

Grant Agreement N°825619

AI4EU Deliverable D5.4

AI4EU Value-based Design Methodology

WP	5	Promoting European ethical, legal, cultural and socio-economic values for AI
Task	5.3	The AI4EU value-based design methodology

Dissemination level¹	PU	Due delivery date	31/12/2020
Nature²	R	Actual delivery date	08/01/2021

Lead beneficiary	UMU
Contributing beneficiaries	BSC

Document Version	Date	Author	Comments
1.0	24/05/2020	Zahoor UI Islam	Survey analysis
1.1	03/12/2020	Andreas Theodorou, Juan Carlos Nieves, Virginia Dignum	Draft version for reviewing
1.2	05/01/2021	Andreas Theodorou, Juan Carlos Nieves, Virginia Dignum, Ulises Cortes	Updated version following peer review
1.3	07/01/2021	Finalized version after quality review	

¹ Dissemination level: **PU** = Public, **PP** = Restricted to other programme participants (including the JU), **RE** = Restricted to a group specified by the consortium (including the JU), **CO** = Confidential, only for members of the consortium (including the JU)

² Nature of the deliverable: **R** = Report, **P** = Prototype, **D** = Demonstrator, **O** = Other

Deliverable abstract

This report aims to provide a framework for AI Assessment for Trustworthy AI based on a study of existing approaches and requirements set by the consortium and the literature.

All of our investigatory activities indicate the same conclusion: while ethical guidelines are well received by industry and academia, there is a lack of effective means of assessing a system's compliance to those guidelines. As we discuss in detail in Section 5, existing assessment lists are high-level generic and far from being actionable.

Understanding how AI systems are more than just the sum of their software components [3], [4], but rather they comprise the socio-technical system around them, we introduce a new assessment framework to aid organisations on performing AI assessment on compliance with Ethical Legal Socio-Economic Cultural (ELSEC) issues. Moreover, we offer the necessary tools to enable those organisations to audit their systems.

Contents

1. Introduction	4
1.1 Document Context.....	4
1.2 Objective.....	4
2. Glossary	5
2.1 AI-specific Terminology	5
2.2 Software Engineering.....	5
2.3 Policy Terms.....	6
2.4 Stakeholders	6
3. Survey on Software Engineering Practices within the AI4EU Consortium	8
3.1. Survey Objectives and Design.....	8
3.2. Results Analysis	9
3.3 Discussion	10
4. Current Practices in the Literature	11
5. AI Assessment	14
6. Framework for Trustworthy AI Evaluation.....	15
6.1 The use of a Design for Values approach.....	15
6.2 Overall Assessment Framework	17
7. ELSEC Evaluation Tools	19
7.1 Tool for Self-Evaluation.....	19
7.2 Abbreviated Assessment list for AI applications in the AI4EU catalogue.....	19
8. Conclusion	23
9. References.....	24
10. Appendix A: Survey of AI4EU.....	26
11. Appendix B.....	40
12. Appendix C: The Responsible AI Canvas.....	42
13. Appendix D.....	43

1. Introduction

1.1 Document Context

This document is part of the work package 5 “Promoting European ethical, legal, cultural and socio-economic values for Artificial Intelligence”. It is titled “The AI4EU value-based design methodology” and reports on the activities of Task T5.3, “The AI4EU value-based design methodology”. It has a relationship to Deliverable 5.6, which provides a qualitative evaluation of the practices and assessment requirements of 50 organisations across Europe [2]. This document is organised as follows:

- Section 2 ensures a common vocabulary with the reader by providing definitions to the various terms throughout the report.
- Section 3 contains a summary of key findings from a survey of the consortium on their existing practices. It explores considerations, practices, and processes that the different partners of the AI4EU project currently have in-place during the different phases of design, development, and deployment of AI systems.
- Section 4 is a survey of the existing academic literature on how existing AI-development methodologies consider issues related to Trustworthy AI.
- Section 5 is a mini-review of the current state of AI Assessment based on publicly accessible documents and AI4EU’s Deliverable 5.6.
- Section 6 describes a framework for continuous checking of the adherence to ethical socio-legal values, such as the one required by Task 5.3.
- Section 7 introduces self-assessment and compliance checking tools, as suggested in the deep dive piloting of the HLEG Trustworthy AI guidelines, reported in Section 2 and Deliverable 5.6.
- Section 8 concludes the report by providing a summary of the document and its contributions.

1.2 Objective

This report aims to provide a framework for AI Assessment for Trustworthy AI based on a study of existing approaches and requirements set by the consortium and the literature.

During the project, we conducted two independent surveys: one quantitative survey of the project partners, discussed in Chapter 2, and another of the existing literature –a summary of which can be found in Chapter 3. In addition, we led the qualitative evaluation of the “Assessment List for Trustworthy AI”, reported in Deliverable 5.6.

All of our investigatory activities indicate the same conclusion: while ethical guidelines are well received by industry and academia, there is a lack of effective means of assessing a system’s compliance to those guidelines. As we discuss in detail in Section 5, existing assessment lists are high-level generic and far from being actionable.

Understanding how AI systems are more than just the sum of their software components [3], [4], but rather they comprise the socio-technical system around them, we introduce a new assessment framework to help organisations on performing AI assessment on compliance with Ethical Legal Socio-Economic Cultural (ELSEC) issues. Moreover, we offer the necessary tools to enable those organisations to audit their systems.

2. Glossary

Some of the greatest challenges of appropriately regulating AI are social rather than technical [5] especially as we cannot even agree on the definition of terms. In line with the overwhelming request to establish a common vocabulary and avoid drawn out semantic debates and minimise risk of adverse outcomes due to misunderstanding, widely used terms are defined here.

2.1 AI-specific Terminology

Agency refers to the ability of an entity to effect change; through the usage of sensors, an agent can perceive its environment and internal state and using actuators to change them.

Intelligence implies the *selection of the right action at the right time*.

Intelligent agents/systems are a special category of entities characterised by their ability to select rationally and as optimally as possible their actions.

Artificial Intelligence refers either to either to the human-made action-selection systems that exhibit signs of intelligence or a field of research and development of intelligent agents. When referring to the former, we use it in the form of *AI system*. There is no such thing as ‘an AI.’

2.2 Software Engineering

Software Engineering (SE) is defined as “a systematic, disciplined, quantifiable approach to the analysis, design, assessment, implementation, test, operation, maintenance and development of software, that is, the application of engineering to software” [8]. In a traditional SE approach, several design methods and models for the software life cycle are defined, and many methodologies for the definition and assessment of the different phases of a life-cycle model are applied [9, 10]. SE should not be considered an isolated activity, but rather it is part of a broader process that involves management tasks, stakeholders’ engagement, and technological solutions.

System refers to the combination of the feasible output, i.e. the deliverable product or service, and the processes used for its design, development, and deployment alongside with the regulatory framework that guides said processes. We only refer to the software—or the combination of hardware and software—explicitly, e.g. *intelligent system*.

Development Methodology is the process of dividing development work into distinct phases to improve design, product management, and project management. It is also known as a system’s life cycle.

Waterfall is a traditional development methodology. The process begins by working on the analysis of requirements, designing architectures, developing the system, and subsequently testing the system before the final deployment. The method is well planned and tends to be well documented. Traditionally, this leads to major releases of versions with a considerable number of new features. The key stages of the waterfall model are closely attributed to the fundamental development activities of software projects.

Agile approaches are designed to adapt to evolving needs, minimise development costs and provide applications of acceptable quality. Agile projects are characterised by multiple incremental releases, each created within a very short timeframe. Usually, all members of the team are involved in the process of preparation, implementation, and testing. Small teams usually communicate face-to-face with each other regularly. Scrum, introduced below, is the most common implementation of agile principles.

SCRUM is an agile methodology in which software features are added to short iterations, call sprints. SCRUM is known for its short, and frequent meetings that keep people focused. Tasks are normally backed up on a scrum board. The group is self-organising and collaboratively managed, although a scrum master must enforce the rules and buffer the team from outside distractions.

Extreme Programming (XP) is a frequent release development methodology in which developers work in pairs for continuous code review. This gives very robust, high-quality software at the expense of twice the development cost. There is a strong emphasis on test-driven development.

Feature-Driven Development (FDD) is an iterative development process with more emphasis on planning out the overall architecture, followed by implementing features in a logical order.

2.3 Policy Terms

Standards They are consensus-based agreed-upon ways of doing things by providing what they consider to be the minimum universally acknowledged good practices. They formalise design guidelines, technical specifications, and even ethical principles into a structure that could be used to either guide or evaluate the level of compliance a company or system has against standards [1], [6]. Compliance provides confidence in a system's efficacy in areas important to users, such as safety, security, and reliability. Most standards are considered *soft governance*, i.e. non-legally mandatory policy to follow. Yet, it is often in the best interest of companies to follow them to demonstrate due diligence and limit their legal liability in case of an incident. Moreover, standards can ensure user-friendly integration between products. In fact, various standards, e.g. USB, were developed through collaborative work of multiple large corporations.

Legislation deals with the enforcement of standards and provides policy to ensure attribution and distribution of responsibility and accountability in case of an incident. Products and services may be required by legislation to follow specific standards to operate, provided, or be sold in a country [6]. In other words, legislation can make standards *hard governance*.

Ethical Guidelines (or Principles) is the framework for examining moral values. Ethical principles that may underpin standards either explicitly or implicitly or produces laws to ensure the enforcement of those values. On their own, they are *soft governance*, and it should be clear that ethical compliance does not imply legal compliance nor substitutes it [7]. Instead, ethical compliance should be complementary to legal compliance.

Internal Policy is the policy used within an organisation—or collection of organisations—to regulate the development and deployment of any product or service they provide. The policy *should* adhere at least to national-level legislation, but it can also include---or at least be informed by---soft governance, i.e., ethical guidelines, and standards. The policy is context-specific itself; different products/services within the same organisation may adhere to different policies. It ensures quality assurance and provides the means of demonstrating due diligence—including its existence being part of an organisation's due diligence.

2.4 Stakeholders

We consider all people affected either by the development or the deployment of the system as stakeholders. Responsibility includes not only the person who writes the code or “presses the deploy button”, but everyone involved in the system's life cycle. This may even include researchers, like ourselves, whose publications influenced the system. While a systematic review of all possible actors is far beyond this report's scope, our proposed evaluation methodology, and for the purposes of our work, we grouped different stakeholders into 6 high-level groups.

Policymakers are the persons responsible for or involved in formulating the system's formal policy and not the legislators or crafters of the external guidelines or standards.

Development Team consists of the actual developers, i.e., people who write the code for the system. Moreover, quality assurance, like continuous integration testers, are also part of this team. They both work towards the delivery of a product or service for third-party consumption.

Management persons are responsible for various decisions related to the status of the system, e.g., product and project managers. Such decisions include approving deployment, features to be included in the system, or even the acceptable risks that could arise by the development and deployment of the system.

Expert Users Not all external stakeholders are end-users. Many autonomous systems will be subject to certification. Such processes will be provided by agencies external to the creators of the system. Certification might be a legal requirement, or it might be a voluntary scheme providing some mark as a guarantor of quality [6]. Systems when they go wrong, either by accident or misuse, can cause harm or injury. Incident investigators may be required to carry systematic formal inquiries into the causes of the accidents. They may require not only application-domain knowledge but also access to company-held information regarding the system.

End Users is a person or organisation, i.e. a legal person who voluntarily and purposely engages with the system. We identified two business models, *business-to-business* and *business-to-consumer*, that determine who the end-user is. In a business-to-consumer model, the end-user can be any individual who directly operates or uses the system, e.g. a passenger of an autonomous vehicle. In a business-to-business model, the end-user can be another organisation that directly operates the system to provide a service to others or to produce goods for consumption.

General Public Members of the wider society, who do not directly engage with the system, but can nevertheless be affected directly or indirectly by it. An example is members of the public who are not using a company's service, but are tracked by it.

3. Survey on Software Engineering Practices within the AI4EU Consortium

At the start of AI4EU, we conducted a survey to gain insights on what are the existing practices that the different partners of AI4EU project employ in regards to the software development methodology (SDM) and what do they expect to be included in the future SDM for developing, deploying and using trustworthy AI systems. This survey reports on current practices, based on our results gathered in April – June 2019, in different application areas, and analyses engineering practices and formal design methodologies used in developing AI systems in industrial and academic settings in which AI4EU partners operate. The survey contributes to define a basis for Deliverable 5.4 in terms of needs and requirements for an ethically aligned SDM. In this section, we discuss the survey design, key results, and reflections. The full survey and its results can be found in Appendix A.

3.1. Survey Objectives and Design

Software development lifecycle is the name given to systematic approaches used to design, develop, test, and update software systems [8]. The software industry has developed and uses a plethora of different approaches, referred to as *software engineering methodologies* (e.g. *waterfall*, *agile*, etc.), to deliver and maintain software. All of these different methodologies share some commonalities: i) they are all composed of a number of clearly defined and distinct work phases; ii) they all provide explicit instructions when each of these phases starts and ends, and iii) they all offers guidance on how to manage resources during said phase. Yet, each of these different methodologies offers a different order of conducting those work phases. Depending on the approach's focus, e.g. XP aims for the rapid development of a prototype, methodologies place different emphasis at each stage of a system's lifecycle. Considering the particularities of AI systems compare to traditional software, any survey into existing practices should explore the methodology most used by its participants.

In addition to lifecycle methodologies, various architectural and code design patterns have been established and adopted by the software industry [9], [10]. Such practices provide generalised guidance to common technical problems encountered by software development teams. They aim to provide high-level reusable instructions on how to structure and connect the different system's components or even blocks of code. Moreover, in professional software development, clean-code practices, aiming to improve the quality of the final deliverables further, have been gaining momentum [11]. In AI, we have multiple examples of architectures and practices, discussed in detail in Section 3, designed explicitly to develop autonomous systems. AI4EU consortium members include multiple software development companies. A goal of our survey is to see if: i) 'traditional' software practices are used –and are adequate– for AI systems development; ii) awareness for AI-specific practices; iii) rational behind the use–or not–of architectural and design patterns.

To achieve our objectives, we designed an online survey with 46 questions, seen in Appendix A, covering the complete life cycle: from the requirements gathering and analysis to the system's design and implementation to its quality assurance (testing), auditing, and compliance to its final deployment and maintenance. To better cluster our results, we also asked the responders for information regarding their organisations, e.g. type of organisation (academia, industry, NGO), size (start-up, small, medium, or large organisations). We further asked for information directly related to the responders' background: i.e. if they have a technical or managerial background, if they worked with AI before, and their current position within the organisation. We advertised the survey to the AI4EU partners through emails to all partners and at the Joint WP5/6/7/9 workshop on May 2 and 3, 2019 at Sorbonne Université Paris.

3.2. Results Analysis

We have received answers from 18 participants. We clustered our results into two categories: *academia* and *industry*. Due to the small number of participants, we did not cluster further based on any other demographic. Luckily, we had an even split with 9 answers from participants affiliated with universities and academic research institute (academia) and 9 from industry, including large-sized companies, small-sized companies, start-ups and Non-Governmental Organisations (NGO's). Participants reported various positions researchers, AI developers, data analysts, software developers, and project and product managers on AI-related products and services.

The survey results, found in full in Appendix A, indicate both commonalities and differences between academic and industry practices in the development of AI systems. Amongst our findings are the following:

- Both industry and academia agree on the need for standardised definitions for both technical and ethical socio-legal terms and values.
- Scrum is the most used methodology, used by both industry (67%) and academia (44%).
- In terms of ethical guidelines to be included in the methodology: industry's top requirements are documentation (89%) and standardised reporting (78%). At the same time, academia stated Challenger benchmarks (Production model benchmarking against other similar systems or purpose-made challenges, e.g. ImageNet) (89%) and Adversarial example testing (Systematically testing the model's behaviour on developer generated potential extreme / rare instances) (89%).
- In terms of the documentation contents, academia stated High-level description of the decision-making system used, e.g. methods (100%), Identified risks (100%). In contrast, industry stated Requirements (89%), High-level description of the decision-making system used, e.g. methods (89%), and System boundaries (78%).
- From the survey, it appeared that no AI-specific architecture is being used in the industry. The methodology to make an explicit recommendation on what architecture system should have, both academia and industry results stated "No" with (56%) average. In terms of methodology to integrate a specific architecture into the methodology, academia (78%) and industry (67%) stated "No".
- Both academia and industry stated that methodology should include recommendations and instructions on why, how and, when to use version control with (66%) and (78%) acceptance, respectively.
- All respondents agreed on limiting system performance to ensure the legal compliance of a system. However, a lower number of participants from academia would do the same for ethical compliance (89%) –unlike industry, where 100% of the respondents agreed on the need to do so.
- In terms of auditing, academia (67%) and industry (66%) are auditing mostly for legal compliance. In terms of auditing projects compliance with organisational policy academia stated (56%) acceptance ratio on requirements gathering while industry stated during development with (56%) acceptance ratio.
- The use of tools for verification and validation is non-existent in both academia (78%) and industrial (89%) practices.

- In terms of keeping internal documentation for data governance in academia (56%) stated “Yes” while in the industry (56%) stated “No”.
- In terms of methodology to require specific standards to follow in academia (78%) stated “No” while in the industry (56%) stated “Yes”. In terms of the methodology to include procedures of software updates and provide a specification for testing both academia (89%) and industry (100%) stated “Yes” respectively.
- In terms of producing reports of the effects of products, services, research output both in academic and research environment is mostly done when relevant, in academia (78%) stated for environmental. In contrast, the industry (78%) opted for Risk of misuse.

3.3 Discussion

This survey provided indicative results on the current practices used by AI4EU members for the development of AI systems. The lack of any AI-specific architecture or methodology is also indicative of the requirements of the AI4EU consortium. We hypothesise that the reason is that existing software engineering methodologies used integrate with existing organisational practices and tools. Furthermore, in larger applications, ‘the AI part’ of a system might only be a small part compared to other ‘dumb’ components, e.g. networking services, user interface, etc., and teams working in the different components are required to use the same development methodology.

Both academic and industry-based participants appear to prefer the use of iterative development methodologies, e.g. *scrum*. These approaches focus on the quick implementation and testing of prototypes. In principle, the team takes care of quality assurance during sprints –i.e. iterations– and as part of daily stand-ups, sprint reviews, and retrospectives. A major drawback of agile methods is the often use of informal evaluation techniques, which often are insufficient or incompatible with quality-assurance practices mandated by standards often used by safety-critical systems or other applications in heavily regulated industries [12], [13]. Yet, the majority of our respondents agree on the need for standardisation and assessment efforts. Based on our results, participants are willing to limit a system’s performance to achieve any necessary ethical socio-legal compliance.

However, the survey suggests that an effective framework for checking effective verification and validation methods and tools is missing. Here, we can take input for the qualitative results presented in Deliverable 5.6 [2], where similar concerns have been raised: While there is consensus between our interviewees of the need for ethical considerations in a system’s development, deployment, and use, our 50 interviewees all raised the need for mechanisms for evaluating policy compliance and system’s compliance with any ethical policy.

Finally, we will like to emphasise that the survey highlighted some of the major processes that must be integrated into development methodology including effective management, communication, use of appropriate standards, sufficient documentation, a mechanism for version control, ethical and legal compliance mechanisms, the introduction of use cases, and finally the support of tools and libraries for policy evaluation and compliance mechanisms. Transparency seems to be a significant concern for both the industry and academia.

Ultimately, the survey’s outcome is that any methodological framework developed as part of WP5 needs to complement existing practices by offering much-needed tools for AI assessment.

4. Current Practices in the Literature

In parallel to the survey on the current practices, we conducted a systematic review of the literature for existing methodologies for AI systems development. These methodologies have facilitated the development of systems in various application domains, from robotics to games to recommendation systems, influencing developers and indirectly affecting society at large. In a comprehensive analysis reported in [14]³, we systematically explored how existing AI development methods include and support software development life cycle methodologies, stakeholder participation, and provide support for ethical considerations in terms of support for explainability, traceability, and transparency. The remainder of this section provides a summary of that analysis.

We evaluate the extent to which current AI development approaches cover aspects central to the development process's alignment towards trustworthy AI systems. These aspects are identified through a formal literature analysis of AI engineering methods applied to different AI implementation fields, including socio-technical systems, game AI, machine learning approaches, and virtual embodied agents. We assess these AI engineering approaches based on their application of software development life cycle methodologies, participation support, and trustworthiness criteria. The concept of surveying software engineering methods for AI outlined here provides a starting point for identifying, designing and implementation approaches to ensure that the future AI systems are robust, comply with the ethical requirements and established standards set by relevant authorities and policymakers.

Several development methods and engineering processes for AI systems development have been proposed in various AI application domains, including *Agent-Oriented Software Engineering* (AOSE) methodologies; *Knowledge-Based Systems* (KBS); *STS frameworks* with design methods for virtual agents; *Behaviour-based AI* (BBAI); and, ML methodologies. These methods share several standard steps: context understanding, knowledge acquisition, data curation, data modelling, optimise performance, production, and monitoring. We exclusively focused on architecture methods that have been applied in recent times. A list of the methods evaluated, citations, and short descriptions can be found in **Table 1**.

It is essential to understand a standard Software Development Life Cycle, stages, processes, methods, participation support, and trustworthiness and how they are adapted and practically applied by AI development methods. These concepts are used to develop criteria for evaluating the suitability of SDMs for existing AI development methods in different AI application domains.

³ The paper is under review; a preprint will be made publicly available after the reviewing process.

Table 1 Application domains, development methods, citations, and a short description of the methods reviewed.

Application Domain	Example Method	Citation	Short Description
Socio-Technical Systems	OperA	[15]	The OperA is a framework for analysis, design and simulation of organisational systems.
	EI/EIDE	[16]	The EI/EIDE is a framework based on formal meta-model (EI) for electronic institutions (EI), and a particular development environment (EIDE) for implementing EI-based models.
	JaCaMo	[17]	JaCaMo is a multi-agent-oriented programming execution platform that provides programming constructs that match the abstractions used at the levels of a multi-agent system for expressing social coordination.
	ABSS Methods	[18]	Agent-based social simulation is a computational modelling approach for social simulation.
Game AI	Behaviour Tree	[19]	Behaviour Tree is a structure that allows the organisation of behaviours in a hierarchical manner, decomposing an initially broad task in several sub-trees of reduced complexity. BT are a formalism used to describe tasks and their switching in a modular fashion.
	Goaloriented action planning (GOAP) F.E.A.R Orkin	[20]	Goal-Oriented Action Planning refers to a simplified STRIPS-like planning architecture specifically designed for real-time control of autonomous character behaviour in games.
Machine Learning	SEMMA	[21]	SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. SEMMA offers an easy process to understand, allowing Data Management projects to be developed and maintained in an organised and appropriate way. Through this way, it confers a framework for its design, development and growth, helping to offer approaches to business challenges as well as identifying business goals.
Virtual Embodied Agents	Behaviour-Oriented Design	[22]	BOD is a methodology for the development of AI that combines an iterative process of development, based on variable state. It emphasises producing flexible designs for extendable and sustainable AI projects.
	FAtiMA	[23]	FAtiMA is an Agent Architecture with planning capabilities designed to use emotions and personality to influence the agent's behaviour. During the last years, the architecture was used in several effective agents including robotics.

Table 2 Results of our analysis of the different methodologies on how much focus they place on the different stages of the life cycle and Trustworthy AI issues. ‘++’ indicates an emphasis on, ‘+’ denotes moderate focus (e.g. through an implicit reference and contribution to the criteria), and ‘n/a’ suggests that there is no explicit reference or otherwise focus.

Criteria	Methods								
	OperA	EI/EIDE	JaCaMo	ABSS	Behaviour Tree	Goap	SEMMA	BOD	FAtiMA
Life-Cycle									
Requirements	++	n/a	n/a	n/a	n/a	+	n/a	++	n/a
Design	++	+	++	+	+	+	n/a	++	++
Implementation	++	++	+	+	+	++	+	++	+
Testing	n/a	+	n/a	n/a	+	+	+	+	n/a
Run-time	+	+	+	n/a	+	+	+	+	n/a
Verification	++	+	n/a	n/a	+	n/a	n/a	n/a	n/a
Validation	++	+	n/a	n/a	+	n/a	++	n/a	n/a
Participation									
Criteria									
Real-world use	+	+	+	+	+	+	+	+	+
Support participants	+	+	+	+	n/a	n/a	n/a	+	n/a
Support adaptation	++	+	+	+	+	n/a	n/a	+	+
Trustworthiness									
Ethical Consideration	+	+	+	+	n/a	n/a	n/a	n/a	+
Explainability	n/a	n/a	n/a	n/a	n/a	n/a	+	n/a	n/a
Traceability	n/a	+	n/a	n/a	n/a	n/a	+	+	n/a
Transparency	n/a	n/a	n/a	n/a	n/a	n/a	+	++	n/a

5. AI Assessment

At the time of writing, over 100 different guidelines are released by prominent intergovernmental organisations, professional bodies, national-level committees and other public organisations, non-governmental, and private for-profit companies; a list of which can be found as Appendix B. A qualitative analysis of these guidelines shows a convergence of expert opinion into the key high-level principles of *accountability*, *responsibility*, *transparency*, *robustness*, *data governance and privacy*, and *fairness* [24].

Identifying which principles we want to uphold is only the first step in a lengthy process [25]. There is a need for more robust and concrete requirements and actions to ensure the responsible development, deployment, and usage of the technology [5]. Any commitment or claim of adherence to any value needs to be assessed by monitoring and verifying the claims.

Existing approaches to AI Assessment focus on the use of extensive checklists, e.g. “Assessment List for Trustworthy AI” (ALTAI) [26]. Our activities during—and subsequent reporting of—the “Piloting Evaluation of Assessment List for Trustworthy AI” indicate a general positive reception to this European initiative from AI practitioners [2]. However, it is also clear that there is still work to be done for any AI Assessment to be considered usable by organisations.

First, the current approach of producing high-level ‘system-agnostic’ checklists, where you “answer once and forget,” is not a suitable way of assessing. Lists, such as the ALTAI, are system agnostic, i.e., they are meant to be used by all systems, regardless of their application domain, AI techniques used, and current stage in its development lifecycle. While arguably by being agnostic, they can accommodate current and future applications. They also result in being generic and excessive, often with duplicate questions, making answering them a laborious activity. Moreover, the questions used make heavy references of the actual values they are assessing compliance of. This is counterproductive and increases confusion. That is as each of these values have highly contextualised with interpretations that vary between cultures and other socio-economic backgrounds (e.g. educational background). In our interviews, we often encountered the issue that each interviewee in each session was using their interpretation of an ethical value. Sometimes the interpretations used by a group of interviewees were even contradicting each other, e.g. when *fairness* was defined as equality, i.e., everyone gets the same treatment, and *equity*, i.e., access to equal opportunities through positive discrimination. This was no surprise; there is no such thing as universal morality to enable us to define universally accepted interpretations of ethical and social values [27].

By realising that morality is culture-driven rather than universal, the question of ethics may be seen in a different—and probably humbler—light. Instead of working for the frivolous task of seeking a “one-size-fits-all” solution by finding widely acceptable interpretations for ethical and social values, we can *make interpretations of such ELSEC values explicit and transparent*[5], [28]. This goal serves a dual purpose: to both make the values actionable and allow the subsequent verification and validation of a system’s adherence to values

During the piloting evaluation, we also pinpointed challenges in using the ALTAI beyond the specifics of the format and questions used. In our analysis, discussed in detail in Deliverable 5.6, we identify the lack of awareness and appropriate education to identify and mitigate ethical socio-legal issues by AI developers as ongoing challenges of adopting AI assessment. Further, the clustering of business operations into separate teams, e.g., *legal*, *development*, *etc.*, without the necessary bridges for multi-stakeholder communication and analysis of the system is also a catalyst for organisations not being able to perform adequate AI assessments.

6. Framework for Trustworthy AI Evaluation

The aim of task T5.3 is to survey and develop a usable methodology for developing AI systems that take ELSEC into account. It builds further on the existing initiatives reported in section 5, particularly on the guidelines for Trustworthy AI developed by the European High-Level Expert Group on AI [26]. We develop our methodology using the recommendations, requests, and findings reported in the previous sections. Rather than completely replacing traditional software development methodologies, e.g. scrum, we supplement them through our framework of using *Design for Values* for creating an ethical policy and the related assessment framework presented in this section.

6.1 The use of a Design for Values approach

Design for Values is a methodological design approach that aims at making moral values part of technological design, research, and development [29]. Values are typically high-level abstract concepts that are difficult to incorporate in software design. To design systems that are able to deal with moral values, values need to be interpreted in concrete operational rules. However, given their abstract nature, values can be interpreted in different ways. The Design for Values approach ensures that the link between values and their concrete interpretations in the design and engineering of systems can be traced and evaluated.

Responsible use of AI reduces risks and burdens and ensures that societal and ethical values are central to development. However, it is not always obvious to organisations and developers how best to approach Responsible AI in their development processes. Most software development methodologies follow a development life cycle that includes the steps of analysis, design, implementation, evaluation, and maintenance. However, a responsible approach to AI systems design requires the evaluation process to be continuous during the whole development process and not just a step in the development sequence. Moreover, the dynamic and adaptable nature of AI systems also requires continuous evaluation because the system is continuously evolving.

Therefore, the responsible development life cycle for AI systems must ensure that the whole process is centered on evaluation and justification processes, as depicted in Figure 1 The **Responsible Development Life Cycle for AI systems** During the development of AI systems, taking a Design for Values approach means that the process needs to include activities for (i) the identification of societal values, (ii) deciding on a moral deliberation approach (e.g., through algorithms, user control, or regulation), and (iii) linking values to formal system requirements and concrete functionalities [30].

A Design for Values approach provides guidelines to how AI applications should be designed, managed, and deployed, so that values can be identified and incorporated explicitly into the design and implementation processes. Design for Values methodologies, therefore provide means to support the following processes:

1. Identify the relevant stakeholders;
2. Elicit values and requirements of all stakeholders;
3. Provide means to aggregate the values and value interpretations from all stakeholders;
4. Maintain explicit formal links between values, norms, and system functionalities that enable adaptation of the system to evolving perceptions and justification of implementation decisions in terms of their underlying values;
5. Provide support to choose system components based on their underlying societal and ethical conceptions when these components are built or maintained by different organisations, holding potentially different values.

These issues point to the need for a multi-layered approach to software development where the links to the values are kept explicit. In the following, we present a possible design methodology for Responsible AI, based on the Value Sensitive Software Development (VSSD) framework proposed by [30]. In software design, architectural decisions capture key design issues and the rationale behind chosen solutions. These are conscious and purposeful development decisions concerning the application, which impact (non-functional) characteristics such as software quality attributes. A fundamental result of software engineering methods is ensuring that these architectural decisions are made explicit.

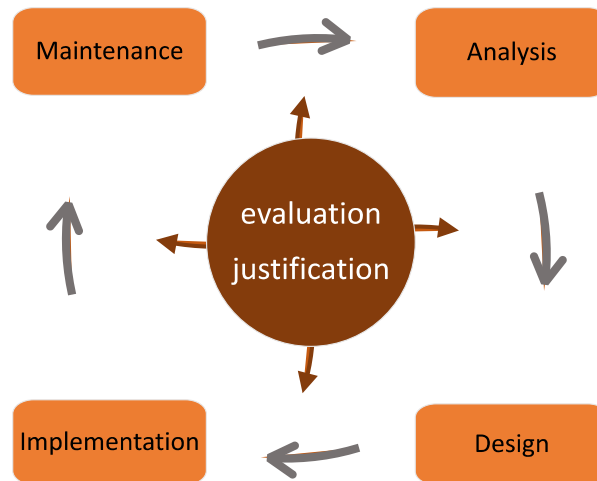


Figure 1 The Responsible Development Life Cycle for AI systems

This framework, depicted in *Figure 2*, connects traditional software engineering concerns with a Design for Values approach to inform AI systems' design. On the one hand, and as described above, Design for Values (left-hand side of the figure) describes the links between values, norms, and system functionalities. On the other hand, domain requirements (right-hand side of the figure) shape software systems' design in terms of the functional, non-functional, and physical/operational demands of the domain. An AI system must obey both orientations, i.e., meet domain demands and at the same time, ensure alignment with social and ethical principles.

By structuring the design of an AI system in terms of high-level motives and roles, specific goals, and concrete plans and actions, it becomes possible to align with both the Design for Values and Software Engineering approaches. As such, at the top level, values and non-functional requirements will inform the specification of the system's motives and roles by clarifying the intention and scope of the system. Norms will provide the (ethical- societal) boundaries for the system's goals, which at the same time need to guarantee that functional requirements are met. Finally, the implementation of plans and actions follows a concrete platform/language instantiation of the functionalities identified by the Design for Values process while ensuring operational and physical domain requirements. These decisions are grounded on both domain characteristics and the values of the designers and other stakeholders involved in the development process. Taking a Design for Values perspective, it becomes possible to make explicit the link to the values behind architectural decisions. In parallel, the system architecture must also reflect the domain requirements, which describe specific contextual concerns. Making these links explicit allows for improvements in the traceability of values throughout the development process, which increases the maintainability of the application. That is if a value is to be interpreted differently, having explicit links between the values and the functionalities that contribute to the realisation of that value makes it much easier to determine which parts of the application should be updated.

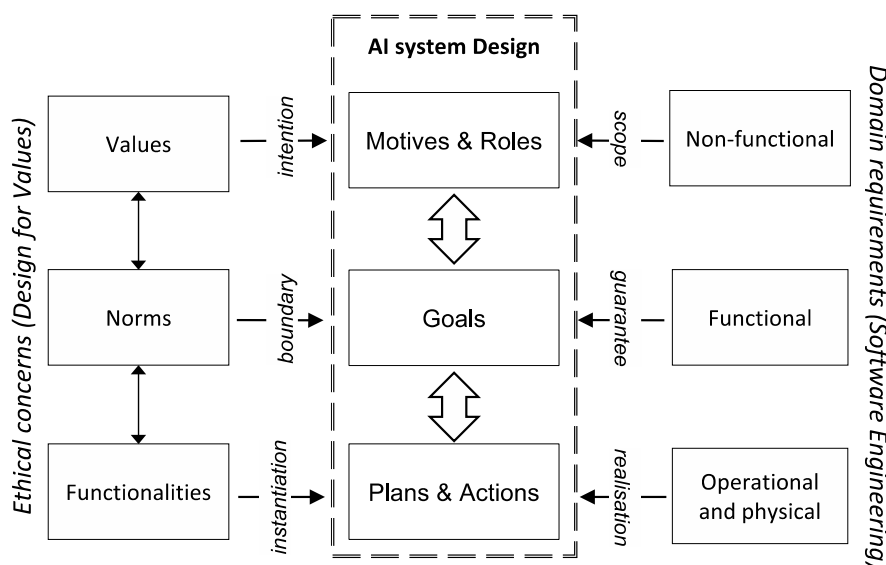


Figure 2 Responsible AI design: integrating ethical concerns and domain considerations in the design of AI applications

6.2 Overall Assessment Framework

To ensure that the HLEG requirements for Trustworthy AI are fully integrated in the development and use of AI systems demands a complex integration of different aspects of design and impact assessment and need to be aligned with the many different types of design processes, AI-based systems and other products, and sectors of application. Several organisations are currently developing approaches to evaluation, risk assessment, and support developers and users assess AI impact. This report proposes an overarching proposal identifying the different aspects to be considered, including stage of development and use, type of AI application, stakeholder, and system properties. This framework is seen in Table 3. Depending on the sector and the overall expected impact of the application, the next stage is to provide concrete measures or actions to operationalise the Trustworthy AI requirements.

During the next stage of T5.3 in months 24-35, and in collaboration with T5.4 (evaluation), we will be completing and testing this approach, depicted in the table below, take as examples concrete applications in collaboration with AI4EU partners. This is also the process that has been advised by the organisations that participated in the deep dive evaluation of the trustworthy AI guidelines, reported in D5.6.

Moreover, a trustworthy AI assessment should go further than a mere checklist and provide means to measure the current maturity of a given application and the means or suggestions for improvement.

As an example of refinement of this approach, depicted in Table 4, we consider a B2C application (e.g. a robot welcoming costumers into a shop) and the requirement of transparency, and considering the needs and expertise of different stakeholders, a possible operationalisation of Table 3 could provide the following support for maturity assessment (where the green boxes indicate the current level of the application being evaluated). Different sectors can demand different types of maturity evidence and/or demand a minimum maturity level.

Table 3 Our assessment framework with the 7 Key Requirements for Trustworthy AI, as identified by the HLEG [26], with the different aspects to be considered, including stage of development and use, type of AI application, stakeholder, and system properties.

	Human agency and oversight	Technical robustness and safety	Privacy and data governance	Transparency	Diversity, non-discrimination and fairness	Environmental and societal well-being	Accountability
stage							
Analysis							
Conceptual design							
Implementation							
Testing							
System							
Deployed product							
Policy/ document							
Concept / research							
stakeholder							
developer							
management							
legal							
End user (expert)							
End user (layperson)							
System properties							
Autonomous							
Adaptable							
Interactive							

Table 4 Example requirements for a B2C application for the value of *Transparency*. Minimum maturity levels in grey.

	Developers	Managers	Legal Experts	End Users	Public
3.	Compliance with standard IEEE P7001	Compliance with standard ISO9001:2015	Specialised Tools	Tools for real-time transparency	N/A
2.	Debugging Tools	Auditing Reports	Access to Auditing Reports	Advance Documentation	Active Indication, e.g. LEDs.
1.	- Version Control - Documentation	Documentation of business cases	Access to logs	- System identifies itself as AI - Basic Documentation	Passive Indication, e.g. stickers.
0.	No Compliance	No Compliance	No Compliance	No Compliance	No Compliance

Finally, for the support of design teams, we are also considering the development of Trustworthy AI design canvas, such as the one depicted in *Appendix C*. This canvas has been used for educational purposes, and has been adapted to the support of the UNICEF guidance for AI and Children⁴ and will be further tested and evaluated in concrete collaborations with AI4EU partners.

⁴ For more information on this project, see <https://www.unicef.org/globalinsight/featured-projects/ai-children>. The support tools, including the design canvas are available at: <https://www.unicef.org/globalinsight/media/1166/file/UNICEF-Global-Insight-tools-to-operationalize-AI-policy-guidance-2020.pdf>

7. ELSEC Evaluation Tools

As part of WP5 Task 5.3, a toolbox of instruments to support the design and evaluation of ethical, legal and societal impact of AI systems was developed. These tools implement the framework described in Section 6 and are aligned with the Trustworthy AI requirements defined by the European High-Level Expert Group on AI, and will be further refined and tested during the remainder of Task 5.3 (months 24-35).

7.1 Tool for Self-Evaluation

We developed a tool to enable organisations to check their compliance with the Trustworthy AI requirements defined by the European High-Level Expert Group on AI [26]. The tool, to be made available in the AI4EU catalogue and on Github, provides self-evaluation support to multiple stakeholders from within or outside the organization with specific competences on each of the requirements and related questions for trustworthy AI. Our expected users include management and development teams.

Following the recommendations made by the interviewees and the key reflections of Deliverable 5.6, we allow users of our tool to customise their experience. Instead of a static experience, i.e. a simple digitise version of the ALTAI, we allow the users to rank requirements and indicate any particularities, e.g., application domain, technique used, of the product they are checking the compliance for. The user can then see questions in order of relevance or even provide search. The answer to each question is clearly indicate to the user and saved for future sessions. Through an intuitive user interface, we empower non-technical managerial personnel to continuously monitor their process. A screen of the tool can be seen in *Figure 3*.

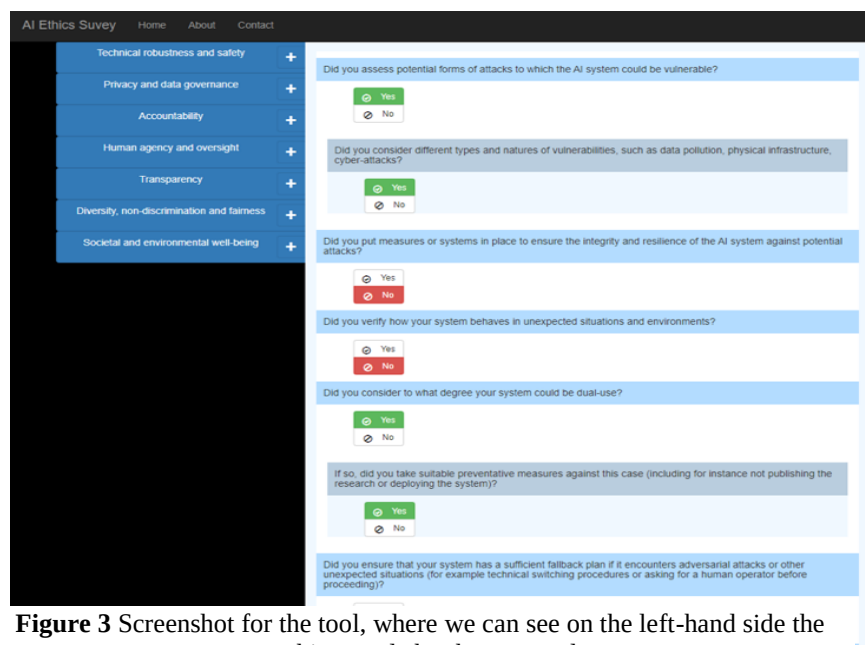


Figure 3 Screenshot for the tool, where we can see on the left-hand side the ranking made by the user and

7.2 Abbreviated Assessment list for AI applications in the AI4EU catalogue

To support the evaluation of applications shared through the AI4EU platform, we propose a framework which is adapted from the ones used in our recent articles on artificial intelligence (AI) and the Sustainable Development Goals (SDGs) [31] and the evaluation of covid19 tracing apps

[32], to evaluate current approaches and concerns related to development, deployment and usage of tracking apps.

The choice for a simplified 3-point scale is meant to facilitate the evaluation. Organisations sharing an AI-based system to the AI4EU platform are asked to indicate whether they consider having sufficiently and appropriately addressed evaluation criteria (2), have given it some thought but possibly not sufficient to guarantee the satisfaction of the criteria (1), or have not considered it or see it as non-applicable to their case (0). This method does not fully measure the 'quality' of the AI-based system but merely reflects the understanding and concern with its trustworthiness by the organisation sharing it.

This abbreviated assessment is not a substitute to the full evaluation of trustworthy AI, as defined by the European High-Level Expert Group on AI in their ALTAI report and supported by the tool presented in Section 7.1. The abbreviated assessment aims to provide a means to quickly determine which applications should be shared in the AI4EU platform. The abbreviated assessment list should replace the current two questions about GDPR compliance and Trustworthy AI. Upon acceptance of an application, it is desirable that such application will realise and publish the full assessment with X months after publication in the AI4EU catalogue.

7.2.1 Abbreviated Assessment List for applications included in AI4EU catalogue

The abbreviated assessment list below is aligned with the 7 requirements for trustworthy AI, as follows:

1. Human Agency and Oversight. (questions 1, 5, 8, 9)
2. Technical Robustness and Safety. (questions 7, 8, 10)
3. Privacy and Data Governance (questions 2, 6, 11)
4. Transparency. (questions 3, 12, 13, 14)
5. Diversity, Non-discrimination, and Fairness. (questions 1, 4)
6. Societal and Environmental Well-being. (questions 5, 9, 14)
7. Accountability. (questions 13, 15)

Any organisation intending to add applications to the AI4EU catalogue should be required to fill in this abbreviated assessment and make the results available openly, alongside the application. AI4EU management can decide on a minimum threshold for inclusion of the candidate application. We suggest that all criteria should be graded 1 or higher for all included applications.

Impact on users

1. Inform on how it is respecting individuals' fundamental rights: This includes considering the effect of the application on the rights to safety, health, non-discrimination, and freedom of association. A clear analysis and response to these principles is desired (2). Unclear information/only partially respecting these rights (1), or not respecting them (0) is not adequate.
2. Privacy and data protection: Data collection should be compliant with the General Data Protection Regulation (GDPR) [33] and respect the individual's privacy. A Data Protection Impact Assessment (DPIA) must be carried out before the deployment of any system. The

purpose of the AI system and the mechanisms to assess its usage need to be clearly defined. All these requirements should be fulfilled (2), whereas fulfilling them only partially (1) or not at all (0) is not adequate.

3. Transparency rights: They include the right of users to be notified, to control their own data, transparency regarding which personal data are collected, and explanation of results produced by the system. It should be clear who, when, and how the system can be audited. Fulfilment of all requirements (2) is suitable, whereas fulfilling them only partially (1) or not at all (0) is not adequate.
4. Accessibility: Possibility to be used by all regardless of demographics, language, disability, digital literacy, and financial accessibility. All these requirements should be fulfilled (2), whereas addressing them only partially (1) or not at all (0) is not adequate.
5. Education and tutorials: Ensure that users are informed and capable of using the system correctly, including e.g., in-system help (2), or external materials, e.g., website (1). Absence (0) is not adequate.

Technology

6. Data management: Ensure data-minimization principle, i.e., usage of local and temporary storage and encryption, based on principles of data protection by design. Ensure that only data strictly necessary are captured and processed. All these requirements are needed (2), whereas unclear documentation (1) or lack of compliance with all of them are not adequate (0).
7. Security: User authentication to prevent risks such as access, modification, or disclosure of the data. Use unique and pseudo-random identifiers, renewed regularly and cryptographically strong. Compliance with these requirements is needed (2), whereas unclear (1) or lack of compliance are not satisfactory (0).
8. Ease to deactivate/remove: Either through clear instructions or automatically by sunset clause (2). Unclear (1) or difficulties for removing the system and the data are not sufficient (0).
9. Ease to access services without using the AI system: In case of AI-systems aimed to replace or complement public services, full non-system alternatives (2), partial (1), none (0)
10. Open-source code: Participatory and multidisciplinary development, access to the code and methods used for adaptation to new knowledge on the virus (2). Open-source code without the possibility of contributing (1) is not recommended, and non-open-source code is undesirable (0).

Governance

11. Ownership: Openness over app (including code, data, use) ownership is preferable (2), unclear (1) or unknown (0) are less adequate.
12. Openness over Data governance: Open data governance is preferable (2), while intermediate (1) or private/opaque settings (0) are not suitable.
13. Legislation and Policy: Clear, broader legal framework (2), partial governmental policy (1) whereas no policy or unknown is not desirable (0).
14. Design Impact Assessment and Open Development Process: Explicit design process, including a clear description about aims and motivation, stakeholders, public consultation

process and impact assessment (2). Unclear information (1) or the lack of such an assessment (0) is not adequate.

15. Right to contest/liability. Users need to be able to contest decisions or demand human intervention (2). Partial/unclear compliance (1) or the lack of this feature (0) are not adequate.

7.2.1 Example use of the abbreviated checklist

To better illustrate the use of the abbreviated assessment list, we considered two resources of the AI4EU Catalogue. These two resources are classified as “AI in Health” in the AI Catalogue (the list of “AI in Health” resources at the time of analysis is given in Appendix 4).

The first resource that we consider is *r.AID.ologist*.⁵ *r.AID.ologist* is described as a solution to provide writing assistance on reports to radiologists. By considering the information in the Catalogue, we assess *r.AID.ologist* against the criteria of the abbreviated assessment list. The result of the assessment can be seen in Table 5.

Table 5 Grades of the *r.AID.ologist* resource

AI in health resource	Criteria of the abbreviated assessment															
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
<i>r.AID.ologist</i>	1	1	1	1	2	1	1	0	1	1	2	2	1	0	1	1

To provide a visual representation of these grades of the abbreviated assessment list, we provide a Radar chart of the grades of the *r.AID.ologist* resource in Figure 4.

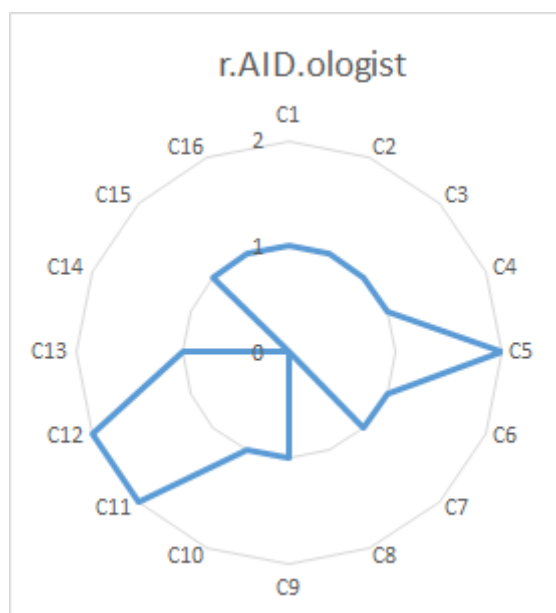


Figure 4 Radar chart of the grades of the *r.AID.ologist* resource

⁵ Found on the AI4EU website: <https://www.ai4eu.eu/resource/raidologist>

The second resource that we consider is to illustrate the abbreviated assessment list is: *ISIC CNN*.⁶ ISIC CNN is presented as an AI resource to classify of dermoscopic images. By considering the documentation of this resource in the AI Catalogue, we allocate the grades seen in *Table 6* to each of the abbreviated assessment list criteria. A Radar chart of these grades is presented in *Figure 5*.

Table 6 Grades of the ISIC CNN resource after applying our abbreviated assessment list.

AI in health resource	Criteria of the abbreviated assessment															
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
ISIC CNN	0	1	0	1	1	0	0	0	0	1	2	2	0	0	0	0

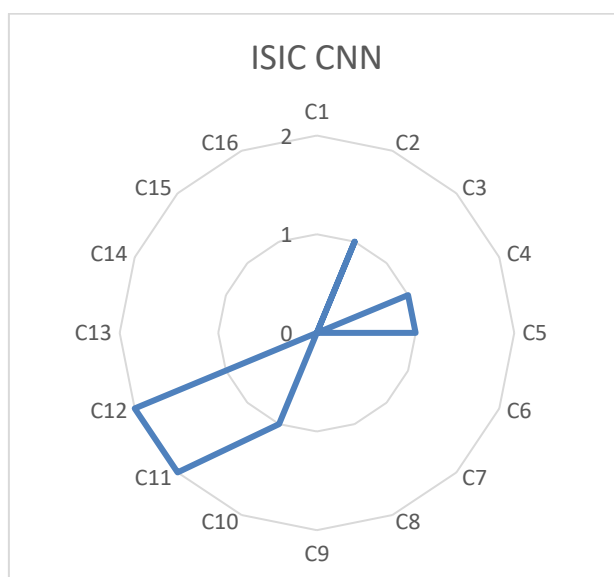


Figure 5 Radar chart of the grades of the ISIC CNN resource.

8. Conclusion

In this document, we present the results of a survey we conducted on the existing and required practices of the AI4EU consortium. Further, we provided summaries of a literature survey and a discussion of the current state of AI Assessment.

Considering the results presented here and in Deliverable 5.6, we developed a framework to help organisations integrate ELSECLEC values into their systems' development. Moreover, to empower organisations perform the necessary validation and verification, we presented practical tools. One of those tools is an abbreviated assessment list, which we plan to incorporate into the platform. The inclusion and use of the abbreviated assessment list should introduce a minimum level of compliance to ELSEC values – or serve as evidence of the lack of– for any AI products and services submitted for distribution in the AI4EU platform.

⁶ <https://www.ai4eu.eu/resource/isic-cnn>

9. References

- [1] J. J. Bryson and A. Winfield, "Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems," *Computer (Long. Beach. Calif.)*, vol. 50, no. 5, pp. 116–119, May 2017.
- [2] A. Cortés *et al.*, "AI4EU Deliverable D5.6 Report on the Deep Dive Piloting the Trustworthy AI Assessment List," 2019.
- [3] V. Dignum, "Ethics in artificial intelligence: introduction to the special issue," *Ethics Inf. Technol.*, vol. 20, no. 1, pp. 1–3, Mar. 2018.
- [4] V. Dignum, *Responsible Artificial Intelligence*. Cham: Springer International Publishing, 2019.
- [5] A. Theodorou and V. Dignum, "Towards ethical and socio-legal governance in AI," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 10–12, Jan. 2020.
- [6] A. F. T. Winfield and M. Jirotko, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 376, no. 2133, p. 20180085, Nov. 2018.
- [7] M. Hildebrandt, "11. Closure: on ethics, code and law," in *Law for Computer Scientists*, 2019.
- [8] I. Sommerville, *Software Engineering*, 9th ed. Pearson, 2011.
- [9] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1995.
- [10] M. Fowler, D. Rice, M. Foemmel, E. Hieeatt, R. Mee, and R. Stafford, *Patterns of Enterprise Application Architecture*. Addison-Wesley Professional, 2002.
- [11] R. C. Martin, *The Clean Coder: A Code of Conduct for Professional Programmers*. Prentice Hall, 2011.
- [12] D. Turk, R. France, and B. Rumpe, "Limitations of Agile Software Processes," *Third Int. Conf. Extrem. Program. Flex. Process. Softw. Eng.*, Sep. 2002.
- [13] M. W. Mwadulo, "Suitability of Agile Methods for Safety-Critical Systems Development: A Survey of Literature," 2016.
- [14] Z. Ul Islam, A. Theodorou, and V. Dignum, "A Survey of Software Engineering Methods for Artificial Intelligence."
- [15] V. Dignum, "A model for organizational interaction: based on agents, founded in logic," SIKS, 2004.
- [16] P. Noriega and D. de Jonge, "Electronic institutions: the EI/EIDE framework," in *Social coordination frameworks for social technical systems*, Springer, 2016, pp. 47–76.
- [17] O. Boissier, R. H. Bordini, J. F. Hübner, A. Ricci, and A. Santi, "Multi-agent oriented programming with JaCaMo," *Sci. Comput. Program.*, vol. 78, no. 6, pp. 747–761, 2013.
- [18] N. Gilbert, "Agent-based social simulation: dealing with complexity," *Complex Syst. Netw. Excell.*, vol. 9, no. 25, pp. 1–14, 2004.
- [19] O. Biggar, M. Zamani, and I. Shames, "A principled analysis of Behavior Trees and their generalisations." 2020.
- [20] J. Orkin, "Three states and a plan: the AI of FEAR," *Game Dev. Conf.*, vol. 2006, no. 1, pp.

1–18, 2006.

- [21] A. I. R. L. Azevedo and M. F. Santos, “KDD, SEMMA and CRISP-DM: a parallel overview,” *IADS-DM*, 2008.
- [22] J. J. Bryson, “The behavior-oriented design of modular agent intelligence,” in *System*, vol. 2592, 2002, pp. 61–76.
- [23] J. Dias, S. Mascarenhas, and A. Paiva, “Fatima modular: Towards an agent architecture with a generic appraisal framework,” in *Emotion modeling*, Springer, 2014, pp. 44–56.
- [24] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI,” 2020.
- [25] J. Whittlestone, A. Alexandrova, R. Nyrop, and S. Cave, “The role and limits of principles in AI ethics: Towards a focus on tensions,” in *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 195–200.
- [26] “Ethics guidelines for trustworthy AI,” European Commission, Brussels, Apr. 2019.
- [27] E. Turiel, *The Culture of Morality*. 2001.
- [28] A. Aler Tubella, A. Theodorou, F. Dignum, and V. Dignum, “Governance by Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5787–5793.
- [29] M. J. Van Den Hoven, “Design For Values And Values For Design,” *Inf. Age*, vol. August/Sep, no. 2, pp. 4–7, 2005.
- [30] H. Aldewereld, V. Dignum, and Y. H. Tan, “Design for values in software development,” in *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*, M. J. van Den Hoven, P. E. Vermaas, and I. van De Poel, Eds. Springer, 2014, pp. 831–845.
- [31] R. Vinuesa *et al.*, “The role of artificial intelligence in achieving the Sustainable Development Goals,” *Nat. Commun.*, vol. 11, no. 1, p. 233, Dec. 2020.
- [32] R. Vinuesa, A. Theodorou, M. Battaglini, and V. Dignum, “A socio-technical framework for digital contact tracing,” *Results Eng.*, vol. 8, p. 100163, Dec. 2020.
- [33] European Parliament and Council of European Union, *General Data Protection Regulations (GDPR)*. EU, 2016.

10. Appendix A: Survey of AI4EU

Table 7 Demographic information for Affiliation group

Affiliations	n	%
University	5	28%
Large-Sized Company	5	28%
Academic Research institution	4	22%
Non-Academic Government Organization	2	11%
Small-Sized Company	1	5.55%
NGO	1	5.55%

Affiliations	Roles	Total
	Academic Researcher	11
	Industry based researcher	5
	AI Developer/ Engineer/ Architect	7
	Product/Project Manager on an AI-related product/service	7
	End user (commercial)	2
	End user (personal)	1
	Consultant	3
	n (Total no of responses)	36
University	Academic Researcher	100%
Academic Research Institute	Academic Researcher	100%
Large-Sized Company	Industry based researcher	80%
	AI Developer/ Engineer/ Architect	80%
Non-Academic Government Organization	AI Developer/ Engineer/ Architect	50%
Small Sized Company	Product/Project Manager on AI related product and services	100%
NGO		

Table 8 Participants Roles based on affiliations and assembled in Categories (academia/industry):

Categories	Affiliations	Academic Researchers	Industry based researchers	AI developers Engineers Architecture	Product Project Manager on AI
Academia	University Academic Research institute	100%	0%	11%	22%
Industry	Large sized company Non-Academic Govt. Organization Small sized company NGO	22%	55%	67%	67%

Table 9 Development Experience:

	Traditional System Development (non-autonomous) systems		AI System Development	
	Past	Actively working	Past	Actively working
Academia	100%	78%	100%	100%
Industry	55%	33%	88%	67%

Table 10 Areas of AI that would benefit from standard definitions and taxonomies:

Areas of AI	Categories	Yes	No	Maybe
Fundamental technical terms (machine learning, agent, cognitive architecture etc.)	Academia	100%	0%	0%
	Industry	78%	11%	11%
Low Level Technical terms (reasoning, learning, functions, goals etc.)	Academia	78%	11%	11%
	Industry	22%	11%	67%
Guideline Terms (transparency, explainability, accountability etc.)	Academia	78%	0%	22%
	Industry	33%	0%	67%
Job Roles and Expertise (AI ethics, AI auditor, AI architecture etc.)	Academia	67%	0%	33%
	Industry	56%	0%	44%

Table 11 Standards and certifications followed during development:

	Yes	No	NA
<i>Academia</i>	0%	33%	67%
<i>Industry</i>	33%	56%	11%

Table 12 Development Frameworks/Paradigms currently in practice:

	<i>Scrum</i>	<i>Waterfall</i>	<i>Extreme Programming</i>	<i>Test-driven development</i>	<i>DevOps</i>	<i>Rapid Application development</i>	<i>None</i>
<i>Academia</i>	44%	22%	33%	11%	11%	22%	44%
<i>Industry</i>	56%	22%	0%	33%	33%	11%	22%

Table 13 Frameworks and methods Preferred:

<i>Frameworks/Paradigms</i>		<i>Ratings 3-4</i>	<i>Ratings 0-1-2</i>
<i>Scrum</i>	<i>Academia</i>	44%	56%
	<i>Industry</i>	67%	33%
<i>Waterfall</i>	<i>Academia</i>	22%	78%
	<i>Industry</i>	22%	78%
<i>Extreme Programming</i>	<i>Academia</i>	11%	89%
	<i>Industry</i>	0%	100%
<i>Test-Driven Development</i>	<i>Academia</i>	22%	78%
	<i>Industry</i>	44%	56%
<i>DevOps</i>	<i>Academia</i>	0%	100%
	<i>Industry</i>	33%	67%
<i>Rapid Application Development</i>	<i>Academia</i>	33%	67%
	<i>Industry</i>	44%	56%
<i>None</i>	<i>Academia</i>	44%	56%
	<i>Industry</i>	11%	89%

Table 14 Which of the following to be included in ethical guidelines on scale of low medium high must include not important?

		<i>Must include High</i>	<i>Medium low</i>	<i>Not Important</i>
<i>System failure early warning systems (e.g. Threshold)</i>	<i>Academia</i>	78%	22%	0%
	<i>Industry</i>	67%	33%	0%
<i>Business as usual System maintenance (E.g. understanding model drift, retraining frequency, training batch lookback, online / streaming training efficacy)</i>	<i>Academia</i>	56%	33%	11%
	<i>Industry</i>	67%	33%	0%
<i>Defining a natural variance of model accuracy (In case of ML systems, if different training set -but same features, what is the variance in model performance due to variation in data - gives a better estimate of future performance)</i>	<i>Academia</i>	67%	33%	0%
	<i>Industry</i>	56%	44%	0%
<i>Challenger benchmarks (Production model benchmarking against other similar systems or purpose-made challenges, e.g. ImageNet)</i>	<i>Academia</i>	89%	11%	0%
	<i>Industry</i>	56%	33%	11%
<i>Standardised Reporting (Standard sets of statistics to be reported for the system's performance)</i>	<i>Academia</i>	33%	56%	11%
	<i>Industry</i>	78%	22%	0%
<i>Documentation</i>	<i>Academia</i>	56%	44%	0%
	<i>Industry</i>	89%	11%	0%
<i>Robustness to unintended feedback loops through specific design considerations and diagnostic monitoring</i>	<i>Academia</i>	78%	22%	0%
	<i>Industry</i>	56%	44%	0%
<i>Adversarial example testing (Systematically testing the model's behaviour on developer generated potential extreme / rare instances)</i>	<i>Academia</i>	89%	11%	0%
	<i>Industry</i>	33%	67%	0%
<i>Adversarial Attack Hardening (Reasonable safeguards against adversarial attacks)</i>	<i>Academia</i>	67%	33%	0%
	<i>Industry</i>	56%	44%	0%

Table 15 Contents of the documentation:

		<i>Average</i>
<i>Requirements</i>	<i>Academia</i>	89%
	<i>Industry</i>	89%

<i>High-level description of the decision-making system used, e.g. method</i>	<i>Academia</i>	100%
	<i>Industry</i>	89%
<i>Low-level details of the decision-making system, e.g. rules in classical AI systems / Plans in Behaviour-Based AI / Training data used for ML</i>	<i>Academia</i>	44%
	<i>Industry</i>	67%
<i>Identified risks</i>	<i>Academia</i>	100%
	<i>Industry</i>	78%
<i>System boundaries</i>	<i>Academia</i>	89%
	<i>Industry</i>	78%
<i>Testing and stress-testing results, e.g. accuracy, tested uptime, etc.</i>	<i>Academia</i>	56%
	<i>Industry</i>	67%

Table 16 Do you use any AI-specific Architecture for AI systems?

	Yes	No
<i>Academia</i>	44%	56%
<i>Industry</i>	0%	100%

Table 17 Should the methodology make explicit recommendations on what architecture system should have?

	Yes	No
<i>Academia</i>	44%	56%
<i>Industry</i>	44%	56%

Table 18 Should the methodology integrate a specific architecture into our methodology?

	Yes	No
<i>Academia</i>	22%	78%
<i>Industry</i>	33%	67%

Table 19 Should the methodology include any recommendations/ instructions on why, how, and when to use version control?

	Yes	No
<i>Academia</i>	66%	33%
<i>Industry</i>	78%	22%

Table 20 Would you limit performance of a system to ensure legal compliance and Ethical compliance?

	<i>Legal Compliance</i>		<i>Ethical Compliance</i>	
	Yes	No	Yes	No
<i>Academia</i>	100%	0%	89%	11%
<i>Industry</i>	100%	0%	100%	0%

Table 21 If you are performing any auditing what are you auditing for?

	<i>Ethics policy compliance</i>	<i>Legal compliance</i>	<i>Finance</i>	<i>Corruption</i>	<i>Disuse of resources</i>	<i>Implementation/testing issues</i>	<i>Others</i>
<i>Academia</i>	55%	67%	22%	11%	22%	55%	44%
<i>Industry</i>	22%	66%	11%	22%	11%	44%	0%

Table 22 How often do you audit a project's compliance with your organisation's policy?

	<i>At requirements gathering phase</i>	<i>Design phase</i>	<i>During Development</i>	<i>During Testing</i>	<i>Prior to First Deployment</i>	<i>Prior to every Deployment</i>	<i>Annual Review</i>
<i>Academia</i>	56%	44%	33%	22%	22%	0%	22%
<i>Industry</i>	33%	44%	56%	11%	22%	22%	22%

Table 23 Do you use any tools for verification and validation?

	Yes	No
<i>Academia</i>	22%	78%
<i>Industry</i>	11%	89%

Table 24 Do you keep internal documentation on your data governance practices (if any)?

	Yes	No
<i>Academia</i>	56%	44%
<i>Industry</i>	44%	56%

Table 25 Methodology Requirements:

	Require specific standards to follow		Include procedures for software updates		Provide specification for testing (minimum required tests)	
	Yes	No	Yes	No	Yes	No
<i>Academia</i>	22%	78%	89%	11%	89%	11%
<i>Industry</i>	56%	44%	100%	0%	100%	0%

Table 26 Reporting: Do you produce reports of the effects of your products/services/research output?

		Always	No	Only when relevant
<i>Social Implication</i>	<i>Academia</i>	0%	33%	67%
	<i>Industry</i>	0%	44%	56%
<i>Jobs Market</i>	<i>Academia</i>	0%	44%	56%
	<i>Industry</i>	0%	56%	44%
<i>Legal Implications (e.g. it is in a grey area of legislation.)</i>	<i>Academia</i>	11%	33%	56%
	<i>Industry</i>	11%	33%	56%
<i>Environmental</i>	<i>Academia</i>	0%	22%	78%
	<i>Industry</i>	0%	33%	67%
<i>Sustainability</i>	<i>Academia</i>	0%	22%	78%

	<i>Industry</i>	<i>0%</i>	<i>33%</i>	<i>67%</i>
<i>Risk of Misuse</i>	<i>Academia</i>	<i>0%</i>	<i>33%</i>	<i>67%</i>
	<i>Industry</i>	<i>11%</i>	<i>11%</i>	<i>78%</i>

Table 27 Disclosure of the system architecture Technical Details:

	<i>List of agent pre-defined goals</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	<i>44%</i>	<i>44%</i>	<i>11%</i>
<i>Industry</i>	<i>33%</i>	<i>56%</i>	<i>11%</i>

	<i>Indication of sensors location & capabilities</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	<i>33%</i>	<i>56%</i>	<i>11%</i>
<i>Industry</i>	<i>22%</i>	<i>56%</i>	<i>22%</i>

Table 28 Disclosure of the machine learning based systems:

	<i>Disclosure of agent using machine learning (including prior training)</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	<i>33%</i>	<i>56%</i>	<i>11%</i>
<i>Industry</i>	<i>11%</i>	<i>56%</i>	<i>33%</i>

	<i>Disclosure if the agent has online learning</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	<i>44%</i>	<i>44%</i>	<i>11%</i>
<i>Industry</i>	<i>0%</i>	<i>67%</i>	<i>33%</i>

	<i>Disclosure of any data filtering and manipulation used</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	<i>22%</i>	<i>67%</i>	<i>11%</i>
<i>Industry</i>	<i>0%</i>	<i>67%</i>	<i>33%</i>

	<i>Disclosure of training Data Used</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	11%	56%	33%
<i>Industry</i>	0%	56%	44%

	<i>Disclosure trained/production model's characteristics, including accuracy and tuning.</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	22%	56%	22%
<i>Industry</i>	0%	56%	44%

Table 29 Disclosure Related to Development and testing of the system:

	<i>Standards followed during development, e.g. ISO9001.</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	44%	44%	11%
<i>Industry</i>	22%	44%	33%

	<i>Disclosure of security testing used</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	22%	78%	0%
<i>Industry</i>	22%	44%	33%

	<i>Disclosure of high-level description of decision-making systems used, e.g. method.</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	44%	44%	11%
<i>Industry</i>	44%	33%	22%

	<i>Disclosure of Low-level details of the decision-making system, e.g. rules in classical AI systems / Plans in Behaviour-Based AI / Training data used for ML</i>		
--	--	--	--

	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	11%	56%	33%
<i>Industry</i>	22%	44%	33%

	<i>Disclosure of identify risks of using or deployment</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	67%	22%	11%
<i>Industry</i>	0%	56%	44%

	<i>Disclosure of system boundaries/limitations found.</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	44%	44%	11%
<i>Industry</i>	22%	56%	22%

	<i>Disclosure of testing conducted at high-level description</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	44%	44%	11%
<i>Industry</i>	0%	67%	33%

	<i>Disclosure of testing and stress testing results e.g. accuracy, tested uptime, etc.</i>		
	<i>For all Users</i>	<i>For expert Users</i>	<i>For none</i>
<i>Academia</i>	22%	67%	11%
<i>Industry</i>	11%	44%	44%

Table 30 Methodology

	<i>Development methodology to contain use cases</i>	<i>Would you Provide use cases on personal experience?</i>	<i>Methodology to contain sample tools/libraries</i>

	Yes	No	Yes	No	Yes	No
<i>Academia</i>	100%	0%	44%	56%	89%	11%
<i>Industry</i>	78%	22%	67%	33%	89%	11%

Q1 - Affiliation

Q2 - Organization type

Q3 - Are you actively working in the development of AI systems?

Q4- Are you actively working in the development of traditional (non-autonomous) systems?

Q5 - Have you in the past worked in the development of an AI system?

Q6 - Have you in the past worked in the development of traditional (non-autonomous) systems?

Q7 - What is / are your role(s) in the AI community?

Q8 - Which (if any) of the following areas of AI, would benefit from a standardised taxonomy and definitions? Fundamental technical terms; e.g. artificial intelligence, machine learning, agent, cognitive architecture, etc.?

Q8 - Which (if any) of the following areas of AI, would benefit from a standardised taxonomy and definitions? Low-level technical terms; e.g. reasoning, learning function, goals, etc.?

Q8 - Which (if any) of the following areas of AI, would benefit from a standardised taxonomy and definitions? Guideline terms; e.g. transparency, explainability, accountability, etc.?

Q8 - Which (if any) of the following areas of AI, would benefit from a standardised taxonomy and definitions? Job roles and expertise; e.g. AI ethicist, AI auditor, AI architect, etc.?

Q9 - Do you follow any certification/standard during development?

Q10 - If so, which and why?

Q11 - If not, why?

Q12 - Which of the following development paradigms/frameworks do you currently use?

Q13 - If none, why?

Q14 - Which of the following development paradigms/frameworks do you prefer? Waterfall

Q14 - Which of the following development paradigms/frameworks do you prefer? Extreme Programming

Q14 - Which of the following development paradigms/frameworks do you prefer? Scrum

Q14 - Which of the following development paradigms/frameworks do you prefer? Rapid Application Development

Q14 - Which of the following development paradigms/frameworks do you prefer? Test-driven Development

Q14 - Which of the following development paradigms/frameworks do you prefer? DevOps

Q14 - Which of the following development paradigms/frameworks do you prefer? None

Q15 - If none, why?

Q16 - How important are the following to be included in the guidelines on a scale of Low / Medium / High? System's 'failure' early warning systems (E.g. threshold)

Q16 - How important are the following to be included in the guidelines on a scale of Low / Medium / High? Business as Usual (BAU) system's maintenance (E.g. understanding model drift, retraining frequency, training batch lookback, online / streaming training efficacy)

Q16 - How important are the following to be included in the guidelines on a scale of Low / Medium / High?: Defining a natural variance of model accuracy (In case of ML systems, if different training set -but same features, what is the variance in model performance due to variation in data - gives a better estimate of future performance)

Q16 - How important are the following to be included in the guidelines on a scale of Low / Medium / High? Challenger benchmarks (Production model benchmarking against other similar systems or purpose-made challenges, e.g. ImageNet)

Q16 - How important are the following to be included in the guidelines on a scale of Low / Medium / High? Standardised reporting (Standard sets of statistics to be reported for the system's performance)

Q16 - How important are the following to be included in the guidelines on a scale of Low / Medium / High? Documentation

Q16 - How important are the following to be included in the guidelines on a scale of Low / Medium / High? Robustness to unintended feedback loops through specific design considerations and diagnostic monitoring

Q16 - How important are the following to be included in the guidelines on a scale of Low / Medium / High? Adversarial Example testing (Systematically testing the model's behaviour on developer generated potential extreme / rare instances)

Q16 - How important are the following to be included in the guidelines on a scale of Low / Medium / High? Adversarial Attack Hardening (Reasonable safeguards against adversarial attacks)

Q17 - If you think documentation is important, do you think that we should provide specifications for the formatting and minimum context?

Q18 - If you think that documentation is important, then what do you think the documentation should contain?

Q19 - Do you use any AI-specific architecture for your intelligent systems?

Q20 - If so, which and why?

Q21 - If not, why?

Q22 - Should the methodology make explicit recommendations on what architecture the system should have?

Q23 - Should the methodology integrate a specific architecture into our methodology?

Q24 - If none, why?

Q25 - Should the methodology include any recommendations/ instructions on why, how, and when to use version control?

Q26 - If none, why?

- Q27 - Would you limit performance of a system to ensure legal compliance?
- Q28 - Would you limit performance of a system to ensure ethical compliance?
- Q29 - If you are performing any auditing, what are you auditing for?
- Q30 - If other, please elaborate.
- Q31 - How often do you audit a project's compliance with your organisation's policy?
- Q32 - Do you use any tools for verification & validation?
- Q33 - If so, can you name the tools used?
- Q34 - Do you keep internal documentation on your data governance practices (if any)?
- Q35 - If you have answered 'yes' to the previous question, please elaborate.
- Q36 - Other than recommending standards, do you think that the methodology -and the AI4EU project- should require specific standards to be followed?
- Q37 - Do you think the methodology should include procedures for software updates?
- Q38 - Do you think that the methodology should provide specifications for testing, i.e. the minimum required tests?
- Q39 - Do you produce reports of the effects of your product/service/research output in relation to...: societal implications?
- Q39 - Do you produce reports of the effects of your product/service/research output in relation to...: jobs market?
- Q39 - Do you produce reports of the effects of your product/service/research output in relation to...: legal implications, e.g. it is in a grey area of legislation.
- Q39 - Do you produce reports of the effects of your product/service/research output in relation to...: environmental.
- Q39 - Do you produce reports of the effects of your product/service/research output in relation to...: sustainability?
- Q39 - Do you produce reports of the effects of your product/service/research output in relation to...: risks of misuse?
- Q40 - Disclosures for the system's architecture/technical details: List of agent's predefined goals.
- Q40 - Disclosures for the system's architecture/technical details: Clear indication of sensors' locations and capabilities.
- Q41 - Disclosures for machine-learning based systems: Disclosure if the agent uses any machine learning (including prior training).
- Q41 - Disclosures for machine-learning based systems: Disclosure if the agent has online learning.
- Q41 - Disclosures for machine-learning based systems: Disclosure of any data filtering and manipulation used.
- Q41 - Disclosures for machine-learning based systems: Disclosure of training data used.
- Q41 - Disclosures for machine-learning based systems: Disclosure trained/production model's characteristics, including accuracy and tuning.

Q42 - Disclosures related to development and testing of the system: Standards followed during development, e.g. ISO9001.

Q42 - Disclosures related to development and testing of the system: Security testing used.

Q42 - Disclosures related to development and testing of the system: High-level description of the decision-making system used, e.g. method.

Q42 - Disclosures related to development and testing of the system: Low-level details of the decision-making system, e.g. rules in classical AI systems / Plans in Behaviour-Based AI / Training data used for ML

Q42 - Disclosures related to development and testing of the system: Identified risks of using or of deployment.

Q42 - Disclosures related to development and testing of the system: System boundaries/limitations found.

Q42 - Disclosures related to development and testing of the system: Testing conducted (high-level description).

Q42 - Disclosures related to development and testing of the system: Testing and stress-testing results, e.g. accuracy, tested uptime, etc.

Q43 - Do you want the development methodology to contain use cases?

Q44 - Would you be interested in providing use cases based on your own experience?

Q45 - Do you want the development methodology to contain sample tools/libraries?

Q46 - Can we contact you at the future for further information?

11. Appendix B

List of guidelines released by prominent intergovernmental organisations, professional bodies, national-level committees and other public organisations, non-governmental, and private for-profit companies.

Name of Document	Author	Country/IGO
Artificial Intelligence. Australia's Ethics Framework. A discussion Paper	Department of Industry Innovation and Science	Australia
AI - Our approach	Microsoft	USA
AI Guidelines	Deutsche Telekom	Germany
AI in the UK: ready, willing and able?	UK House of Lords, Select Committee on Artificial Intelligence	UK
AI Now 2018 Report	AI Now Institute	USA
AI Principles	Future of Life Institute	USA
AI Principles & Ethics	Smart Dubai	UAE
AI Principles of Telefónica	Telefonica	Spain
AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	AI4People	EU
Artificial Intelligence (AI) in Health	Royal College of Physicians	UK
Artificial Intelligence and Machine Learning: Policy Paper	Internet Society	international
Artificial intelligence and privacy	The Norwegian Data Protection Authority	Norway
Artificial Intelligence. The Public Policy Opportunity	Intel Corporation	USA
Artificial Intelligence: open questions about gender inclusion	W20	international
Automated and Connected Driving: Report	Federal Ministry of Transport and Digital Infrastructure, Ethics Commission	Germany
Big data, artificial intelligence, machine learning and data protection	Information Commissioner's Office	UK
Business Ethics and Artificial Intelligence	Institute of Business Ethics	UK
Charlevoix Common Vision for the Future of Artificial Intelligence	Leaders of the G7	international
Commitments and principles	OP Group	Finland
Declaration on ethics and data protection in Artificial Intelligence	ICDPPC	international
DeepMind Ethics & Society Principles	DeepMind Ethics & Society	UK
Digital Decisions	Center for Democracy & Technology	USA
Discussion Paper on Artificial Intelligence (AI) and Personal Data - Fostering Responsible Development and Adoption of AI	Personal Data Protection Commission Singapore	Singapore
Discussion Paper: National Strategy for Artificial Intelligence	National Institution for Transforming India (Niti Aayog)	India
Draft AI R&D Guidelines for International Discussions	Institute for Information and Communications Policy (IICP), The Conference toward AI Network Society	Japan
Dutch Artificial Intelligence Manifesto	Special Interest Group on Artificial Intelligence (SIGAI), ICT Platform Netherlands (IPN)	Netherlands
Ethical Principles for Artificial Intelligence and Data Analytics	Software & Information Industry Association (SIIA), Public Policy Division	international
Ethical, social, and political challenges of Artificial Intelligence in Health	Future Advocacy	UK
Ethically Aligned Design. A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, version 2	Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	international
Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Autonomous and Intelligent Systems, First Edition (EAD1e)	Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	international
Ethics Framework - Responsible AI	Machine Intelligence Garage Ethics Committee	UK
Ethics Guidelines for Trustworthy AI	High-Level Expert Group on Artificial Intelligence	EU
Ethics of AI in Radiology: European and North American Multisociety Statement	American College of Radiology; European Society of Radiology; Radiology Society of North America; Society for Imaging; Informatics in Medicine; European Society of Medical Imaging; Informatics; Canadian Association of Radiologists; American Association of Physicists in Medicine	international
Ethics Policy	Icelandic Institute for Intelligent Machines (IIIM)	Iceland
Ethique de la recherche en robotique	CERNA (Allistene)	France
European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment	Council of Europe: European Commission for the efficiency of Justice (CEPEJ)	EU
Everyday Ethics for Artificial Intelligence. A practical guide for designers & developers	IBM	USA
For a meaningful Artificial Intelligence. Towards a French and European strategy	Mission Villani	France

Governing Artificial Intelligence. Upholding Human Rights & Dignity	Data & Society	USA
How can humans keep the upper hand? Report on the ethical matters raised by AI algorithms	French Data Protection Authority (CNIL)	France
Human Rights in the Robot Age Report	The Rathenau Institute	Netherlands
IBM's Principles for Trust and Transparency	IBM	USA
Initial code of conduct for data-driven health and care technology	UK Department of Health & Social Care	UK
Intel's AI Privacy Policy White Paper. Protecting individuals' privacy and data in the artificial intelligence world	Intel Corporation	USA
Introducing Unity's Guiding Principles for Ethical AI – Unity Blog	Unity Technologies	USA
ITI AI Policy Principles	Information Technology Industry Council (ITI)	international
L'intelligenza artificiale al servizio del cittadino	Agenzia per l'Italia Digitale (AGID)	Italy
Machine learning: the power and promise of computers that learn by example	The Royal Society	UK
Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society	Government of the Republic of Korea	South Korea
Montréal Declaration: Responsible AI	Université de Montréal	Canada
OpenAI Charter	OpenAI	USA
Our principles	Google	USA
Policy Recommendations on Augmented Intelligence in Health Care H-480.940	American Medical Association (AMA)	USA
Position on Robotics and Artificial Intelligence	The Greens (Green Working Group Robots)	EU
Preparing for the future of Artificial Intelligence	Executive Office of the President; National Science and Technology Council; Committee on Technology	USA
Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Fairness, Accountability, and Transparency in Machine Learning (FATML)	n.a.
Principles of robotics	Engineering and Physical Sciences Research Council UK (EPSRC)	UK
Privacy and Freedom of Expression In the Age of Artificial Intelligence	Privacy International & Article 19	international
Report of COMEST on Robotics Ethics	COMEST/UNESCO	international
Report on Artificial Intelligence and Human Society (Unofficial translation)	Advisory Board on Artificial Intelligence and Human Society	Japan
Report with recommendations to the Commission on Civil Law Rules on Robotics	European Parliament	EU
Responsible AI and robotics. An ethical framework.	Accenture UK	UK
Responsible bots: 10 guidelines for developers of conversational AI	Microsoft	USA
SAP's guiding principles for artificial intelligence	SAP	Germany
Science, Law and Society (SLS) Initiative	The Future Society	USA
Sony Group AI Ethics Guidelines	SONY	Japan
Statement on Algorithmic Transparency and Accountability	Association for Computing Machinery (ACM)	USA
Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems	European Commission, European Group on Ethics in Science and New Technologies	EU
Tenets	Partnership on AI	n.a.
The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	AI Now Institute	USA
The Ethics of Code: Developing AI for Business with Five Core Principles	Sage	UK
The Japanese Society for Artificial Intelligence Ethical Guidelines	Japanese Society for Artificial Intelligence	Japan
The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation	Future of Humanity Institute; University of Oxford; Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; OpenAI	international
The National Artificial Intelligence Research and Development Strategic Plan	National Science and Technology Council; Networking and Information Technology Research and Development Subcommittee	USA
The responsible AI framework	PriceWaterhouseCoopers UK	UK
The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems	Access Now; Amnesty International	international
Tieto's AI ethics guidelines	Tieto	Finland
Top 10 Principles for Ethical Artificial Intelligence	UNI Global Union	international
Unified Ethical Frame for Big Data Analysis. IAF Big Data Ethics Initiative, Part A	The Information Accountability Foundation	UK
Universal Guidelines for Artificial Intelligence	The Public Voice	international
Work in the age of artificial intelligence. Four perspectives on the economy, employment, skills and ethics	Ministry of Economic Affairs and Employment	Finland

12. Appendix C: The Responsible AI Canvas

RAI Canvas		Project title:	Date:	Authors:
Purpose/motivation Who is affected? Individuals/collectives/communities/categories	(development) Processes Engineering - Method - Reporting - Version control - Version: $\alpha - \beta$ – production ? Requirements - List - Elicitation method AI methods - ML: SL/UL/RL/CNN... - KR: XS/causal/bayes... Evaluation - Provability - Robustness tests Interpretability - Transparency - Explanation Bias and fairness - Detection - Intervention	Data Sources - Which sources and owners - Collection and curation process Sharing - IP - Constraints and contracts Legislative context - GDPR compliance Governance - Access - Storage - Aging Ethics approval - Yes – no – n/a – pending ?	Application/use Behaviour - Positive effects - Negative effects Conflicts / dilemmas - Identification - Handling - Human control processes Contextualisation - Explicit? - Limitations Failure handling - Redress - Consequences - Security Beneficial (SDGs) 	Evidence Attached – can produce - n/a - none
Who decides? - Stakeholder involvement - Decision processes - Openness	Values/principles - Value elicitation - Value representation (norms) - Methods (e.g. DfV)	Responsibility chain - Who is responsible for what? - Communication - Liability	Evidence Attached – can produce - n/a - none	Evidence Attached – can produce - n/a - none
Short description project:				

13. Appendix D

List⁷ of resources in AI4EU catalogue classified as “AI in Health”, illustration of the quality of current ethical evaluation (in comparison with the evaluation according to the abbreviated assessment proposed in Section 7.2).

AI in health				
Project name		URL	GDPR requirements	Trustworthy AI
Covid-19 chest CT-scan Dataset	This is a sample of the Covid-19 chest CT-scan Dataset	https://www.ai4eu.eu/resource/covid-19-chest-ct-scan-dataset	The dataset has been irreversibly de-identified. It is GDPR-compliant	The dataset has been irreversibly de-identified. Data reflect only the population tested to COVID-19 by CT-scans and does not contain any kind of ethnic information.
CVD-Dataset	Docker container providing access to the repository storing the Cardio-Vascular Disease dataset.	https://www.ai4eu.eu/resource/cvd-dataset	No impact.	N/A
r.AID.ologist	Case-Based Reasoning, Deep Learning supported framework to provide writing support to radiologists	https://www.ai4eu.eu/resource/raidologist	As it is specifically devised for medical data, the risks are high, as this type of data is highly sensitive. Therefore, all data introduced into the framework should be properly anonymized.	The development has been created from a research project using only public and available datasets and libraries.
ABELE	Adversarial Black box Explainer generating Latent Exemplars	https://www.ai4eu.eu/resource/abele	The method provides an explainer capable of creating an explanation in the form of augmented images. This approach addresses the	The explanation layer implemented in ABLE is intended to augment the trust and confidence of the analyst towards AI-based decision support systems. In particular, for black-box

⁷ As per November 2020.

			"explicability" requirement of the GDPR, where a requirement is fixed for automated decision processes that have an impact on humans.	decision systems, the system may improve the insight of the analyst on the internal strategy of the AI algorithm.
ISIC CNN	Deep Neural Network trained for the classification of dermoscopic images into nine categories.	https://www.ai4eu.eu/resource/isic-cnn	No impact	We followed the guidelines: (1) lawful - respecting all applicable laws and regulations (2) ethical - respecting ethical principles and values (3) robust - both from a technical perspective while taking into account its social environment
VERIFAI	VERIFAI (Verifiable and Explainable Risk Forecasting Artificial Intelligence) Framework	https://www.ai4eu.eu/resource/verifai	We do not disclosure any confidential data.	We are aware and respect the 7 requirements. We contribute in specific for further specification in terms of accountability and transparency.
Clinical Use Cases	3 Clinical Use Cases with supporting databases are offered to the community	https://www.ai4eu.eu/resource/clinical-use-cases	Data is anonymised	