder inclusion, transparency, etc). In what follows, the proposed ethical requirements serve as a guide for software developers during the development process. The aim is that these requirements can make developers more aware of ethical considerations during the development process on top of the values embedded within the Agile philosophy.

Secondly, in providing ethical guidelines lies in the horizontal structure embedded within the Agile philosophy. Imposing ethical requirements in Agile implementations suggests a top-down hierarchical structure and is therefore at odds with this horizontal level of power. This implies that when following a fully Agile spirit, ethical incorporation into Agile must come from the developer himself instead of being imposed externally (S. Braams, personal communication, December 11, 2019). Nonetheless, decisions made by the developers relating to algorithmic settings may have an influence on society (Huldtgren 2015) and should therefore be well-considered.

Thirdly, due to a lack in clear methodologies there are no all-encompassing ethical guidelines for Agile software development. It cannot be emphasized enough that the way Agile development is done is context-dependent (R. Duchoba, personal communication, February 24, 2020). While *generally* the five phases as represented in *Figure 5* fit most development processes, it does not apply to *all* projects and/or organizations.

The most common way of software development is still a hybrid of waterfall and agile (*The 13th annual State of Agile Report*, 2019; S. Braams, personal communication, December 11, 2019). The waterfall mindset is so deeply rooted in society – and in many cases necessary for big organizations to work efficiently – that Agile may be adopted by a *team* but not by the entire *organization*. The organization then still imposes restrictions on the team. It is for such hybrid situations that these requirements are most applicable.

One key aspect to increase ethical development, as mentioned by an expert in the field, is to have a work environment where the team feels comfortable to communicate their thoughts and concerns. In order to provide such an environment, there needs to a be moment where the team can indeed share their comments; it is suggested that there is someone who can help the team to share their comments. It is hence important that during the design process there are such opportunities and people (e.g., in Scrum this could be the Scrum Master).

As noted, not all projects include a clear separation between the following phases. It is acknowledged that this is a simplified understanding of Agile development, and that in practice there tends to be an interplay between these phases. Annex 1 elaborates in more detail on requirements related to Agile in general or specific to the separate phases.

Applying ethics to the Agile approach

Agile in general

Requirement 1: Adhere to the Agile values and principles

It is important that throughout the entire software development process the Agile mindset is prevailing, therefore it is important that the following values are adhered to. These values in turn result in other ethical requirements, such as stakeholder inclusion, flexibility, and transparency within the team.

- *Individuals and interactions*
- *Working software*
- *Customer collaboration*
- *Responding to change*

Requirements Gathering phase

Requirement 2: Inclusion of ethical requirements and ethical assessment of business objectives

The Requirements Gathering phase assesses requirements for the end-product according to the desires and needs of the client. During this phase, it is useful to consider what ethical values may potentially be at stake in order to adjust the design process to insure these values with respect to the final product. The client's business objectives are assessed to see if they are compatible with the high-level ethical requirements. For example, an objective such as 'surveillance of people' results in a structural infringement on one's privacy. It is recommended to include Annex 1 in the assessment. If the client adapts his/her wishes based on a previous cycle, the new requirements should be assessed similarly. The assessment may result in several outcomes:

1. The business objectives are compatible with the ethics requirements. Proceed to next step.

2. The business objectives are inherently incompatible with ethics requirements. The development of the system should be terminated.

3. The business objectives are incompatible with ethics requirements, but modifications of the business objectives are possible to ensure compatibility. Modify business objectives and proceed to next step.

4. It is unclear whether business objectives are compatible with ethics requirements. Cautiously proceed to the next step, and keep monitoring closely.

Planning & Designing phase

The Planning & Designing phase concerns the design of the final product. Note that the design is an emergent process. Depending on changing needs of the client or obstacles during the development process, the final product is only a preliminary estimate and changes throughout the process. Design is therefore dependent on the evolution of the product. Nevertheless, a good plan and design increases the efficiency of the project. Furthermore, it may be necessary during each cycle to adjust the design (cycle within a cycle) due to complications during the development. The Planning & Designing and Development phases are closely linked, as the former may require coding to check whether the design indeed works[7].

Requirement 3a: Appropriate task distribution

It is important that the plans made are feasible in practice. The project should therefore be managed in an appropriate way. For example, the Product Owner (PO) in Scrum should have enough experience and understanding that he can estimate technical limitations of the design and of the clients' requirements. If the PO does so correctly, s/he can estimate the time needed for certain tasks which in turn allows the PO to assemble an achievable backlog for the iteration.

Requirement 3b: Transparency concerning task distribution

To increase transparency and trust among the team, each member's task should be displayed in the project room. This increases accountability among the team members and increases the possibility to adjust a mistake when the origin of the error is clear (note that some errors have an unclear origin).

Requirement 3c: Ethical assessment of technical requirements, methods and models

In addition to checking whether the client's business objectives are compatible with the HLEG requirements, it is necessary to assess whether there are technical limitations embedded within the intended system that may give reason for concern. This includes the model and methods used planned to design the system. This may result in the following outcomes:

[7] See e.g., http://www.agilemodeling.com/essays/agileDesign.htm, specifically #7, #8, #11, #12.

1. All of the technical requirements are compatible with the ethics requirements. One can proceed to next step.

2. Some of the technical requirements are incompatible with the ethics requirements, but modifications to these requirements are possible to ensure compatibility. One should modify requirements and proceed to next step.

3. It is unclear whether all the technical requirements are compatible with the ethics requirements. One should cautiously proceed to the next step, and keep monitoring closely.

Development phase

Requirement 4a: Ethical data integration and usage

The Development phase uses data, hence all issues related to data should be carefully considered. During this part of the development process, issues such as bias, discrimination, fairness and diversity, privacy, and data quality are of particular importance when constructing and using the data set.

Requirement 4b: Incorporating ethical values

During the development of the system, it is important that ethical values are kept in mind while developing the system. Small programming decisions may have a societal impact, and it is thus important that such decisions are well-considered.

Requirement 4c: Peer Review of Code

While there is little documentation in Agile, this is partly because the code *is* the documentation. This means that the code should be well-explained by the coder and should be understandable for other coders. If a bug appears in the code, this should be mendable not only by the original programmer of the code, but also by others. A peer review of the code is then essential as this indicates whether the code is well written and explainable.

Testing phase

The Testing phase is important as it allows to discover potential defects in the product. This phase may be concerned with the violation of ethical values. The team is at liberty in this phase to assess potential arising risks. The Testing phase is closely linked with both the Development phase as well as the Evaluation phase. The Testing phase is however more limited to the development team. Here, the product is less likely to already be evaluated by the client.

Requirement 5: Ethical Testing

During testing, make sure that the testing is done appropriately so that certain problematic consequences are avoided. Perhaps the system works well for one racial group or gender and is less effective/biased towards another. In situations where indeed gender, race, etc. play a role, make sure all groups are assessed in equal manners to avoid discrimination.

Evaluation phase

Requirement 6a: Inclusion of an ethical checklist

The Evaluation phase is focused on evaluating whether the project is meets the requirements of the client. Not only could this phase focus on technical quality of the product, but it could include a checklist (e.g., definition of done in Scrum) to secure that certain ethical values are adhered to. Although ethics should be embedded throughout the entire development process, a checklist after each iteration ensures that the team does not neglect certain issues. In cases where some ethical problems are overlooked, this can then be rectified in the next iteration. Note that Agile development has the potential to lead to "haphazard and harmful creations that are flung into the world before their potential impacts are assessed" (Alix, 2017). It is thus of utmost important that the Evaluation phase is done in a secure and proper fashion, as from this phase a product may potentially be sent off to the final market release. Several outcomes are possible:

1. The product(s) adhere fully to the ethics requirements.

2. Some elements of the product(s) do not adhere to the ethics requirements, but modifications to the design are a practical solution that will ensure adherence to the ethics requirements. One should modify the design and product in the next iteration.

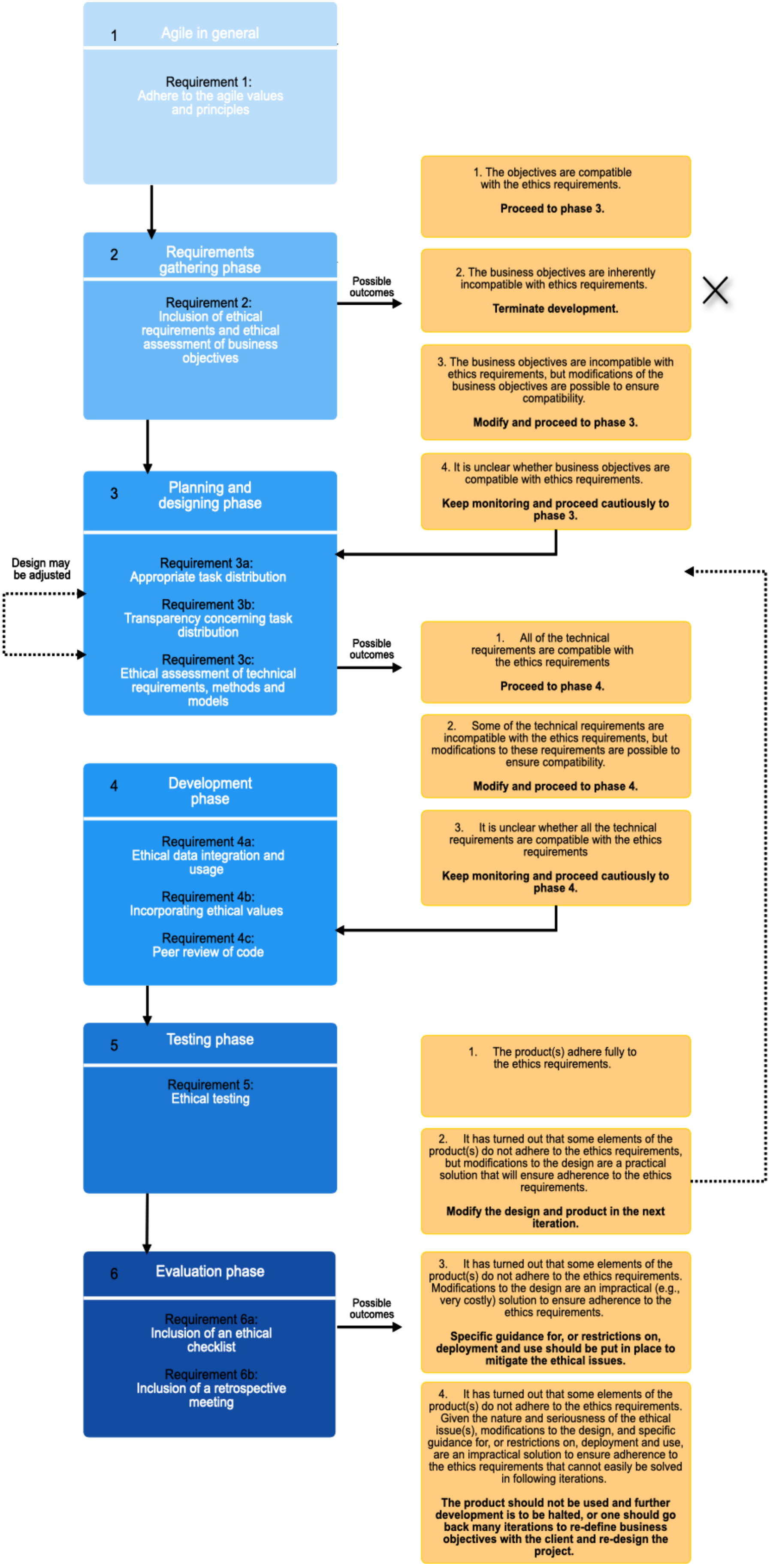
3. Some elements of the product(s) do not adhere to the ethics requirements. Modifications to the design are an impractical (e.g., very costly) solution to ensure adherence to the ethics requirements. However, specific guidance for, or restrictions on, deployment and use can be put in place to mitigate the ethical issues. One should take such measures.

4. Some elements of the product(s) do not adhere to the ethics requirements. Given the nature and seriousness of the ethical issue(s), modifications to the design, and specific guidance for, or restrictions on, deployment and use, are an impractical solution to ensure adherence to the ethics requirements that cannot easily be solved in following iterations. The product should not be used and further development is to be halted, or one should go back many iterations to re-define business objectives with the client and re-design the project. In case when this is not the first iteration, this option is unlikely to occur thanks to Agile's flexible character. Nevertheless, ethical concerns may have been overlooked in previous iterations.

Requirement 6b: Inclusion of a retrospective meeting

All Agile approaches should integrate a retrospective meeting (or something similar) in their evaluation phase. Such a meeting allows for deeper discussions on what went right/wrong in the iteration. Whereas technical issues are discussed on the spot, when they occur, ethical issues tend to be addressed in such longer meetings (R. Duchoba, pers. comm., February 24, 2020). These meetings thus provide a good place and moment for an ethical discussion, and are therefore strongly recommended.

Figure 6 contains a complete flowchart for the Ethics by Design approach for Agile.



1 Agile in general

Requirement 1:
Adhere to the agile values and principles

2 Requirements gathering phase

Requirement 2:
Inclusion of ethical requirements and ethical assessment of business objectives

Possible outcomes

1. The objectives are compatible with the ethics requirements.

Proceed to phase 3.

2. The business objectives are inherently incompatible with ethics requirements.

Terminate development.

3. The business objectives are incompatible with ethics requirements, but modifications of the business objectives are possible to ensure compatibility.

Modify and proceed to phase 3.

4. It is unclear whether business objectives are compatible with ethics requirements.

Keep monitoring and proceed cautiously to phase 3.

3 Planning and designing phase

Design may be adjusted

Requirement 3a:
Appropriate task distribution

Requirement 3b:
Transparency concerning task distribution

Requirement 3c:
Ethical assessment of technical requirements, methods and models

Possible outcomes

1. All of the technical requirements are compatible with the ethics requirements

Proceed to phase 4.

2. Some of the technical requirements are incompatible with the ethics requirements, but modifications to these requirements are possible to ensure compatibility.

Modify and proceed to phase 4.

3. It is unclear whether all the technical requirements are compatible with the ethics requirements

Keep monitoring and proceed cautiously to phase 4.

4 Development phase

Requirement 4a:
Ethical data integration and usage

Requirement 4b:
Incorporating ethical values

Requirement 4c:
Peer review of code

5 Testing phase

Requirement 5:
Ethical testing

6 Evaluation phase

Requirement 6a:
Inclusion of an ethical checklist

Requirement 6b:
Inclusion of a retrospective meeting

Possible outcomes

1. The product(s) adhere fully to the ethics requirements.

2. It has turned out that some elements of the product(s) do not adhere to the ethics requirements, but modifications to the design are a practical solution that will ensure adherence to the ethics requirements.

Modify the design and product in the next iteration.

3. It has turned out that some elements of the product(s) do not adhere to the ethics requirements. Modifications to the design are an impractical (e.g., very costly) solution to ensure adherence to the ethics requirements.

Specific guidance for, or restrictions on, deployment and use should be put in place to mitigate the ethical issues.

4. It has turned out that some elements of the product(s) do not adhere to the ethics requirements. Given the nature and seriousness of the ethical issue(s), modifications to the design, and specific guidance for, or restrictions on, deployment and use, are an impractical solution to ensure adherence to the ethics requirements that cannot easily be solved in following iterations.

The product should not be used and further development is to be halted, or one should go back many iterations to re-define business objectives with the client and re-design the project.

Fig. 6.

3.3 The V-Model

Description of the approach

The V-Model consists of a number of developmental phases, which each yield a number of predefined products that constitute input for subsequent phases. Grouped, these phases can be graphically presented in the form of a "V" (see *Figure 2*). In this "V", phases on the left side represent the decomposition of requirements and development of system specifications ("Project Definition" in *Figure 2*), and phases on the right side represent the integration of parts and their verification and validation ("Project Test and Integration" in *Figure 2*).

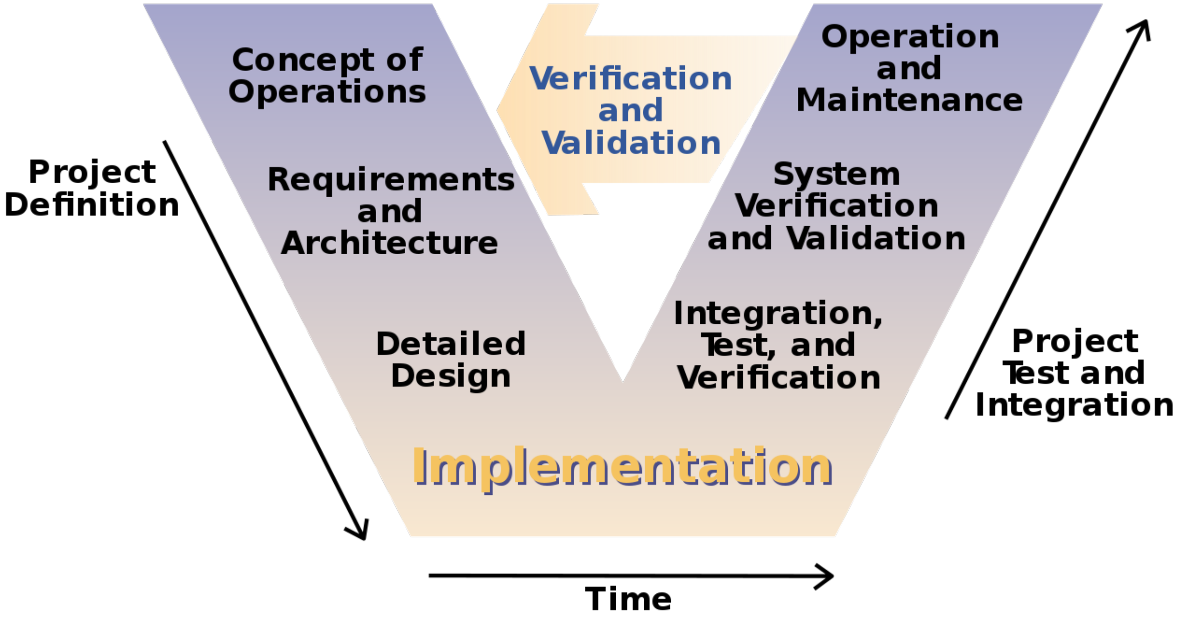

Concept of Operations
Verification and Validation
Operation and Maintenance
Project Definition
Requirements and Architecture
System Verification and Validation
Detailed Design
Integration, Test, and Verification
Project Test and Integration
Implementation
Time

Figure 7: Graphical representation of a typical V-Model approach to systems engineering[8]

Let us consider each of the phases of a typical V-Model in a bit more detail. According to the V-Model, any robotics development project begins with the creation of high-level descriptions of what is expected of the system under development ("Concept of Operations" in *Figure 2*). During this phase, there is ample discussion with the client(s) about (their needs and wishes regarding) the operational benefits and (development, deployment and operational) costs of the final system. Moving down the left side of the "V", the high-level descriptions get evermore concrete, as design requirements and specifications are formulated and a detailed system design is drawn up. So, during the "Requirements and Architecture" phase, functional and non-functional requirements are formulated for the system to be accepted by the client(s), as well as specifications, in terms of quantitative values, for the to-be-

[8] Public domain image extracted from Wikipedia.com at https://en.wikipedia.org/wiki/V-Model#/media/File:Systems_Engineering_Process_II.svg on January 21, 2020 (Author: Wikipedia user "Slashme"). Redrawn from original in: Leon Osborne, Jeffrey Brummond, Robert Hart, Mohsen (Moe) Zarean Ph.D., P.E, Steven Conger. *Clarus Concept of Operations.* Publication No. FHWA-JPO-05-072, Federal Highway Administration (FHWA), 2005.

developed hardware and software that will fulfil these requirements. In addition, a system architecture is drawn up, that conveys the fundamental organization of the system in terms of the characteristics of its major components, as well as their relationships to one another and their environment. Subsequently, during the "Detailed Design" phase, the entire system, down to its individual components and their interrelations, is designed on the basis of the architecture and specifications. At the bottom of the "V" (in *Figure 2*), the system design is then implemented, meaning that all of its individual hardware and software components are constructed.

After implementation, a process of integration and testing is started. The phases in this process move up along the right side of the "V". First, there is a phase of "Integration, Test and Verification", where the system's constructed components are integrated, and where it is verified whether the individual components' functional properties and their interoperation are in accordance with the system design as defined during the "Detailed Design" phase. Testing can be done by the developers or by specialists. Subsequently, during the "System Verification" phase, it is investigated whether the system as a whole meets the conditions set out during the "Requirements and Specifications" phase. Finally, during the "Operation and Maintenance" phase it is evaluated whether the entire system performs in accordance with the initial high-level descriptions of it and the client(s)'s expectations during real-world operation.

Importantly, during each of the integration and testing phases, there is a feedback loop to a corresponding phase on the left side of the "V" ("Project Definition"). Taken together, these feedback loops are the "Verification and Validation" part of the model, where verification means testing against the technical requirements and specifications, and validation means testing in the real world and against the user(s)'s needs. If it turns out during integration and testing that a component is not functioning the way it should, developers are advised to retrace their steps and rebuild the component so that its performance is in accordance with the requirements, or (if this is not possible) adjust the expectations, requirements and/or detailed design of the component or the system in question.

Stated advantages of using the V-Model approach

There are a number of advantages associated with the use of the V-Model in robotics development. First, the approach puts significant emphasis on product verification and validation, meaning that the final product of any given development project is likely to meet the requirements set out at the beginning, and the expectations of the client.

Second, by offering a structured and fairly regimented development process with distinct phases and clear instructions, recommendations and detailed explanations for each of these phases, the V-Model helps to bring large, complicated robotics development projects to a successful end.[9]

Limitations and criticisms of the V-Model approach

In spite of the above-mentioned advantages, the V-Model has a number of limitations and has faced criticism from Agile proponents (mainly in software engineering) and others.[10] The limitations and criticisms include the following:

[9] Interview with anonymous university-appointed robotics expert conducted in the Netherlands on January 20, 2020.

[10] See Kneuper (2018), Liversidge (2015) and Wellens (2008).

- The V-Model might be too linear an approach in that it does not easily allow one to go back and forth between different developmental phases in situations where such action can be helpful. It does not deal well with changing circumstances (e.g., when the client's wishes have changed), on the basis of which earlier decisions might have to be revisited. This issue has mainly been raised in relation to software engineering.
- The V-Model, in its most general, abstract form, verifies and validates only the physical product that is being constructed, not the underlying requirements and design of the product, which may also be prone to errors. Fixing any errors in the project definition early on might save a lot of trouble later.
- The V-Model might lead to verification and validation procedures being pressed into tight windows at the very end of a development project, when earlier phases have exceeded their scheduled times, and while the project's finalisation date remains fixed.
- The V-Model might lead to inefficient and ineffective verification and validation procedures by advocating regimented standard testing procedures and not supporting ad hoc exploratory testing. This issue has mainly been raised in relation to software engineering.

Notwithstanding these limitations and criticisms, the V-Model remains the most commonly used general approach in robotics development.

Applying ethics to the V-Model approach

In this section, we describe how ethics can be integrated into the V-Model approach for its use in robotics engineering. We lay out the basic procedures that should be followed by developers during each of the developmental phases of the V-Model approach (see *Figure 2*). This section references a more detailed list of ethical requirements that are provided in Annex 2 of this report.

Concept of operations phase

Requirement 1a: Ethics check of business objectives

To integrate ethics into the concept of operations phase, it is necessary to test the business objectives formulated during this phase against the seven high-level ethics requirements. The aim is to establish whether there are any tensions between these business objectives and ethics requirements. In addition, it is recommended to test the business objectives against the operationalised requirements in Annex 2, which also detail a number of special issues that are particular to the development of robotic systems.

Sometimes, the objectives of a project and ethics requirements are incompatible with one another. For instance, an objective of a particular robotics project may be to engage in covert surveillance of people, which would violate principles of privacy and autonomy. Given such situations, there are a number of possible outcomes of the ethics check and ways to proceed:

1. All of the business objectives are compatible with the ethics requirements. One can proceed to next step.

2. Some of the business objectives are incompatible with the ethics requirements, and modifications to these business objectives are not possible. The development of the system should be terminated.

3. Some of the business objectives are incompatible with the ethics requirements, but modifications to these business objectives are possible to ensure compatibility. One should modify business objectives and proceed to next step.

4. It is unclear whether all of the business objectives are entirely compatible with the ethics requirements. One should cautiously proceed to the next step and keep monitoring closely.

As part of the ethics check, specific ethical issues that could be at play in the project in relation to the business objectives should be documented.

During the concept of operations phase, special attention should be given to ethical issues involving privacy, and individual, societal and environmental wellbeing (the latter including issues of mass unemployment and human obsolescence).

Requirement 1b: Stakeholder analysis or involvement in the concept of operations phase

Inclusion of ethical criteria in the development process could benefit from a stakeholder analysis, in which the stakeholders of the project are identified and their values and interests are assessed. This makes it easier to evaluate the importance of the different ethical requirements in the context of the business objectives, and identify more specific requirements that are important to test the objectives against. Going one step further, stakeholders could also be consulted or be involved in decision-making during the concept of operations phase.

Requirements and architecture phase

Requirement 2a: Ethics check of the list of requirements

During the requirements and architecture phase, one should test the list of (technical) requirements for the system under development (derived from the business objectives) against the seven high-level ethics requirements for possible tensions. It is advised to also test these against the operationalised ethics requirements in Annex 2. Even if the business objectives are compatible with the ethics requirements, certain requirements and specifications may be introduced that are incompatible with the ethics requirements. Furthermore, issues that can be overlooked during the assessment of the business objectives may now be noticed more easily given the concreteness and level of detail that the requirements and specifications provide.

If there are tensions between the technical requirements and the high-level ethics requirements (and the operationalised ethics requirements), one should make modifications to the technical requirements to reduce these tensions. There are three possible outcomes and ways to proceed:

1. All of the technical requirements are compatible with the ethics requirements. One can proceed to next step.

2. Some of the technical requirements are incompatible with the ethics requirements, but modifications to these requirements are possible to ensure compatibility. One should modify requirements and proceed to next step.

3. It is unclear whether all the technical requirements are compatible with the ethics requirements. One should cautiously proceed to the next step, and keep monitoring closely.

During the requirements and architecture phase, special attention should be given to ethical issues involving privacy, safety, dual use and misuse, and justice and fairness. All ethical issues that could be at play in relation to the requirements should be documented.

Finally, one should add the high-level ethics requirements to the list of requirements to ensure adherence to them in subsequent phases where this list of requirements is referenced.

Requirement 2b: Ethics check of the system architecture
During the requirements and architecture phase, one should also test the architecture (or high-level design) of the system under development against the seven high-level ethics requirements (and the operationalised ethics requirements) for possible tensions. If there are tensions between the system architecture and the high-level ethics requirements, one should make modifications to the architecture to reduce these tensions. The possible outcomes of the check are analogous to those of Requirement 2a.

Requirement 2c: Stakeholder involvement in the requirements and architecture phase
During the formulation of requirements and the design of a system architecture, the values and interests of stakeholders should be actively considered. One can compare the drawn-up requirements and system architecture with the results of the stakeholder analysis performed in Requirement 1b, and/or consult stakeholders directly during this phase.

Detailed design phase

Requirement 3: Ethics check of the detailed design
During the detailed design phase, one should test the design(s) for the system under development against the seven high-level ethics requirements (and the operationalised ethics requirements) for possible tensions. Ideally, one should not just test the design, but design with the ethics requirements in mind. One should carefully consider how a product based on a particular design may be used, and what the ethical implications are of such use.

If there are tensions between the design and the high-level ethics requirements (and the operationalised ethics requirements), one should make modifications to the design to reduce these tensions. There are, again, three possible outcomes and ways to proceed:

1. All elements of the design are compatible with the ethics requirements. One can proceed to next step.

2. Some elements of the design are incompatible with the ethics requirements, but modifications to the design are possible to ensure compatibility. One should modify the design and proceed to next step.

3. Some elements of the design are incompatible with the ethics requirements, and no modifications can be made to ensure compatibility. One should make a new design (based on different technical principles).

If multiple designs are being considered that perform similarly in functional and financial terms, the ethics requirements should be of key importance in selecting the final design. In cases where best design (out of multiple designs) from a functional and financial perspective is an acceptable but not best solution from an ethical perspective, a cost-benefit analysis may be performed to select the final design.

During the detailed design phase, special attention should be given to ethical issues involving privacy, safety, dual use and misuse, justice and fairness (especially algorithmic bias), and transparency. All ethical issues that could be at play in relation to the design(s) should be documented.

Specific ethical issues maybe associated with the different subsystems of robots—their sensor systems, actuator systems, and control systems:

- Sensor systems may give rise to issues of privacy (e.g., photographing or filming the external environment with a camera), issues of safety (e.g., laser-based sensors), and issues of reliability and error.
- Actuator systems may give rise to issues of safety, health and bodily harm (e.g., speakers that produce loud sounds, robots bumping into humans, risk of psychological harm due to the menacing appearance and movements of some robots' actuator systems).
- Control systems may give rise to issues of safety (e.g., safety risks due to unpredictable robot behaviour), responsibility and accountability (e.g., due to robot autonomy), transparency (e.g., the use of machine learning in robot control systems), privacy (e.g., the collection and use of massive amounts of data, including personal data), and discrimination (e.g., algorithmic bias in robot control systems).

Implementation phase

Requirement 4: Ensuring safety during the implementation phase
During the implementation phase, the system design is being implemented, meaning that all of its individual hardware and software components being are constructed. One should take all necessary precautions to ensure the safety of the individuals constructing the robotics hardware components.

Verification and validation phases

Requirement 5a: Ethics check during the verification and validation phases
As part of the verification and validation phases of the V-Model (which include "Integration, Test and Verification", "System Verification" and "Operation and Maintenance"), ethical checks should be performed of the project's results. During the "Operation and Maintenance" phase, the system's individual components are ethically checked; during the "System Verification" phase, the whole system is ethically evaluated; and during the "Operation and Maintenance", the system's real-world performance is ethically appraised.

Possible outcomes of these checks are the following:

1. The product(s) adhere fully to the ethics requirements. No further development is needed.

2. It has turned out that some elements of the product(s) do not adhere to the ethics requirements, but modifications to the design are a practical solution that will ensure

adherence to the ethics requirements. One should modify the design and product and re-execute verification and validation phases.

3. It has turned out that some elements of the product(s) do not adhere to the ethics requirements. Modifications to the design are an impractical (e.g., very costly) solution to ensure adherence to the ethics requirements. However, specific guidance for, or restrictions on, deployment and use can be put in place to mitigate the ethical issues. One should take such measures.

4. It has turned out that some elements of the product(s) do not adhere to the ethics requirements. Given the nature and seriousness of the ethical issue(s), modifications to the design, and specific guidance for, or restrictions on, deployment and use, are an impractical solution to ensure adherence to the ethics requirements. The product should not be used and further development is to be halted, or one should start again at the design stage.

Requirement 5b: Ensuring safety during the verification and validation phases

During the verification and validation phases, developers work in close proximity to the robotics hardware that they have constructed. One should take all necessary precautions to ensure the safety of individuals working with the robotics hardware.

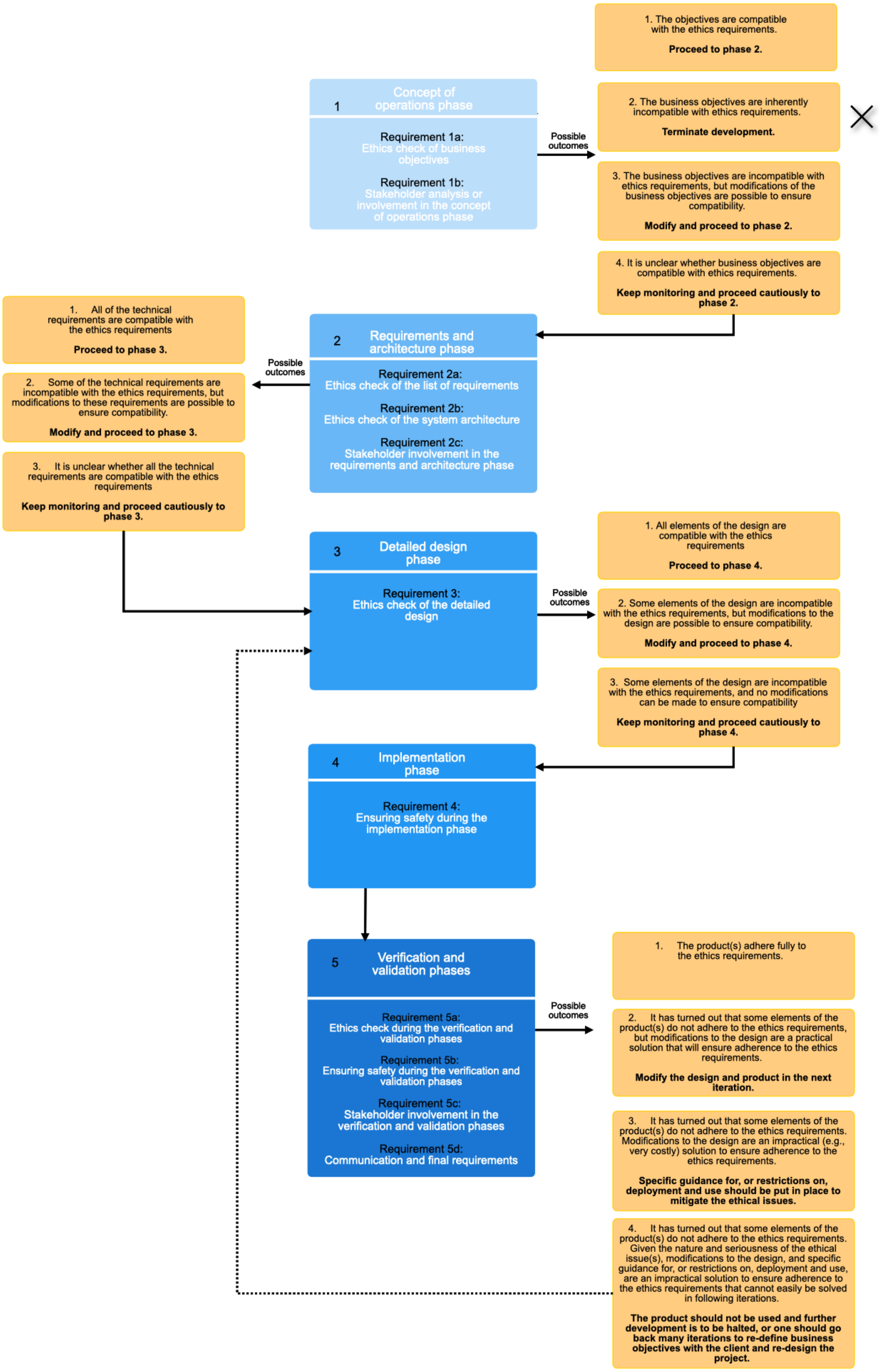
Requirement 5c: Stakeholder involvement in the verification and validation phases

As part of the ethics check in the verification and validation phases of the V-Model, a stakeholder analysis could be performed, or stakeholders could be consulted or involved in the decision-making.

Requirement 5d: Communication and final requirements

Prior to delivering the final product to the client, it is important to communicate all relevant facts and limitations to the client, and to ensure that the system includes ethically required functionality beyond the model, such as mechanisms for human oversight, audibility, and redress.

Figure 8 contains a complete flowchart for the Ethics by Design approach for V-Model.



1 Concept of operations phase

Requirement 1a: Ethics check of business objectives

Requirement 1b: Stakeholder analysis or involvement in the concept of operations phase

1. The objectives are compatible with the ethics requirements. **Proceed to phase 2.**

2. The business objectives are inherently incompatible with ethics requirements. **Terminate development.**

3. The business objectives are incompatible with ethics requirements, but modifications of the business objectives are possible to ensure compatibility. **Modify and proceed to phase 2.**

4. It is unclear whether business objectives are compatible with ethics requirements. **Keep monitoring and proceed cautiously to phase 2.**

2 Requirements and architecture phase

Requirement 2a: Ethics check of the list of requirements

Requirement 2b: Ethics check of the system architecture

Requirement 2c: Stakeholder involvement in the requirements and architecture phase

1. All of the technical requirements are compatible with the ethics requirements **Proceed to phase 3.**

2. Some of the technical requirements are incompatible with the ethics requirements, but modifications to these requirements are possible to ensure compatibility. **Modify and proceed to phase 3.**

3. It is unclear whether all the technical requirements are compatible with the ethics requirements **Keep monitoring and proceed cautiously to phase 3.**

3 Detailed design phase

Requirement 3: Ethics check of the detailed design

1. All elements of the design are compatible with the ethics requirements **Proceed to phase 4.**

2. Some elements of the design are incompatible with the ethics requirements, but modifications to the design are possible to ensure compatibility. **Modify and proceed to phase 4.**

3. Some elements of the design are incompatible with the ethics requirements, and no modifications can be made to ensure compatibility **Keep monitoring and proceed cautiously to phase 4.**

4 Implementation phase

Requirement 4: Ensuring safety during the implementation phase

5 Verification and validation phases

Requirement 5a: Ethics check during the verification and validation phases

Requirement 5b: Ensuring safety during the verification and validation phases

Requirement 5c: Stakeholder involvement in the verification and validation phases

Requirement 5d: Communication and final requirements

1. The product(s) adhere fully to the ethics requirements.

2. It has turned out that some elements of the product(s) do not adhere to the ethics requirements, but modifications to the design are a practical solution that will ensure adherence to the ethics requirements. **Modify the design and product in the next iteration.**

3. It has turned out that some elements of the product(s) do not adhere to the ethics requirements. Modifications to the design are an impractical (e.g., very costly) solution to ensure adherence to the ethics requirements. **Specific guidance for, or restrictions on, deployment and use should be put in place to mitigate the ethical issues.**

4. It has turned out that some elements of the product(s) do not adhere to the ethics requirements. Given the nature and seriousness of the ethical issue(s), modifications to the design, and specific guidance for, or restrictions on, deployment and use, are an impractical solution to ensure adherence to the ethics requirements that cannot easily be solved in following iterations. **The product should not be used and further development is to be halted, or one should go back many iterations to re-define business objectives with the client and re-design the project.**

Fig. 8.

4. Conclusion

The aim of this report was to propose a comprehensive strategy for ethical AI and robotics. In addition, it was an aim to present an approach for Ethics by Design, as part of that strategy. These two aims were undertaken in two major sections of the report, "A strategy for Ethical AI and Robotics" (section 2) and "A framework for Ethics by Design" (section 2).

In section 2, it was claimed that a strategy for ethical AI and robotics should contain three components: (1) an identification of relevant actors; (2) an identification of methods that these actors can use to contribute to ethical AI & robotics, and (3) proposals of ways in which these methods can be made available to these actors, and ways to motivate them to use them. Subsequently, these three components were given content in the report. Six main classes of relevant actors were defined, including AI & robotics developers; AI & robotics development support organizations; organizations that deploy and use AI & robotics technology; governance and standards organizations; educational and media organizations; and civil society organizations and the general public.

Next, six types of methods for ethical AI & robotics were discussed and related to these classes of actors: methods for ethical development and design, methods for ethical deployment and use, corporate responsibility policies and cultures, national and international guidelines, standards and certification, policy and regulation actions (by governments), and education, training and awareness raising. Finally, it was briefly discussed how these methods can be made available to actors.

In section 3, we propose a framework for Ethics by Design, which is a key component of our strategy for ethical AI & robotics. This proposal builds on the SHERPA project deliverable on Ethics by Design (Brey et al., 2019). We propose a general approach for Ethics by Design, with guidelines for including ethical criteria into development processes for AI & robotics, and then specific approaches in relation to the popular CRISP-DM, Agile and V-Model methodologies. In two annexes, we moreover propose extended, detailed guidelines for the incorporation of ethics into Agile and V-Model.

Our discussion of methods for ethical AI & robotics in section 2 is only brief, and we did not have the room to arrive at detailed proposals for many of the methods that we discuss. For many of the proposed methods, however, we refer to both past and planned deliverables that we have completed or are preparing within the SIENNA project, or to other initiatives in which these methods have been or are being developed.

As stated earlier, this strategy is only a first step towards ethical AI & robotics, and a second step consists of its implementation. This requires both the further specification and operationalisation of the methods described in it, the mobilisation of stakeholders and the implementation of the strategy together with these stakeholders. This is what we will spend much of the remainder of the SIENNA project on.

Annex 1

ANNEX 1: Detailed Recommendations for Agile

Agile: Agile Software Development in general

In all phases, assess and ensure that:

VALUE	RECOMMENDATION
HUMAN AGENCY, LIBERTY AND DIGNITY	*1. Fundamental Rights* • 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).
HUMAN AGENCY, LIBERTY AND DIGNITY	*2. Respect for Human Dignity* • the system does not affect human dignity negatively (e.g., by treating individuals means for other goals, rather than as goals in themselves; by disrespectin individuality, e.g., in profiling and data processing; by objectifying or dehumanizin individuals; or by causing harmful effects on human psychology or identity, e.g., b harming their self- control or their sense of self-worth, which may be rooted in th meaning-creation of various human activities such as work); • the system is developed to promote human capacity (e.g., by enabling individu self- development) and humans' intrinsic value is respected in the design proce and by the resulting system; • any individual is aware whether they are interacting with an AI, particularly if the are interacting with an autonomous system.
TECHNICAL ROBUSTNESS AND SAFETY	*3. Security, design, testing, and verification* (specifically in the requirement gathering phase) • you have evaluated the possible security risks and that the system is protected against cybersecurity attacks both during the design process and when implemented; • security is implemented into the system's architecture and that the security of the system is tested and, whenever possible, verified before, during, and after deployment; • security measures are designed to benefit humans.
TECHNICAL ROBUSTNESS AND SAFETY	*4. Resilience* (specifically in the requirement gathering phase) • 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.
TECHNICAL ROBUSTNESS AND SAFETY	*5. Safety and verification* (specifically in the requirement gathering phase, development and evaluation phases) • those responsible for the development of the system have the necessary

	skills to understand how they function and their potential impacts; • mechanisms to safeguard user safety and protect against substantial risks are implemented; • the system is tested before, during, and after deployment, to remain safe and secure throughout its lifetime; • safety measures are designed to benefit humans.
TECHNICAL ROBUSTNESS AND SAFETY	*6. Fallback* (specifically in the requirement gathering phase, development and evaluation phases) • if the system fails it does so safely (e.g., by shutting down safely or going into a safe mode).
PRIVACY AND DATA GOVERNANCE	*7. Clarify roles and responsibilities towards information use, security and privacy* (specifically in requirements gathering, planning & designing and development phases) • there are clear and precise descriptions of the roles and responsibilities of users toward information, media and network usage, security, and privacy; • a common culture is established and encouraged that strongly promotes ethical behaviour for all individuals in the enterprise, and establishes a low tolerance threshold for unethical behaviours.
PRIVACY AND DATA GOVERNANCE	*8. Develop cultures of security and privacy awareness* (specifically in requirements gathering, planning & designing and development) • a culture of security and privacy awareness is established and encouraged that positively influences desirable behaviour and actual implementation of security and privacy policy in daily practice; • a validated log is maintained of who has access to any information that could have implications for security or privacy; • sufficient security and privacy guidance is provided to the developing team during the development process, and to relevant stakeholders both during development and after deployment; • security and privacy champions are indicated (including C-level executives, leaders in HR, and security and/or privacy professionals) and proactively support and communicate security and privacy programs, innovations and challenges; • a culture is established and encouraged that facilitates awareness regarding user responsibility to maintain security and privacy practices; • 'privacy by design' is a core part of the development process and that the end-product abides by these design principles.
PRIVACY AND DATA GOVERNANCE	*9. Personal data use, reduction, and elimination* (specifically in requirements gathering, planning & designing and development) • alternatives that minimize or eliminate the use of personal data or sensitive data are considered and used whenever possible and, in line with the GDRP, that all personal data held is strictly necessary, reasonable and proportionate for the successful execution of business objectives;

	• there are protections against the risk that previously non-sensitive and/or non-personal data may become sensitive or personal (e.g., through the use of aggregation technology).
PRIVACY AND DATA GOVERNANCE	*10. Personal data storage* (specifically in requirements gathering, planning & designing and development) • 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.
PRIVACY AND DATA GOVERNANCE	*11. Data review and minimization* (specifically in requirements gathering, planning & designing, development and testing) • consideration is given to develop the system or train the model with or without minimal use of potentially sensitive or personal data, and applied whenever possible (note that it is questionable whether any data is ever fully anonymized—see Recommendation 2 in Testing); • 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.
PRIVACY AND DATA GOVERNANCE	*12. Alignment with existing standards* • the system is aligned with relevant and appropriate standards (e.g., ISO, IEEE) and/or widely adopted protocols for daily data management and governance.
PRIVACY AND DATA GOVERNANCE	*13. Oversight of access to data* • persons who can access particular data under particular conditions are qualified and required to access the data, and that they have the necessary competence to understand the details of the data protection policy; • there is an embedded oversight mechanism to log when, where, how, by whom, and for what purpose data was accessed, as well as for data collection, storage, processing, and use.
TRANSPARENCY	*14. Responsibility for Traceability* • there is a “human in control” when needed, especially when the system may cause harmful outcomes (e.g., an AI playing a game like chess, which may have no harmful outcomes, would not necessarily require a human in control, unless there was the potential for negative effects); • a balanced prioritisation for human control, related to the plausibility and/or severity of the outcome; • there are measures to enable audit and to remedy issues related to governing the system and allow organisations using your technology the ability to identify

	when there is an issue or harm, and the ability to prevent these issues from occurring, and stop it when these issues are identified; • there are appropriate remedial steps for detection and response mechanisms if something goes wrong, by closely liaison with the organisational user, or end-user.
TRANSPARENCY	*15. Explanations of rationale* • whenever possible, the process of, and rationale behind, the choices made by the system are explainable upon request to an organisational user and/or auditing body in situations where there is a potential and/or existent harm; • the reasons for the collection and use of particular data sets are explainable upon request to organisational users and/or auditing bodies; • in situations where the system-development organisations provide these technologies directly to the end-user, there is redress and explanations of how the system arrived at those decisions, if there is harm caused to the end-user by the system's decisions; • decisions made about individuals are understandable in colloquial language terms for an ordinary (end-)user or stakeholder (e.g., 'You have been put into this category because of x, y, and z').
TRANSPARENCY	*16. Trade-offs* • trade-offs between explainability/transparency and best performance of the system are appropriately balanced based on the systems context of application (e.g., in healthcare the accuracy and performance of the system may be more important than its explainability; whereas, in policing, explainability is much more crucial to justify behaviours and outcomes of law enforcement; and in other areas, such as recruitment, both accuracy and explainability are similarly valued).
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*17. Engagement with users to identify harmful bias* • a mechanism allows others to flag issues related to harmful bias, discrimination, or poor performance of the system and establish clear steps and ways of communicating on how and to whom such issues can be raised (i.e,. during the design, development, and deployment of the system); • there is transparency about how the algorithms may affect individuals to allow for effective stakeholder feedback and engagement; • the implementation of methods for redress and feedback from users at all stages of the system's life-cycle.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*18. Anticipating harmful functional bias* • whenever possible, the potential of the system being used for harmful or illegal purposes is avoided, and that if the system can be used for unintended purposes, then consider potential implications of this likelihood and develop mitigation procedures in the event of potential ethical issues arising; • the system is not designed for bad purposes and attempt to eliminate, whenever

	possible, ways that they can be misused (one way to do this is to use tried-and-tested general models, rather than building all models from scratch).
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*19. Avoiding harmful automation bias* • an appropriate level of human control for the system (by including respective task allocations between the system and humans for meaningful interactions and appropriate human oversight and control); • safeguards are embedded to prevent overconfidence in or overreliance on the system through education and training to be more aware of harmful bias in the system.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*20. Accessibility and Usability* (specifically in the requirement gathering and evaluation phases) • the system is understandable and accessible to users of assistive technologies, users with special needs or disabilities, or groups otherwise at risk of exclusion; • the system is usable by users of assistive technologies, users with special needs or disabilities, or groups otherwise at risk of exclusion (or if the system cannot be *used* properly, attempt to make improvements and ensure that any limitations are fully understood by these groups); • you seek feedback from teams or groups that represent different backgrounds and experiences (including but not limited to users of assistive technologies, users with special needs, or disabilities), and that this process should be accommodating to include different variations and users; • no persons or groups are disproportionately negatively affected by the system. Or if that cannot be ensured, then attempt to minimize the negative effects and ensure that these people and groups fully understand these negative effects before using the system, and that those at risk of being negatively affected are adequately represented in the design process by including feedback from those likely to be affected in the design of the system.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*21. Whistleblowing* • 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.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*22. Diversity* (specifically in the requirement gathering, testing and evaluation phases) • a process to include the participation of different stakeholders in the development, use, and review of the system; • that efforts are made so that a wide diversity of the public, including different sexes, ages, and ethnicities, are represented; • that this is applied within the organization, by informing and involving impacted workers and their representatives in advance.

DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*23. Inclusion* • an adequate inclusion of diverse viewpoints during the development of the system; • that development is based on an acknowledgement that different cultures may respond differently, have different thought processes and patterns, and express themselves differently.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*24. Engagement with stakeholder community* (specifically in requirement gathering and evaluation) • the broader societal impact of the AI system's use beyond the individual (end-)users (such as potentially indirectly affected stakeholders) is evaluated; • the social impacts of the system are well understood (e.g., assess whether there is a risk of job loss , deskilling of the workforce, or changes to occupational structure) and record any steps taken to counteract such risks; • a culture is established and encouraged to ensure timely communication of IT change requests to affected groups, and consult the affected groups regarding implementation and testing of changes; • stakeholders are involved throughout the system's life cycle, and foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.
ACCOUNTABILITY	*25. Engagement and reporting* • incidents are identified and reported on a correct and timely basis and implement appropriate internal and external escalation paths; • incidents are responded to and resolved immediately; • a culture of proactive problem management (detection, action and prevention), with clearly defined roles and responsibilities, is established and encouraged; • a transparent and open environment for reporting problems is established and encouraged, by providing independent reporting mechanisms and/or rewarding people who bring problems forward; • there is an awareness of the importance of an effective control environment; • a proactive risk- and self-aware culture is established and encouraged, including commitment to self-assessment, continuous learning, and independent assurance reviews; • auditability is built into the system; • performance indications are identified and regularly report on the outcomes, in relation to the auditing system.
ACCOUNTABILITY	*26. Compliance as culture* • a compliance-aware culture is established and encouraged, including disciplinary procedures for noncompliance with legal and regulatory requirements; • a culture that embraces internal audit, assurance findings, and recommendations (based on root cause analysis) is established and encouraged;

	• leaders take responsibility to ensure that internal audit and assurance are involved in strategic initiatives and recognize the need for (and value of) audit and assurance reports; • mechanisms that facilitate the system's auditability (such as ensuring traceability and logging of the AI system's processes and outcomes); • in applications affecting fundamental rights (including safety-critical applications) the system can be audited independently; • the developing team attempts to learn to avoid situations requiring accountability in the first place, by ensuring ethical best practices.
ACCOUNTABILITY	*27. Code of ethics* • 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.
ACCOUNTABILITY	*28. Impact on business* • there is an ability to evaluate the degree to which the system's decision influences the organisation's decision-making processes, why this particular system was deployed in this specific area, and how the system creates value for the organization and the general public; • a clear rationale is established by your organization about why you are designing and creating the system, and the intended purpose that it will serve.
ACCOUNTABILITY	*29. Redress mechanisms* • the contextual meaning of accountability is clear for different roles in the development chain (e.g., data scientists, CDOs, board members, business managers), including what form of sanctions are in place for whom, and which roles should take personal responsibility, with redress mechanisms in case of negative impacts; • a set of mechanisms that allows for redress in case the occurrence of any harm or adverse impact is established; • where possible, embed mechanisms to provide information to (end-)users/third parties about opportunities for redress.
ACCOUNTABILITY	*30. Avoiding automation bias* • an appropriate level of human control for the system and use case, including respective task allocations between the system and humans for meaningful interactions and appropriate human oversight and control; • safeguards are embedded to prevent overconfidence in or overreliance on the system for work processes.

ACCOUNTABILITY	*31. Responsibility* • the “human in control”, and the moments or tools for human intervention, are clearly identified; • there are measures to enable audit and to remedy issues related to governing AI autonomy; • there is a human-in-the-loop to control the system, to ensure and protect the autonomy of human beings; • detection and response mechanisms are appropriate in the event of something going wrong.

Agile: Requirements gathering phase

VALUE	**RECOMMENDATION** **IN THE REQUIREMENT GATHERING PHASE, ASSESS AND ENSURE THAT:**
PRIVACY AND DATA GOVERNANCE	*1. Availability of data* • 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); • there is an embedded process that allows individuals to remove their data from the system and/or correct errors in the data where these occur, and ensure that this process is available at any stage in the process (note that once data is correctly and fully anonymized it is no longer considered personal data, although there may be potential for re-identification through aggregation of data sets). • if previously anonymized data is re-identified (see Recommendation 2 in P&D and Dev; Recommendation 3 in P&D and Dev), then these data are made available once more (note, however, that it is questionable whether any data is ever fully anonymized— see Recommendation 2 in Testing).
PRIVACY AND DATA GOVERNANCE	*2. Clarity on ownership of data.* • 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 (Recommendation 2 in P&D and Dev); • agreements stipulate what the owner and (end-)users of the data are permitted to do with those data.
TRANSPARENCY	*3. Communication regarding interactions with the system* • it is communicated to, and presumably understood by, the (end-)users or other affected persons that they are interacting with a non-human agent and/or that

	a decision, content, advice or outcome is the result of an algorithmic decision, in situations where not doing so would be deceptive, misleading, or harmful to the user.
TRANSPARENCY	*4. Communication with stakeholders* • a culture is established and encouraged in which open and structured communication is provided to stakeholders, in line with their requirements (including organisational users and end-users, if you are dealing directly with them). • information to stakeholders, (end-)users, and other affected persons, about the system's capabilities and limitations, is communicated in a clear, understandable, and proactive manner, that enables realistic expectation setting; • it is clear to stakeholders, (end-)users, and other affected persons, what the purpose of the system is and who or what may benefit from the product/service; • usage scenarios for the product are specified and clearly communicated so that they are understandable and appropriate for the intended audience; • in cases where stakeholders cannot be provided with certain data and answers, there is a full disclosure of that limitation, why there is a limitation, and also what they themselves do and do not know.
TRANSPARENCY	*5. Communication within user and stakeholder community* • a culture is established and encouraged based on mutual trust, transparent communication, open and understandable terms, a common language, ownership, and accountability; • an explanation, which all reasonable users and stakeholders can presumably understand, is given as to why the system took a certain choice resulting in a certain outcome; • mechanisms are established to inform organisational users and end-users (if dealing directly with them) about the reasons and criteria behind the system's outcomes and, in collaboration with users, establish processes that consider users' feedback and use this to adapt the system; • any potential or perceived risks are clearly communicated to the (end-)user (e.g., consider human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue).
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*6. Bias assessment in Planning* • the potential for harmful bias in the business understanding and requirements stage is evaluated and, if possible, avoided (e.g., some requirements may inadvertently favour particular groups in society over others, e.g., if you are using the system to hire a new candidate, there may be more gender- or ethnicity-specific characteristics entered into the criteria for assessment, which would have negatively biased results); • developing teams receive unconscious bias training to assist developers to identify innate biases during the development of systems.

INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*7. Environmental impact* • a mechanism to measure the ecological impact of the system's use (e.g., the energy used by data centres). • where possible, measures to reduce the ecological impact of your system's life cycle; • an adherence to resource-efficiency, sustainable energy-promotion, the protection of the non-human living world around us, and the attempt to ensure biodiversity and the healthy functioning of ecosystems (in particular, decisions made by the system that will directly affect the non-human world around us need to be carefully factored in, with strong emphasis on the impact on these ecological externalities, through a holistic ecosystem-focused outlook); • transparency about ecological impact and, if possible, work with environmental protection organisations to ensure that the system is sustainable, and keep the ecological footprint proportionate to the intended benefit to humanity.
ACCOUNTABILITY	*8. Reporting Impacts* • a risk assessment is conducted, which takes into account different stakeholders that are (in)directly affected by the system and the likelihood of those impacts; • training and education is provided to help develop accountability practices (including teachings of the potential legal framework applicable to the system); • if possible, that an 'ethical AI review board' or a similar mechanism is established to discuss overall accountability and ethics practices, including potentially unclear grey areas; • processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks, or biases in the system, is established.
ACCOUNTABILITY	*9. Minimising negative impact* • a process for minimization of negative impacts (such as external guidance and/or an auditing processes to oversee ethics and accountability), in addition to internal initiatives; • that audit controls are built into the system to check performance, record decisions made about the purpose and functioning of the system (including reporting on the impacts in general, not just occurrences of negative impacts); • an attempt to predict the consequences/externalities of the system's processing.

Agile: Planning & Designing phase

VALUE	**REQUIREMENT** **IN THE PLANNING & DESIGNING PHASE, ASSESS AND ENSURE THAT:**

TECHNICAL ROBUSTNESS AND SAFETY	*1. Accuracy, reliability, and effectiveness* • the accuracy, reliability, and effectiveness of the system.
PRIVACY AND DATA GOVERNANCE	*2. Creation of new personal data* • If needed, further informed consent is acquired (or, if not, that there is an alternative legal basis as set out in Articles 6(1) and 9(2) of GDPR) for the creation of new personal or sensitive information/data (e.g., through estimation of missing data, the production of derived attributes and new records, data integration, or aggregation of data sets); • all newly created personal or sensitive information/data is given at least the same protection and attracts the same rights as previously collected or held personal or sensitive information/data.
PRIVACY AND DATA GOVERNANCE	*3. Subsequent collection and/or creation of new personal data* • no new personal information is or can be collected or created during regular use of the system, unless necessary (e.g., for the function of the system or realization of the business objectives); • if new personal information is collected or created, then limitations are properly imposed to protect individuals' privacy or sensitive information/data, and further informed consent is acquired, if needed.
PRIVACY AND DATA GOVERNANCE	*4. Oversight of data quality* • there are processes to ensure the quality and integrity of all pertinent data, including means of verifying that data sets have not been compromised or hacked (if you are in control of the quality of the external data sources used, to assess to what degree you can validate their quality); • a culture of shared responsibility for the organization's data assets is established and encouraged; • the potential value of data assets is acknowledged, and that roles and responsibilities are clear for governance and management of data assets; • the impact and risk of data loss is continuously communicated; • employees understand the true cost of failing to implement a data quality culture.
PRIVACY AND DATA GOVERNANCE	*5. Availability of data* • 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); • there is an embedded process that allows individuals to remove their data from the system and/or correct errors in the data where these occur, and ensure that this process is available at any stage in the process (note that once data is correctly and fully anonymized it is no longer considered personal data, although there may be potential for re-identification through aggregation of data sets). • if previously anonymized data is re-identified (see Recomendation 2 and 3 Planning & Designing and Development phases), then these data are made

	available once more (note, however, that it is questionable whether any data is ever fully anonymized— see Recommendation 2 in Testing).
TRANSPARENCY	*6. Traceability measures* • measurements to ensure traceability are established through the following methods: ○ Methods used for designing and developing systems (rule-based AI systems: the method of programming or how the model was built; learning-based AI systems: the method of training the algorithm, including which data was gathered and selected, and how this occurred); ○ Methods used to test and validate systems (rule-based AI systems: the scenarios or cases used in order to test and validate; learning-based model: information about the data used to test and validate); ○ Outcomes of the system (outcomes of or decisions taken by the system, as well as potential other decisions that would result from different cases, e.g., for other subgroups of users); ○ A series of technical methods to ensure traceability should be taken (such as encoding the metadata to extract and trace it when required). There should be a way of capturing where the data has come from, and the ability to construct how the different pieces of data relate to one another.
TRANSPARENCY	*7. Training data* • if possible, you can analyse your training data, that your data is representative, and value aligned; • whenever possible, there is an ability to go back to each state the system has been in to determine or predict what the system would have done at time t and, whenever possible, determine which training data was used. • in the event of a system malfunction or harm resulting from the system, as much transparency as is possible of your training data is made available, without violating privacy, to the appropriate authorities.
TRANSPARENCY	*8. Explainable systems* • you know to what degree the decisions and outcomes made by the system can be understood, including whether you have access to the internal workflow of the model; • explainability is guaranteed (through technologies such as Explainable AI), when there is a greater emphasis within its use for explainability over performance, or when there is no trade-off between explainability and performance.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*9a. Bias assessment in data analysis* • an evaluation is performed to determine the diversity and representativeness of users in the data, testing for specific populations or problematic use cases is performed, and that input, training, and output data is analysed for harmful bias; • the potential for harmful bias in the data understanding stage is evaluated (e.g., some data sets may contain harmful biases if they consist solely of the behaviour

	of subclasses of all people, e.g., young white men, and if the system is deployed in situations where groups other than those in the data set will be affected) and, if possible, avoided (e.g., incorporate additional users' data that is not included in the data; look at the alternative or additional supply chains from the data that you are using; or in some cases, the datasets need to be discarded altogether). • data from just one class is not used to represent another class, unless it is justifiably representative.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*9b. Bias assessment in data preparation* • the potential for harmful bias in the data preparation stage is evaluated and, if possible, avoided (e.g., the cleaning of the data set may inadvertently remove data relating to certain minority or under-represented groups, leaving the data set as a whole biased); • you have clearly established what kind of sample you need, what kind of sample you have taken, and that you articulate what it will be used for.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*9c. Bias assessment in modelling* • the potential for harmful bias in the modelling stage is evaluated and, if possible, avoided (e.g., some algorithms make assumptions about universal behaviours and characteristics which are untrue; many behaviours which are assumed to be universal are in fact culturally specific); • a strategy or a set of procedures is established to avoid creating or reinforcing unfair bias in the system regarding the use of input data as well as for the algorithm's design, and that the strategy includes an assessment of the possible limitations stemming from the composition of the used data sets; there is in the design process an awareness of cultural bias to prevent or exacerbate any potential harmful bias.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*10. Intended use* • 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.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*11. Individual wellbeing assessment* • the system is evaluated for its likely and potential impact on individual wellbeing (including consideration of the way in which the system will or could be used which may be detrimental to users or stakeholders). Particular care should be taken for detriments towards vulnerable groups through discussion with them, rather than assuming their needs. • Discussions with vulnerable groups should be made in the planning phase before each period in order to plan it into the period schedule. This may be based on the evaluation of a previous period.

Agile: Development phase

VALUE	REQUIREMENT IN THE DEVELOPMENT PHASE, ASSESS AND ENSURE THAT:
TECHNICAL ROBUSTNESS AND SAFETY	*1. Accuracy, reliability, and effectiveness* • the accuracy, reliability, and effectiveness of the system.
PRIVACY AND DATA GOVERNANCE	*2. Creation of new personal data* • If needed, further informed consent is acquired (or, if not, that there is an alternative legal basis as set out in Articles 6(1) and 9(2) of GDPR) for the creation of new personal or sensitive information/data (e.g., through estimation of missing data, the production of derived attributes and new records, data integration, or aggregation of data sets); • all newly created personal or sensitive information/data is given at least the same protection and attracts the same rights as previously collected or held personal or sensitive information/data.
PRIVACY AND DATA GOVERNANCE	*3. Subsequent collection and/or creation of new personal data* • no new personal information is or can be collected or created during regular use of the system, unless necessary (e.g., for the function of the system or realization of the business objectives); • if new personal information is collected or created, then limitations are properly imposed to protect individuals' privacy or sensitive information/data, and further informed consent is acquired, if needed.
PRIVACY AND DATA GOVERNANCE	*4. Oversight of data quality* • there are processes to ensure the quality and integrity of all pertinent data, including means of verifying that data sets have not been compromised or hacked (if you are in control of the quality of the external data sources used, to assess to what degree you can validate their quality); • a culture of shared responsibility for the organization's data assets is established and encouraged; • the potential value of data assets is acknowledged, and that roles and responsibilities are clear for governance and management of data assets; • the impact and risk of data loss is continuously communicated; • employees understand the true cost of failing to implement a data quality culture.
PRIVACY AND DATA GOVERNANCE	*5. Availability of data* • 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);

	• there is an embedded process that allows individuals to remove their data from the system and/or correct errors in the data where these occur, and ensure that this process is available at any stage in the process (note that once data is correctly and fully anonymized it is no longer considered personal data, although there may be potential for re-identification through aggregation of data sets). • if previously anonymized data is re-identified (see Recommendation 2 P&D and 2 Dev; Recommendation 3 P&D and 3 Dev), then these data are made available once more (note, however, that it is questionable whether any data is ever fully anonymized— see Recommendation 2 in Testing).
TRANSPARENCY	*6. Traceability measures* • measurements to ensure traceability are established through the following methods: ○ Methods used for designing and developing systems (rule-based AI systems: the method of programming or how the model was built; learning-based AI systems: the method of training the algorithm, including which data was gathered and selected, and how this occurred); ○ Methods used to test and validate systems (rule-based AI systems: the scenarios or cases used in order to test and validate; learning-based model: information about the data used to test and validate); ○ Outcomes of the system (outcomes of or decisions taken by the system, as well as potential other decisions that would result from different cases, e.g., for other subgroups of users); ○ A series of technical methods to ensure traceability should be taken (such as encoding the metadata to extract and trace it when required). There should be a way of capturing where the data has come from, and the ability to construct how the different pieces of data relate to one another.
TRANSPARENCY	*7. Understandability of Code* • the code is actively explained and documented within the software program and understandable to fellow programmers; • the code is peer-reviewed by fellow programmers.
TRANSPARENCY	*8. Training data* • if possible, you can analyse your training data, that your data is representative, and value aligned; • whenever possible, there is an ability to go back to each state the system has been in to determine or predict what the system would have done at time t and, whenever possible, determine which training data was used. • in the event of a system malfunction or harm resulting from the system, as much transparency as is possible of your training data is made available, without violating privacy, to the appropriate authorities.
TRANSPARENCY	*9. Explainable systems*

	• you know to what degree the decisions and outcomes made by the system can be understood, including whether you have access to the internal workflow of the model; • explainability is guaranteed (through technologies such as Explainable AI), when there is a greater emphasis within its use for explainability over performance, or when there is no trade-off between explainability and performance.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*10a. Bias assessment in data analysis* • an evaluation is performed to determine the diversity and representativeness of users in the data, testing for specific populations or problematic use cases is performed, and that input, training, and output data is analysed for harmful bias; • the potential for harmful bias in the data understanding stage is evaluated (e.g., some data sets may contain harmful biases if they consist solely of the behaviour of subclasses of all people, e.g., young white men, and if the system is deployed in situations where groups other than those in the data set will be affected) and, if possible, avoided (e.g., incorporate additional users' data that is not included in the data; look at the alternative or additional supply chains from the data that you are using; or in some cases, the datasets need to be discarded altogether). • data from just one class is not used to represent another class, unless it is justifiably representative.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*10b. Bias assessment in data preparation* • the potential for harmful bias in the data preparation stage is evaluated and, if possible, avoided (e.g., the cleaning of the data set may inadvertently remove data relating to certain minority or under-represented groups, leaving the data set as a whole biased); • you have clearly established what kind of sample you need, what kind of sample you have taken, and that you articulate what it will be used for.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*10c. Bias assessment in modelling* • the potential for harmful bias in the modeling stage is evaluated and, if possible, avoided (e.g., some algorithms make assumptions about universal behaviours and characteristics which are untrue; many behaviours which are assumed to be universal are in fact culturally specific); • a strategy or a set of procedures is established to avoid creating or reinforcing unfair bias in the system regarding the use of input data as well as for the algorithm's design, and that the strategy includes an assessment of the possible limitations stemming from the composition of the used data sets; • there is in the design process an awareness of cultural bias to prevent or exacerbate any potential harmful bias.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*11. Intended use* • 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.

Agile: Testing

VALUE	**REQUIREMENT** **IN THE TESTING PHASE, ASSESS AND ENSURE THAT:**
PRIVACY AND DATA GOVERNANCE	1. *Privacy awareness* • mechanisms allowing developers and users to flag issues related to privacy or data protection in the system's processes of data collection (including for training and operation) and data processing; • mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable).
PRIVACY AND DATA GOVERNANCE	2. *Protection against re-identification* • 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).
TRANSPARENCY	3. *Communication regarding interactions with the system* • it is communicated to, and presumably understood by, the (end-)users or other affected persons that they are interacting with a non-human agent and/or that a decision, content, advice or outcome is the result of an algorithmic decision, in situations where not doing so would be deceptive, misleading, or harmful to the user.
TRANSPARENCY	4. *Communication with stakeholders* • a culture is established and encouraged in which open and structured communication is provided to stakeholders, in line with their requirements (including organisational users and end-users, if you are dealing directly with them). • information to stakeholders, (end-)users, and other affected persons, about the system's capabilities and limitations, is communicated in a clear, understandable, and proactive manner, that enables realistic expectation setting; • it is clear to stakeholders, (end-)users, and other affected persons, what the purpose of the system is and who or what may benefit from the product/service; • usage scenarios for the product are specified and clearly communicated so that they are understandable and appropriate for the intended audience; • in cases where stakeholders cannot be provided with certain data and answers, there is a full disclosure of that limitation, why there is a limitation, and also what they themselves do and do not know.
TRANSPARENCY	5. *Communication within user and stakeholder community*

	• a culture is established and encouraged based on mutual trust, transparent communication, open and understandable terms, a common language, ownership, and accountability; • an explanation, which all reasonable users and stakeholders can presumably understand, is given as to why the system took a certain choice resulting in a certain outcome; • mechanisms are established to inform organisational users and end-users (if dealing directly with them) about the reasons and criteria behind the system's outcomes and, in collaboration with users, establish processes that consider users' feedback and use this to adapt the system; • any potential or perceived risks are clearly communicated to the (end-)user (e.g., consider human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue).
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*6. Decision variability* • a measurement or assessment mechanism, of the potential impact of decision variability on fundamental rights, is established based on an evaluation of the system's possibility for decision variability that can occur under the same conditions; • variability is explained to the organisational user of the system and/or the end-user (if they are using it directly). For example, in medicine this should be explained to doctors that use it.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*7. Distributing the system to organisational users* • the user interface is clearly presented, including information about potential errors and the accuracy of the system (including the underlying certainty).
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*8. Environmental impact* • a mechanism to measure the ecological impact of the system's use (e.g., the energy used by data centres). • where possible, measures to reduce the ecological impact of your system's life cycle; • an adherence to resource-efficiency, sustainable energy-promotion, the protection of the non-human living world around us, and the attempt to ensure biodiversity and the healthy functioning of ecosystems (in particular, decisions made by the system that will directly affect the non-human world around us need to be carefully factored in, with strong emphasis on the impact on these ecological externalities, through a holistic ecosystem-focused outlook); • transparency about ecological impact and, if possible, work with environmental protection organisations to ensure that the system is sustainable, and keep the ecological footprint proportionate to the intended benefit to humanity.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*9. Individual wellbeing assessment* • the system is evaluated for its likely and potential impact on individual wellbeing (including consideration of the way in which the system will or could be used which may be detrimental to users or stakeholders). Particular care should be

	taken for detriments towards vulnerable groups through discussion with them, rather than assuming their needs. • Discussions with vulnerable groups should be made in the planning phase before each period in order to plan it into the period schedule. This may be based on the evaluation of a previous period.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*10. Mitigation of impacts on democracy* • an evaluation of whether the system is intended, or could be used for, supporting, organizing or influencing political processes, including political messaging and communication, and if so, take measures to ensure that the system supports democratic processes and protects against interventions that manipulates, misleads or excludes voters and distorts democratic processes; • compliance with higher authorities of AI development and implement an ethical officer to ensure corporate social responsibility within the company; • that external ethics audits are carried out to guarantee that system development is not harming democratic processes.
ACCOUNTABILITY	*11. Reporting Impacts* • a risk assessment is conducted, which takes into account different stakeholders that are (in)directly affected by the system and the likelihood of those impacts; • training and education is provided to help develop accountability practices (including teachings of the potential legal framework applicable to the system); • if possible, that an 'ethical AI review board' or a similar mechanism is established to discuss overall accountability and ethics practices, including potentially unclear grey areas; • processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks, or biases in the system, is established.
ACCOUNTABILITY	*12. Minimising negative impact* • a process for minimization of negative impacts (such as external guidance and/or an auditing processes to oversee ethics and accountability), in addition to internal initiatives; • that audit controls are built into the system to check performance, record decisions made about the purpose and functioning of the system (including reporting on the impacts in general, not just occurrences of negative impacts); • an attempt to predict the consequences/externalities of the system's processing.
ACCOUNTABILITY	*13. Identify interests and values at risk* • a mechanism to identify relevant interests and values implicated by the system and potential trade-offs between them, before deployment and during the life-cycle of the system, which should include considerations regarding how trade-offs were decided and documented;

	• the establishment of values and interests at risk, through stakeholder analysis, product testing, discussion groups, external workshops, and a range of diversity and inclusion sessions.
ACCOUNTABILITY	*14. Install systems to allow for internal complaint* • the existence and advertisement (through the companies) of a clear complaints and whistleblowing system (directing employees to a suitable contact venue and setting out the process for registering both anonymous and identifiable complaints); • that employees are aware of a zero-tolerance policy for any recriminations for whistleblowing or the registering of internal complaints.

Agile: Evaluation

VALUE	**REQUIREMENT** **IN THE EVALUATION PHASE, ASSESS AND ENSURE THAT:**
PRIVACY AND DATA GOVERNANCE	*1. Privacy awareness* • mechanisms allowing developers and users to flag issues related to privacy or data protection in the system's processes of data collection (including for training and operation) and data processing; • mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable).
PRIVACY AND DATA GOVERNANCE	*2. Oversight of data quality* • there are processes to ensure the quality and integrity of all pertinent data, including means of verifying that data sets have not been compromised or hacked (if you are in control of the quality of the external data sources used, to assess to what degree you can validate their quality); • a culture of shared responsibility for the organization's data assets is established and encouraged; • the potential value of data assets is acknowledged, and that roles and responsibilities are clear for governance and management of data assets; • the impact and risk of data loss is continuously communicated; • employees understand the true cost of failing to implement a data quality culture.
TRANSPARENCY	*3. Traceability measures* • measurements to ensure traceability are established through the following methods: ○ Methods used for designing and developing systems (rule-based AI systems: the method of programming or how the model was built; learning-based AI systems: the method of training the algorithm, including which data was gathered and selected, and how this occurred);

	○ Methods used to test and validate systems (rule-based AI systems: the scenarios or cases used in order to test and validate; learning-based model: information about the data used to test and validate); ○ Outcomes of the system (outcomes of or decisions taken by the system, as well as potential other decisions that would result from different cases, e.g., for other subgroups of users); ○ A series of technical methods to ensure traceability should be taken (such as encoding the metadata to extract and trace it when required). There should be a way of capturing where the data has come from, and the ability to construct how the different pieces of data relate to one another.
TRANSPARENCY	*4. Understandability of Code* • the code is actively explained and documented within the software program and understandable to fellow programmers; • the code is peer-reviewed by fellow programmers.
TRANSPARENCY	*5. Training data* • if possible, you can analyse your training data, that your data is representative, and value aligned; • whenever possible, there is an ability to go back to each state the system has been in to determine or predict what the system would have done at time t and, whenever possible, determine which training data was used. • in the event of a system malfunction or harm resulting from the system, as much transparency as is possible of your training data is made available, without violating privacy, to the appropriate authorities.
TRANSPARENCY	*6. Explainable systems* • you know to what degree the decisions and outcomes made by the system can be understood, including whether you have access to the internal workflow of the model; • explainability is guaranteed (through technologies such as Explainable AI), when there is a greater emphasis within its use for explainability over performance, or when there is no trade-off between explainability and performance.
TRANSPARENCY	*7. Communication with stakeholders* • a culture is established and encouraged in which open and structured communication is provided to stakeholders, in line with their requirements (including organisational users and end-users, if you are dealing directly with them). • information to stakeholders, (end-)users, and other affected persons, about the system's capabilities and limitations, is communicated in a clear, understandable, and proactive manner, that enables realistic expectation setting; • it is clear to stakeholders, (end-)users, and other affected persons, what the purpose of the system is and who or what may benefit from the product/service; • usage scenarios for the product are specified and clearly communicated so that they are understandable and appropriate for the intended audience;

	• in cases where stakeholders cannot be provided with certain data and answers, there is a full disclosure of that limitation, why there is a limitation, and also what they themselves do and do not know.
TRANSPARENCY	*8. Communication within user and stakeholder community* • a culture is established and encouraged based on mutual trust, transparent communication, open and understandable terms, a common language, ownership, and accountability; • an explanation, which all reasonable users and stakeholders can presumably understand, is given as to why the system took a certain choice resulting in a certain outcome; • mechanisms are established to inform organisational users and end-users (if dealing directly with them) about the reasons and criteria behind the system's outcomes and, in collaboration with users, establish processes that consider users' feedback and use this to adapt the system; • any potential or perceived risks are clearly communicated to the (end-)user (e.g., consider human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue).
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*9. Bias assessment in Planning* • the potential for harmful bias in the business understanding and requirements stage is evaluated and, if possible, avoided (e.g., some requirements may inadvertently favour particular groups in society over others, e.g., if you are using the system to hire a new candidate, there may be more gender- or ethnicity-specific characteristics entered into the criteria for assessment, which would have negatively biased results); • developing teams receive unconscious bias training to assist developers to identify innate biases during the development of systems.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*10a. Bias assessment in data analysis* • an evaluation is performed to determine the diversity and representativeness of users in the data, testing for specific populations or problematic use cases is performed, and that input, training, and output data is analysed for harmful bias; • the potential for harmful bias in the data understanding stage is evaluated (e.g., some data sets may contain harmful biases if they consist solely of the behaviour of subclasses of all people, e.g., young white men, and if the system is deployed in situations where groups other than those in the data set will be affected) and, if possible, avoided (e.g., incorporate additional users' data that is not included in the data; look at the alternative or additional supply chains from the data that you are using; or in some cases, the datasets need to be discarded altogether). • data from just one class is not used to represent another class, unless it is justifiably representative.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*10b. Bias assessment in data preparation* • the potential for harmful bias in the data preparation stage is evaluated and, if possible, avoided (e.g., the cleaning of the data set may inadvertently remove

	data relating to certain minority or under-represented groups, leaving the data set as a whole biased); • you have clearly established what kind of sample you need, what kind of sample you have taken, and that you articulate what it will be used for.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*10c. Bias assessment in modelling* • the potential for harmful bias in the modeling stage is evaluated and, if possible, avoided (e.g., some algorithms make assumptions about universal behaviours and characteristics which are untrue; many behaviours which are assumed to be universal are in fact culturally specific); • a strategy or a set of procedures is established to avoid creating or reinforcing unfair bias in the system regarding the use of input data as well as for the algorithm's design, and that the strategy includes an assessment of the possible limitations stemming from the composition of the used data sets; • there is in the design process an awareness of cultural bias to prevent or exacerbate any potential harmful bias.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*11. Decision variability* • a measurement or assessment mechanism, of the potential impact of decision variability on fundamental rights, is established based on an evaluation of the system's possibility for decision variability that can occur under the same conditions; • variability is explained to the organisational user of the system and/or the end-user (if they are using it directly). For example, in medicine this should be explained to doctors that use it.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*12. Intended use* • 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.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*13. Distributing the system to organisational users* • the user interface is clearly presented, including information about potential errors and the accuracy of the system (including the underlying certainty).
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*14. Environmental impact* • a mechanism to measure the ecological impact of the system's use (e.g., the energy used by data centres). • where possible, measures to reduce the ecological impact of your system's life cycle; • an adherence to resource-efficiency, sustainable energy-promotion, the protection of the non-human living world around us, and the attempt to ensure biodiversity and the healthy functioning of ecosystems (in particular, decisions made by the system that will directly affect the non-human world around us need to be carefully factored in, with strong emphasis on the impact on these ecological externalities, through a holistic ecosystem-focused outlook);

	• transparency about ecological impact and, if possible, work with environmental protection organisations to ensure that the system is sustainable, and keep the ecological footprint proportionate to the intended benefit to humanity.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*15. Individual wellbeing assessment* • the system is evaluated for its likely and potential impact on individual wellbeing (including consideration of the way in which the system will or could be used which may be detrimental to users or stakeholders). Particular care should be taken for detriments towards vulnerable groups through discussion with them, rather than assuming their needs. • Discussions with vulnerable groups should be made in the planning phase before each period in order to plan it into the period schedule. This may be based on the evaluation of a previous period.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*16. Emotional attachment* • if the system is developed to interact directly with humans, evaluate whether it encourages humans to develop unwanted attachment and unwanted empathy towards the system or detrimental addiction to the system, and if so take appropriate action to minimize such effects; • the system clearly communicates that its social interaction is simulated and that it lacks human capacities such as "understanding" and "feelings"; • the system does not make humans believe it has consciousness (e.g., through expressions that simulate emotions).
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*17. Societal impact assessment* • the system's likely and potential impact on social relationships and social cohesion (including consideration of the way in which the system will or could be used which may be detrimental to groups of users or groups of stakeholders) is not inappropriate; • social benefits are determined through social metrics, not simply measurements in terms of GDP (e.g., liveability indexes).
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*18. Mitigation of impacts on democracy* • an evaluation of whether the system is intended, or could be used for, supporting, organizing or influencing political processes, including political messaging and communication, and if so, take measures to ensure that the system supports democratic processes and protects against interventions that manipulates, misleads or excludes voters and distorts democratic processes; • compliance with higher authorities of AI development and implement an ethical officer to ensure corporate social responsibility within the company; • that external ethics audits are carried out to guarantee that system development is not harming democratic processes.
ACCOUNTABILITY	*19. Reporting Impacts* • a risk assessment is conducted, which takes into account different stakeholders that are (in)directly affected by the system and the likelihood of those impacts;

	• training and education is provided to help develop accountability practices (including teachings of the potential legal framework applicable to the system); • if possible, that an 'ethical AI review board' or a similar mechanism is established to discuss overall accountability and ethics practices, including potentially unclear grey areas; • processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks, or biases in the system, is established.
ACCOUNTABILITY	*20. Minimising negative impact* • a process for minimization of negative impacts (such as external guidance and/or an auditing processes to oversee ethics and accountability), in addition to internal initiatives; • that audit controls are built into the system to check performance, record decisions made about the purpose and functioning of the system (including reporting on the impacts in general, not just occurrences of negative impacts); • an attempt to predict the consequences/externalities of the system's processing.
ACCOUNTABILITY	*21. Identify interests and values at risk* • a mechanism to identify relevant interests and values implicated by the system and potential trade-offs between them, before deployment and during the life-cycle of the system, which should include considerations regarding how trade-offs were decided and documented; • the establishment of values and interests at risk, through stakeholder analysis, product testing, discussion groups, external workshops, and a range of diversity and inclusion sessions.
ACCOUNTABILITY	*22. Install systems to allow for internal complaint* • the existence and advertisement (through the companies) of a clear complaints and whistleblowing system (directing employees to a suitable contact venue and setting out the process for registering both anonymous and identifiable complaints); • that employees are aware of a zero-tolerance policy for any recriminations for whistleblowing or the registering of internal complaints.

ANNEX 2: Detailed Requirements for the V-Model

V-Model: Requirements for all V-Model phases

VALUE	REQUIREMENT IN ALL PHASES, ASSESS AND ENSURE THAT:
HUMAN AGENCY, LIBERTY AND DIGNITY	*1. Fundamental Rights* • 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).
HUMAN AGENCY, LIBERTY AND DIGNITY	*2. Respect for Human Dignity* • the system does not affect human dignity negatively (e.g., by treating individuals as means for other goals, rather than as goals in themselves; by disrespecting individuality, e.g., in profiling and data processing; by objectifying or dehumanizing individuals; or by causing harmful effects on human psychology or identity, e.g., by harming their self-control or their sense of self-worth, which may be rooted in the meaning-creation of various human activities such as work); • the system is developed to promote human capacity (e.g., by enabling individual self-development) and humans' intrinsic value is respected in the design process and by the resulting system; • any individual is aware whether they are interacting with a robot, particularly if they are interacting with one that has a large degree of autonomy.
TECHNICAL ROBUSTNESS AND SAFETY	*3. Security, design, testing, and verification* • you have evaluated the possible security risks and that the system is protected against cybersecurity attacks both during the design process and when implemented; • security is implemented into the system's architecture and that the security of the system is tested and, whenever possible, verified before, during, and after deployment; • security measures are designed to benefit humans.
TECHNICAL ROBUSTNESS AND SAFETY	*4. Resilience* • 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.
TECHNICAL ROBUSTNESS AND SAFETY	*5. Safety and verification* • those responsible for the development of the system have the necessary skills to understand how they function and their potential impacts;

	• mechanisms to safeguard user safety and protect against substantial risks are implemented; • the system is tested before, during, and after deployment, to remain safe and secure throughout its lifetime; • safety measures are designed to benefit humans.
TECHNICAL ROBUSTNESS AND SAFETY	*6. Fallback* • if the system fails it does so safely (e.g., by shutting down safely or going into a safe mode).
TECHNICAL ROBUSTNESS AND SAFETY	*7. Dual-use and misuse* • proper authorities are consulted before development, and relevant national and supranational regulations are adhered to, if the system is found to have significant military applications or if it contributes to the proliferation of weapons of mass destruction. • precautions are taken to prevent or counter the effects of potential malicious use (e.g., by criminals) of the system if such use is deemed likely (e.g., the appointment of a security advisor, limiting dissemination, classification, training for staff).
PRIVACY AND DATA GOVERNANCE	*8. Clarify roles and responsibilities towards information use, security and privacy* • there are clear and precise descriptions of the roles and responsibilities of users toward information, media and network usage, security, and privacy; • a common culture is established and encouraged that strongly promotes ethical behaviour for all individuals in the enterprise, and establishes a low tolerance threshold for unethical behaviours.
PRIVACY AND DATA GOVERNANCE	*9. Develop cultures of security and privacy awareness* • a culture of security and privacy awareness is established and encouraged that positively influences desirable behaviour and actual implementation of security and privacy policy in daily practice; • a validated log is maintained of who has access to any information that could have implications for security or privacy; • sufficient security and privacy guidance is provided to the developing team during the development process, and to relevant stakeholders both during development and after deployment; • security and privacy champions are indicated (including C-level executives, leaders in HR, and security and/or privacy professionals) and proactively support and communicate security and privacy programs, innovations and challenges; • a culture is established and encouraged that facilitates awareness regarding user responsibility to maintain security and privacy practices; • 'privacy by design' is a core part of the development process and that the end-product abides by these design principles.

PRIVACY AND DATA GOVERNANCE	*10. Personal data use, reduction, and elimination* • alternatives that minimize or eliminate the use of personal data or sensitive data are considered and used whenever possible and, in line with the GDRP, that all personal data held is strictly necessary, reasonable and proportionate for the successful execution of business objectives; • there are protections against the risk that previously non-sensitive and/or non-personal data may become sensitive or personal (e.g., through the use of aggregation technology).
PRIVACY AND DATA GOVERNANCE	*11. Personal data storage* • 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.
PRIVACY AND DATA GOVERNANCE	*12. Alignment with existing standards* • the system is aligned with relevant and appropriate standards (e.g., ISO, IEEE) and/or widely adopted protocols for daily data management and governance.
PRIVACY AND DATA GOVERNANCE	*13. Data Protection Officers* • a Data Protection Officer (DPO), where one exists, is adequately involved in the development process.
PRIVACY AND DATA GOVERNANCE	*14. Oversight of data quality* • there are processes to ensure the quality and integrity of all pertinent data, including means of verifying that data sets have not been compromised or hacked (if you are in control of the quality of the external data sources used, to assess to what degree you can validate their quality); • a culture of shared responsibility for the organization's data assets is established and encouraged; • the potential value of data assets is acknowledged, and that roles and responsibilities are clear for governance and management of data assets; • the impact and risk of data loss is continuously communicated; • employees understand the true cost of failing to implement a data quality culture.
PRIVACY AND DATA GOVERNANCE	*15. Employment of protocols and procedures for data governance* • appropriate protocols, processes, and procedures are followed to manage and ensure proper data governance; • there are reasonable safeguards for compliance with relevant protocols, processes and procedures for your industry.
PRIVACY AND DATA GOVERNANCE	*16. Oversight of access to data*

	• persons who can access particular data under particular conditions are qualified and required to access the data, and that they have the necessary competence to understand the details of the data protection policy; • there is an embedded oversight mechanism to log when, where, how, by whom, and for what purpose data was accessed, as well as for data collection, storage, processing, and use.
TRANSPARENCY	*17. Trade-offs* • trade-offs between explainability/transparency and best performance of the system are appropriately balanced based on the systems context of application (e.g., in healthcare the accuracy and performance of the system may be more important than its explainability; whereas, in policing, explainability is much more crucial to justify behaviours and outcomes of law enforcement; and in other areas, such as recruitment, both accuracy and explainability are similarly valued).
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*18. Anticipating harmful functional bias* • whenever possible, the potential of the system being used for harmful or illegal purposes is avoided, and that if the system can be used for unintended purposes, then consider potential implications of this likelihood and develop mitigation procedures in the event of potential ethical issues arising; • the system is not designed for bad purposes and attempt to eliminate, whenever possible, ways that they can be misused (one way to do this is to use tried-and-tested general models, rather than building all models from scratch).
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*19. Avoiding harmful automation bias* • an appropriate level of human control for the system (by including respective task allocations between the system and humans for meaningful interactions and appropriate human oversight and control); • safeguards are embedded to prevent overconfidence in or overreliance on the system through education and training to be more aware of harmful bias in the system.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*20. Accessibility and Usability* • the system is understandable and accessible to users of assistive technologies, users with special needs or disabilities, or groups otherwise at risk of exclusion; • the system is usable by users of assistive technologies, users with special needs or disabilities, or groups otherwise at risk of exclusion (or if the system cannot be used properly, attempt to make improvements and ensure that any limitations are fully understood by these groups); • you seek feedback from teams or groups that represent different backgrounds and experiences (including but not limited to users of assistive technologies, users with special needs, or disabilities), and that this process should be accommodating to include different variations and users;

	• no persons or groups are disproportionately negatively affected by the system. Or if that cannot be ensured, then attempt to minimize the negative effects and ensure that these people and groups fully understand these negative effects before using the system, and that those at risk of being negatively affected are adequately represented in the design process by including feedback from those likely to be affected in the design of the system.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*21. Review process* • knowledgeable professionals, both internal and external to the company, examine the development process and the product through a risk assessment procedure.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*22. Whistleblowing* • 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.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*23. Diversity* • a process to include the participation of different stakeholders in the development, use, and review of the system; • that efforts are made so that a wide diversity of the public, including different sexes, ages, and ethnicities, are represented; • that this is applied within the organization, by informing and involving impacted workers and their representatives in advance.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*24. Inclusion* • an adequate inclusion of diverse viewpoints during the development of the system; • that development is based on an acknowledgement that different cultures may respond differently, have different thought processes and patterns, and express themselves differently.
ACCOUNTABILITY	*25. Engagement and reporting* • incidents are identified and reported on a correct and timely basis and implement appropriate internal and external escalation paths; • incidents are responded to and resolved immediately; • a culture of proactive problem management (detection, action and prevention), with clearly defined roles and responsibilities, is established and encouraged; • a transparent and open environment for reporting problems is established and encouraged, by providing independent reporting mechanisms and/or rewarding people who bring problems forward; • there is an awareness of the importance of an effective control environment;

	• a proactive risk- and self-aware culture is established and encouraged, including commitment to self-assessment, continuous learning, and independent assurance reviews; • auditability is built into the system; • performance indications are identified and regularly report on the outcomes, in relation to the auditing system.
ACCOUNTABILITY	*26. Compliance as culture* • a compliance-aware culture is established and encouraged, including disciplinary procedures for noncompliance with legal and regulatory requirements; • a culture that embraces internal audit, assurance findings, and recommendations (based on root cause analysis) is established and encouraged; • leaders take responsibility to ensure that internal audit and assurance are involved in strategic initiatives and recognize the need for (and value of) audit and assurance reports; • mechanisms that facilitate the system's auditability (such as ensuring traceability and logging of the robotic system's processes and outcomes); • in applications affecting fundamental rights (including safety-critical applications) the system can be audited independently; • the developing team attempts to learn to avoid situations requiring accountability in the first place, by ensuring ethical best practices.
ACCOUNTABILITY	*27. Code of ethics* • 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.
ACCOUNTABILITY	*28. Avoiding automation bias* • an appropriate level of human control for the system and use case, including respective task allocations between the system and humans for meaningful interactions and appropriate human oversight and control; • safeguards are embedded to prevent overconfidence in or overreliance on the system for work processes.
ACCOUNTABILITY	*29. Responsibility* • 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 robot autonomy; • there is a human-in-the-loop to control the system, to ensure and protect the autonomy of human beings;

	• detection and response mechanisms are appropriate in the event of something going wrong.

V-Model: Concept of operations phase

VALUE	**REQUIREMENT** **IN THE OPERATIONS PHASE, ASSESS AND ENSURE THAT:**
HUMAN AGENCY, LIBERTY AND DIGNITY	*1. Potential for impact on autonomy* • evaluation of the end-users' awareness about how the system may impact their autonomy is performed to determine if it is appropriate to make people aware of this impact, and if so, then ensure their awareness (e.g., if an end-user is using the system in a medical capacity, you need to ensure that the functionality of the system and the context in which it is used does not undermine their informed consent to any treatment options); • the system does not harm individuals' autonomy (i.e., the freedom and ability to make one's own goals and influence the outcomes of those decisions); • any interference the system has with the stakeholders' decision-making process (e.g., by recommending actions, decisions, or by how it presents stakeholders with options) is justified and minimised.
TRANSPARENCY	*2. Communication regarding interactions with the system* • if the robot is an agent, it is communicated to, and presumably understood by, the (end-)users or other affected persons that they are interacting with a non-human agent and/or that a decision, content, advice or outcome is the result of an algorithmic decision, in situations where not doing so would be deceptive, misleading, or harmful to the user.
TECHNICAL ROBUSTNESS AND SAFETY	*3. Communication with stakeholders* • a culture is established and encouraged in which open and structured communication is provided to stakeholders, in line with their requirements (including organisational users and end-users, if you are dealing directly with them); • information to stakeholders, (end-)users, and other affected persons, about the system's capabilities and limitations, is communicated in a clear, understandable, and proactive manner, that enables realistic expectation setting; • it is clear to stakeholders, (end-)users, and other affected persons, what the purpose of the system is and who or what may benefit from the product/service; • usage scenarios for the product are specified and clearly communicated so that they are understandable and appropriate for the intended audience; • in cases where stakeholders cannot be provided with certain data and answers, there is a full disclosure of that limitation, why there is a limitation, and also what they themselves do and do not know.

DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*4. Bias assessment in concept of operations* • the potential for harmful bias in the concept of operations phase is evaluated and, if possible, avoided (e.g., some expressed goals may inadvertently favour particular groups in society over others); • developing teams receive unconscious bias training to assist developers to identify innate biases during the development of systems.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*5. Environmental impact* • a mechanism to measure the ecological impact of the system's use (e.g., the e-waste produced by robots at the end of their life-cycle). • where possible, measures to reduce the ecological impact of your system's life cycle; • an adherence to resource-efficiency, sustainable energy-promotion, the protection of the non-human living world around us, and the attempt to ensure biodiversity and the healthy functioning of ecosystems (in particular, decisions made by the system that will directly affect the non-human world around us need to be carefully factored in, with strong emphasis on the impact on these ecological externalities, through a holistic ecosystem-focused outlook); • transparency about ecological impact and, if possible, work with environmental protection organisations to ensure that the system is sustainable, and keep the ecological footprint proportionate to the intended benefit to humanity.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*6. Individual wellbeing assessment* • the proposed system is evaluated for its likely and potential impact on individual wellbeing (including consideration of the way in which the system will or could be used which may be detrimental to users or stakeholders). Particular care should be taken for detriments towards vulnerable groups through discussion with them, rather than assuming their needs. In addition, special attention must be paid to the potential impact on individuals' employment, skills, and the quality of work.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*7. Societal impact assessment* • the proposed 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).
ACCOUNTABILITY	*8. Reporting impacts* • a risk assessment is conducted, which takes into account different stakeholders that are (in)directly affected by the system and the likelihood of those impacts; • training and education is provided to help develop accountability practices (including teachings of the potential legal framework applicable to the system);

	• if possible, that an 'ethical robotics/AI review board' or a similar mechanism is established to discuss overall accountability and ethics practices, including potentially unclear grey areas; • processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks, or biases in the system, is established.
ACCOUNTABILITY	*9. Minimising negative impact* • a process for minimization of negative impacts (such as external guidance and/or an auditing processes to oversee ethics and accountability), in addition to internal initiatives; • that audit controls are built into the system to check performance, record decisions made about the purpose and functioning of the system (including reporting on the impacts in general, not just occurrences of negative impacts); • an attempt to predict the consequences/externalities of the system's processing.

V-Model: Requirements and architecture phase

VALUE	**REQUIREMENT** **IN THE REQUIREMENTS AND ARCHITECTURE PHASE, ASSESS AND ENSURE THAT:**
PRIVACY AND DATA GOVERNANCE	*1. Informed consent* • 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.

PRIVACY AND DATA GOVERNANCE	*2. Creation of new personal data* • If needed, further informed consent is acquired (or, if not, that there is an alternative legal basis as set out in Articles 6(1) and 9(2) of GDPR) for the creation of new personal or sensitive information/data (e.g., through estimation of missing data, the production of derived attributes and new records, data integration, or aggregation of data sets); • all newly created personal or sensitive information/data is given at least the same protection and attracts the same rights as previously collected or held personal or sensitive information/data.
PRIVACY AND DATA GOVERNANCE	*3. Subsequent collection and/or creation of new personal data* • no new personal information is or can be collected or created during regular use of the system, unless necessary (e.g., for the function of the system or realization of the business objectives); • if new personal information is collected or created, then limitations are properly imposed to protect individuals' privacy or sensitive information/data, and further informed consent is acquired, if needed.
PRIVACY AND DATA GOVERNANCE	*4. Data review and minimization* • consideration is given to develop the system or train the model with or without minimal use of potentially sensitive or personal data, and applied whenever possible (note that it is questionable whether any data is ever fully anonymized); • 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.
TRANSPARENCY	*5. Trade-offs* • trade-offs between explainability/transparency and best performance of the system are appropriately balanced based on the systems context of application (e.g., in healthcare the accuracy and performance of the system may be more important than its explainability; whereas, in policing, explainability is much more crucial to justify behaviours and outcomes of law enforcement; and in other areas, such as recruitment, both accuracy and explainability are similarly valued).
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*6. Bias assessment in requirements* • the potential for harmful bias in the requirements and architecture phase is evaluated and, if possible, avoided (e.g., some requirements may inadvertently favour particular groups in society over others); • developing teams receive unconscious bias training to assist developers to identify innate biases during the development of systems.

V-Model: Detailed design phase

VALUE	REQUIREMENT IN THE DESIGN PHASE, ASSESS AND ENSURE THAT:
PRIVACY AND DATA GOVERNANCE	*1. Informed consent* • 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.
PRIVACY AND DATA GOVERNANCE	*2. Creation of new personal data* • If needed, further informed consent is acquired (or, if not, that there is an alternative legal basis as set out in Articles 6(1) and 9(2) of GDPR) for the creation of new personal or sensitive information/data (e.g., through estimation of missing data, the production of derived attributes and new records, data integration, or aggregation of data sets); • all newly created personal or sensitive information/data is given at least the same protection and attracts the same rights as previously collected or held personal or sensitive information/data.
PRIVACY AND DATA GOVERNANCE	*3. Subsequent collection and/or creation of new personal data* • no new personal information is or can be collected or created during regular use of the system, unless necessary (e.g., for the function of the system or realization of the business objectives); • if new personal information is collected or created, then limitations are properly imposed to protect individuals' privacy or sensitive information/data, and further informed consent is acquired, if needed.
PRIVACY AND DATA GOVERNANCE	*4. Data review and minimization* • consideration is given to develop the system or train the model with or without minimal use of potentially sensitive or personal data, and applied whenever possible (note that it is questionable whether any data is ever fully anonymized); • 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.

PRIVACY AND DATA GOVERNANCE	*5. Availability of data* • 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); • there is an embedded process that allows individuals to remove their data from the system and/or correct errors in the data where these occur, and ensure that this process is available at any stage in the process (note that once data is correctly and fully anonymized it is no longer considered personal data, although there may be potential for re-identification through aggregation of data sets); • if previously anonymized data is re-identified, then these data are made available once more (note, however, that it is questionable whether any data is ever fully anonymized).
PRIVACY AND DATA GOVERNANCE	*6. Protection against re-identification* • 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).
TRANSPARENCY	*7. Traceability measures* • measurements to ensure traceability are established through the following methods: ○ Methods used for designing and developing robotic AI systems (rule-based AI systems: the method of programming or how the model was built; learning-based AI systems: the method of training the algorithm, including which data was gathered and selected, and how this occurred); ○ Methods used to test and validate robotic AI systems (rule-based AI systems: the scenarios or cases used in order to test and validate; learning-based model: information about the data used to test and validate); ○ Outcomes of the system (outcomes of or decisions taken by the system, as well as potential other decisions that would result from different cases, e.g., for other subgroups of users); ○ A series of technical methods to ensure traceability should be taken (such as encoding the metadata to extract and trace it when required). There should be a way of capturing where the data has come from, and the ability to construct how the different pieces of data relate to one another.
TRANSPARENCY	*8. Responsibility for Traceability* • there is a "human in control" when needed, especially when the system may cause harmful outcomes (e.g., a robot playing a game like chess, which may have no harmful outcomes, would not necessarily require a human in control, unless there was the potential for negative effects); • a balanced prioritisation for human control, related to the plausibility and/or severity of the outcome;

	• there are measures to enable audit and to remedy issues related to governing the system and allow organisations using your technology the ability to identify when there is an issue or harm, and the ability to prevent these issues from occurring, and stop it when these issues are identified; • there are appropriate remedial steps for detection and response mechanisms if something goes wrong, by closely liaison with the organisational user, or end-user.
TRANSPARENCY	*9. Training data* • if possible, you can analyse your training data, that your data is representative, and value aligned; • whenever possible, there is an ability to go back to each state the system has been in to determine or predict what the system would have done at time t and, whenever possible, determine which training data was used; • in the event of a system malfunction or harm resulting from the system, as much transparency as is possible of your training data is made available, without violating privacy, to the appropriate authorities.
TRANSPARENCY	*10. Explainable systems* • you know to what degree the decisions and outcomes made by the system can be understood, including whether you have access to the internal workflow of the model; • explainability is guaranteed (through technologies such as Explainable AI), when there is a greater emphasis within its use for explainability over performance, or when there is no trade-off between explainability and performance.
TRANSPARENCY	*11. Explanations of rationale* • whenever possible, the process of, and rationale behind, the choices made by the system are explainable upon request to an organisational user and/or auditing body in situations where there is a potential and/or existent harm; • the reasons for the collection and use of particular data sets are explainable upon request to organisational users and/or auditing bodies; • in situations where the system-development organisations provide these technologies directly to the end-user, there is redress and explanations of how the system arrived at those decisions, if there is harm caused to the end-user by the system's decisions; • decisions made about individuals are understandable in colloquial language terms for an ordinary (end-)user or stakeholder (e.g., 'You have been put into this category because of x, y, and z').
TRANSPARENCY	*12. Trade-offs* • trade-offs between explainability/transparency and best performance of the system are appropriately balanced based on the systems context of application (e.g., in healthcare the accuracy and performance of the system may be more important than its explainability; whereas, in policing, explainability is much

	more crucial to justify behaviours and outcomes of law enforcement; and in other areas, such as recruitment, both accuracy and explainability are similarly valued).

V-Model: Verification and validation phases

VALUE	**REQUIREMENT** **IN THE VERIFICATION AND VALIDATION PHASES, ASSESS AND ENSURE THAT:**
HUMAN AGENCY, LIBERTY AND DIGNITY	*1. Potential for impact on autonomy* • evaluation of the end-users' awareness about how the system may impact their autonomy is performed to determine if it is appropriate to make people aware of this impact, and if so, then ensure their awareness (e.g., if an end-user is using the system in a medical capacity, you need to ensure that the functionality of the system and the context in which it is used does not undermine their informed consent to any treatment options); • the system does not harm individuals' autonomy (i.e., the freedom and ability to make one's own goals and influence the outcomes of those decisions); • any interference the system has with the stakeholders' decision-making process (e.g., by recommending actions, decisions, or by how it presents stakeholders with options) is justified and minimised.
PRIVACY AND DATA GOVERNANCE	*2. Privacy awareness* • mechanisms allowing developers and users to flag issues related to privacy or data protection in the system's processes of data collection (including for training and operation) and data processing; • mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable).
TRANSPARENCY	*3. Responsibility for Traceability* • there is a "human in control" when needed, especially when the system may cause harmful outcomes (e.g., an AI playing a game like chess, which may have no harmful outcomes, would not necessarily require a human in control, unless there was the potential for negative effects); • a balanced prioritisation for human control, related to the plausibility and/or severity of the outcome; • there are measures to enable audit and to remedy issues related to governing the system and allow organisations using your technology the ability to identify when there is an issue or harm, and the ability to prevent these issues from occurring, and stop it when these issues are identified; • there are appropriate remedial steps for detection and response mechanisms if something goes wrong, by closely liaison with the organisational user, or end-user.
TRANSPARENCY	*4. Communication regarding interactions with the system*

	• if the robot is an agent, it is communicated to, and presumably understood by, the (end-)users or other affected persons that they are interacting with a non-human agent and/or that a decision, content, advice or outcome is the result of an algorithmic decision, in situations where not doing so would be deceptive, misleading, or harmful to the user.
TRANSPARENCY	*5. Communication with stakeholders* • a culture is established and encouraged in which open and structured communication is provided to stakeholders, in line with their requirements (including organisational users and end-users, if you are dealing directly with them); • information to stakeholders, (end-)users, and other affected persons, about the system's capabilities and limitations, is communicated in a clear, understandable, and proactive manner, that enables realistic expectation setting; • it is clear to stakeholders, (end-)users, and other affected persons, what the purpose of the system is and who or what may benefit from the product/service; • usage scenarios for the product are specified and clearly communicated so that they are understandable and appropriate for the intended audience; • in cases where stakeholders cannot be provided with certain data and answers, there is a full disclosure of that limitation, why there is a limitation, and also what they themselves do and do not know.
TRANSPARENCY	*6. Communication within user and stakeholder community* • a culture is established and encouraged based on mutual trust, transparent communication, open and understandable terms, a common language, ownership, and accountability; • an explanation, which all reasonable users and stakeholders can presumably understand, is given as to why the system took a certain choice resulting in a certain outcome; • mechanisms are established to inform organisational users and end-users (if dealing directly with them) about the reasons and criteria behind the system's outcomes and, in collaboration with users, establish processes that consider users' feedback and use this to adapt the system; • any potential or perceived risks are clearly communicated to the (end-)user (e.g., consider human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue).
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*7. Engagement with users to identify harmful bias* • a mechanism allows others to flag issues related to harmful bias, discrimination, or poor performance of the system and establish clear steps and ways of communicating on how and to whom such issues can be raised (i.e., during the design, development, and deployment of the system); • there is transparency about how the algorithms may affect individuals to allow for effective stakeholder feedback and engagement;

	• the implementation of methods for redress and feedback from users at all stages of the system’s life-cycle.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*8. Decision variability* • a measurement or assessment mechanism, of the potential impact of decision variability on fundamental rights, is established based on an evaluation of the system’s possibility for decision variability that can occur under the same conditions; • variability is explained to the organisational user of the system and/or the end-user (if they are using it directly). For example, in medicine this should be explained to doctors that use it.
DIVERSITY, NON-DISCRIMINATION, AND FAIRNESS	*9. Distributing the system to organisational users* • the user interface is clearly presented, including information about potential errors and the accuracy of the system (including the underlying certainty).
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*10. Environmental impact* • a mechanism to measure the ecological impact of the system’s use (e.g., the energy used by data centres). • where possible, measures to reduce the ecological impact of your system’s life cycle; • an adherence to resource-efficiency, sustainable energy-promotion, the protection of the non-human living world around us, and the attempt to ensure biodiversity and the healthy functioning of ecosystems (in particular, decisions made by the system that will directly affect the non-human world around us need to be carefully factored in, with strong emphasis on the impact on these ecological externalities, through a holistic ecosystem-focused outlook); • transparency about ecological impact and, if possible, work with environmental protection organisations to ensure that the system is sustainable, and keep the ecological footprint proportionate to the intended benefit to humanity.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*11. Individual wellbeing assessment* • the system is evaluated for its likely and potential impact on individual wellbeing (including consideration of the way in which the system will or could be used which may be detrimental to users or stakeholders). Particular care should be taken for detriments towards vulnerable groups through discussion with them, rather than assuming their needs.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*12. Emotional attachment* • if the system is developed to interact directly with humans, evaluate whether it encourages humans to develop unwanted attachment and unwanted empathy towards the system or detrimental addiction to the system, and if so take appropriate action to minimize such effects; • the system clearly communicates that its social interaction is simulated and that it lacks human capacities such as “understanding” and “feelings”;

	• the system does not make humans believe it has consciousness (e.g., through expressions that simulate emotions).
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*13. Societal impact assessment* • the system’s likely and potential impact on social relationships and social cohesion (including consideration of the way in which the system will or could be used which may be detrimental to groups of users or groups of stakeholders) is not inappropriate; • social benefits are determined through social metrics, not simply measurements in terms of GDP (e.g., liveability indexes).
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*14. Engagement with stakeholder community* • the broader societal impact of the robotic system’s use beyond the individual (end-)users (such as potentially indirectly affected stakeholders) is evaluated; • the social impacts of the system are well understood (e.g., assess whether there is a risk of job loss, deskilling of the workforce, or changes to occupational structure) and record any steps taken to counteract such risks; • a culture is established and encouraged to ensure timely communication of IT change requests to affected groups, and consult the affected groups regarding implementation and testing of changes; • stakeholders are involved throughout the system’s life cycle, and foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.
INDIVIDUAL, SOCIETAL, AND ENVIRONMENTAL WELL-BEING	*15. Mitigation of impacts on democracy* • an evaluation of whether the system is intended, or could be used for, supporting, organizing or influencing political processes, including political messaging and communication, and if so, take measures to ensure that the system supports democratic processes and protects against interventions that manipulates, misleads or excludes voters and distorts democratic processes; • compliance with higher authorities of robotics and AI development and implement an ethical officer to ensure corporate social responsibility within the company; • that external ethics audits are carried out to guarantee that system development is not harming democratic processes.
ACCOUNTABILITY	*16. Reporting Impacts* • a risk assessment is conducted, which takes into account different stakeholders that are (in)directly affected by the system and the likelihood of those impacts; • training and education is provided to help develop accountability practices (including teachings of the potential legal framework applicable to the system); • if possible, that an ‘ethical robotics/AI review board’ or a similar mechanism is established to discuss overall accountability and ethics practices, including potentially unclear grey areas;

	• processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks, or biases in the system, is established.
ACCOUNTABILITY	*17. Minimizing negative impact* • a process for minimization of negative impacts (such as external guidance and/or an auditing processes to oversee ethics and accountability), in addition to internal initiatives; • that audit controls are built into the system to check performance, record decisions made about the purpose and functioning of the system (including reporting on the impacts in general, not just occurrences of negative impacts); • an attempt to predict the consequences/externalities of the system's processing.
ACCOUNTABILITY	*18. Impact on business* • that there is an ability to evaluate the degree to which the system's decision influences the organisation's decision-making processes, why this particular system was deployed in this specific area, and how the system creates value for the organization and the general public; • a clear rationale is established by your organization about why you are designing and creating the system, and the intended purpose that it will serve.
ACCOUNTABILITY	*19. Identify interests and values at risk* • a mechanism to identify relevant interests and values implicated by the system and potential trade-offs between them, before deployment and during the life-cycle of the system, which should include considerations regarding how trade-offs were decided and documented; • the establishment of values and interests at risk, through stakeholder analysis, product testing, discussion groups, external workshops, and a range of diversity and inclusion sessions.
ACCOUNTABILITY	*20. Install systems to allow for internal complaint* • the existence and advertisement (through the companies) of a clear complaints and whistleblowing system (directing employees to a suitable contact venue and setting out the process for registering both anonymous and identifiable complaints); • that employees are aware of a zero-tolerance policy for any recriminations for whistleblowing or the registering of internal complaints.
ACCOUNTABILITY	*21. Internal Auditor* • the internal auditor(s) within the company is audited to guarantee that it is not abusing their role within the organisation; an internal ethics advisor has the same degree of independence and security as is now envisaged for the DPO under GDPR. Alternatively (or in addition) we encourage organisations to develop sectoral solutions (e.g., an ethics council for their sector; start-ups and microbusinesses may not have the resources to put an ethicist on the payroll, so

	an alternative, such as Ethics-as-a-Service or external ethics auditing, may be implemented instead).
ACCOUNTABILITY	*22. Redress mechanisms* • the contextual meaning of accountability is clear for different roles in the development chain (e.g., data scientists, CDOs, board members, business managers), including what form of sanctions are in place for whom, and which roles should take personal responsibility, with redress mechanisms in case of negative impacts; • a set of mechanisms that allows for redress in case the occurrence of any harm or adverse impact is established; • where possible, embed mechanisms to provide information to (end-)users/third parties about opportunities for redress.

V-Model: Special topics (applicable to all phases)[11]

This subsection presents ethical requirements in relation to specific types of robots, robotic functions or techniques, and robot application areas. These requirements are applicable to most (if not all) phases of the V-Model, and they complimentary to the requirements provided in the previous sections of this annex.

TYPE OF ROBOT	**REQUIREMENT**
SOCIAL ROBOT	*1. Robots that can recognize or express emotions* • When robots 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 robots express emotions, there should be pre-emptive statements that one is interacting with a robot and there should be built-in distinguishability from humans.

[11] For additional special topics, see the specific ethical issues in relation to particular robotics techniques, products and application domains discussed in Jansen et al. (2019).

SOCIAL ROBOT	*2. Covert and deceptive robots* • Human beings should always know if they are directly interacting with another human being or a machine. It is the responsibility of robotics 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 a robotic system (for instance, by issuing clear and transparent disclaimers); • The use of deceptive robots 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.
DEFENCE ROBOT	*3. Robots designed for defence purposes* • For new, robotic 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; • Robotic 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 robotic 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; • Robotic technologies for surveillance 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.
ETHICALLY AWARE ROBOT	*4. Ethically aware robots* • 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 robotic systems are ethically aware and what this implies; • Ethics should be in line with the culture in which it is embedded; • Compliance certification (external) and internal audit should be ensured.

References

13th Annual State of Agile Survey. (2019, May 7). Retrieved from https://www.stateofagile.com/

Alix (2018, May 7). *Working Ethically At Speed*. Retrieved from https://medium.com/@alixtrot/working-ethically-at-speed-4534358e7eed

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., … Thomas, D. (2001). *Manifesto for Agile Software Development*. Retrieved from http://agilemanifesto.org/

Brey, P., Lundgren, B., Macnish, K. and Ryan, M. (2019). *Guidelines for the development and use of SIS.* Deliverable D3.2 of the SHERPA project. https://doi.org/10.21253/DMU.11316833.

CEN (2017). *Ethics assessment for research and innovation - Part 2: Ethical impact assessment framework.* CEN workshop agreement, CWA 17145-2.

Fitzgerald, B., Hartnett, G., & Conboy, K. (2006). Customising agile methods to software practices at Intel Shannon. *European Journal of Information Systems*, 15(2), 200-213.

Friedman, B., Kahn, P. and Borning, A. (2006). 'Value Sensitive Design and Information Systems,' in *Human-Computer Interaction in Management Information Systems: Foundations* (eds. P. Zhang and D. Galletta). Armonk, NY: M.E. Sharpe.

Gonçalves, L. (2019, May 1). *What Is Agile Methodology.* Retrieved from https://luis-goncalves.com/what-is-agile-methodology/

HLEG-AI (High-Level Expert Group on Artificial Intelligence) (2019). *Ethics Guidelines for Trustworthy AI.* Downloaded on 8-3-2020 at https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

Hoven, Jeroen van den, Pieter E. Vermaas, and Ibo van de Poel, eds. (2015). *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*. Springer Netherlands.

Huldtgren, A. (2015). Design for Values in ICT. In *Handbook of ethics and values in technological design. Sources, Theory, Values and Application Domains* (eds. J. van den Hoven, P. Vermaas, and I. van de Poel, eds). Springer.

IEEE (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems) (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. IEEE. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/

Jansen, P., Brey, P., Fox, A., Maas. J., Hillas, B., Wagner, N., Smith, P., Oluoch, I., Lamers, L, Van Gein, H., Resseguier, A., Rodrigues, R., Wright, D. and Douglas, D. (2019). *Ethical Analysis of AI and Robotics Technologies.* D4.4 of the SIENNA project. https://www.sienna-project.eu/publications/.

Kneuper, R. (2018). *Software Processes and Life Cycle Models: An Introduction to Modelling.* Cham, Switzerland: Springer Nature Switzerland.

Liversidge, E. (2015). The Death Of The V-Model, *Harmonic Software Systems*, June 25, 2015. http://harmonicss.co.uk/project/the-death-of-the-v-model/.

Ministry of Science and Technology of China (2019). *Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence*. A translation can be found at: https://perma.cc/V9FL-H6J7.

OECD (2019). *Recommendation of the Council on Artificial Intelligence.* Retrieved on 8-3-2020 at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Ryan, M., Philip Brey, Kevin Macnish, Tally Hatzakis, Owen King, Jonne Maas, Ruben Haasjes, Ana Fernandez, Sebastiano Martorana, Isaac Oluoch, Selen Eren, and Roxanne Van Der Puil (2019). *Ethical Tensions and Social Impacts.* Deliverable D 1.4 of the SHERPA project. https://doi.org/10.21253/DMU.8397134

Wellens, K. (2008). The Seductive and Dangerous V-Model, *Testing Experience* magazine, December 2008. Expanded version of article available at http://www.clarotesting.com/page11.htm.