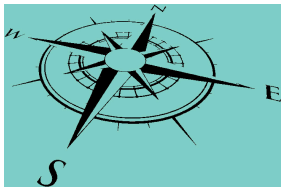
THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI) for self assessment

This document contains the final Assessment List for Trustworthy AI (ALTAI) presented by the AI HLEG. This Assessment List for Trustworthy AI (ALTAI) is intended for self-evaluation purposes. It provides an initial approach for the evaluation of Trustworthy AI.

Contents

Vision....4
Mission....4
Values....4
1. Agency & Oversight....7
1.1. Agency & Autonomy....7
1.2. Oversight....9
1.3. Training....10
1.4. Adverse Effects....11
1.5. Cessation....11
1.6. Oversight & Controls....11
2. Robustness & Safety....13
2.1. Resilience & Security....13
2.2. Safety....15
2.3. Accuracy....17
2.4. Reliability, Fall-Back & Reproducibility....18
3. Privacy & Data Governance....21
3.1. Privacy & Protection....21
3.2. Data Protection....21
4. Transparency....24
4.1. Traceability....24
4.2. Explainability....25
4.3. Communication....26
5. Diversity, Non-Discrimination & Fairness....27
5.1. Bias....27
5.2. Accessibility & Universal Design....30
5.3. Stakeholder Participation....32
6. Well-Being....33
6.1. Environment....33
6.2. Work & Skills....34
6.3. Society & Democracy....35
7. Accountability....37
7.1. Auditability....37
7.2. Risk Management....37
Administrative Information....40



HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE (AIHLEG)

Stakeholder(s):

European Commission

AI Teams :
This Assessment List for Trustworthy AI (ALTAI) is best completed involving a multidisciplinary team of people. These could be from within and/or outside your organisation with specific competences or expertise on each of the requirements and related questions. Among the stakeholders you may find for example the following:

AI Designers :
AI designers bridge the gap between AI capabilities and user needs. For example, they can create prototypes showing some novel AI capabilities and how they might be used if the product is deployed, prior to the possible development of the AI product. AI designers also work with development teams to better understand user needs and how to build technology that addresses those needs. Additionally, they can support AI developers by designing platforms to support data collection and annotation, ensuring that data collection respects some properties (such as safety and fairness)

AI Developers :
An AI developer is someone who performs some of the tasks included in the AI development. AI development is the process of conceiving, specifying, designing, training, programming, documenting, testing, and bug fixing involved in creating and maintaining AI applications, frameworks, or other AI components. It includes writing and maintaining the AI source code, as well as all that is involved between the conception of the software through to the final manifestation and use of the software.

Data Scientists

Procurement Officers

Procurement Specialists

Front-End Staff :
that will use or work with the AI system

Legal/Compliance Officers

Management

Ethics Review Boards :
An AI Ethics Review Board or AI Ethics Committee should be composed of a diverse group of stakeholders and expertises, including gender, background, age and other factors. The purpose for which the AI Ethics Board is created should be clear to the organisation establishing it and the members who are invited to join it. The members should have an independent role that is not influenced by any economic or other considerations. Bias and conflicts of interest should be avoided. The overall size can vary depending on the scope of the task. Both the authority the AI Ethics Review Board has and the access to information should be proportionate to their ability to fulfill the task to their best possible ability.

Data Protection Officers (DPO) :
This denotes an expert on data protection law. The function of a DPO is to internally monitor a public or private organisation's compliance with GDPR. Public or private organisations must appoint DPOs in the following circumstances: (i) data processing activities are carried out by a public authority or body, except for courts acting in their judicial capacity; (ii) the processing of personal data requires regular and systematic monitoring of individuals on a large scale; (iii) the processing of personal data reveals sensitive information like racial of ethnic origin, political opinions, religious or philosophical beliefs, or refers to criminal convictions and offences. A DPO must be independent of the appointing organisation.

End-Users :
An end-user is the person that ultimately uses or is intended to ultimately use the AI system. This could either be a consumer or a professional within a public or private organisation. The end-user stands in contrast to users who support or maintain the product, such as system administrators, database administrators, information technology experts, software professionals and computer technicians.

Red Reams :
Red teaming is the practice whereby a red team or independent group challenges an organisation to improve its effectiveness by assuming an adversarial role or point of view. It is often used to help identify and address potential security vulnerabilities.

Vision

Trustworthy AI

Mission

To facilitate evaluation of AI systems for trustworthiness

Values

Human Rights: This Assessment List (ALTAI) is firmly grounded in the protection of people's fundamental rights, which is the term used in the European Union to refer to human rights enshrined in the EU Treaties, the Charter of Fundamental Rights (the Charter) , and international human rights Law.

Flexibility: This Assessment List for Trustworthy AI (ALTAI) is intended for flexible use: organisations can draw on elements relevant to the particular AI system from this Assessment List for Trustworthy AI (ALTAI) or add elements to it as they see fit, taking into consideration the sector they operate in.

Understanding: It helps organisations understand what Trustworthy AI is, in particular what risks an AI system might generate, and how to minimize those risks while maximising the benefit of AI.

Risk Mitigation: It is intended to help organisations identify how proposed AI systems might generate risks, and to identify whether and what kind of active measures may need to be taken to avoid and minimise those risks.

Awareness: It raises awareness of the potential impact of AI on society, the environment, consumers, workers and citizens (in particular children and people belonging to marginalised groups).

Engagement: Organisations will derive the most value from this Assessment List (ALTAI) by active engagement with the questions it raises, which are aimed at encouraging thoughtful reflection to provoke appropriate action and nurture an organisational culture committed to developing and maintaining Trustworthy AI systems.

Reflection

Involvement: It encourages the involvement of all relevant stakeholders.

Insight: It helps to gain insight on whether meaningful and appropriate solutions or processes to accomplish adherence to the seven requirements (as outlined above) are already in place or need to be put in place.

Solutions: This could be achieved through internal guidelines, governance processes, etc.

Competitiveness: A trustworthy approach is key to enabling 'responsible competitiveness', by providing the foundation upon which all those using or affected by AI systems can trust that their design, development and use are lawful, ethical and robust.

Innovation: This Assessment List for Trustworthy AI (ALTAI) helps foster responsible and sustainable AI innovation in Europe. It seeks to make ethics a core pillar for developing a unique approach to AI, one that aims to benefit, empower and protect both individual human flourishing and the common good of society.

Action

Trust: We believe that this will enable Europe and European organisations to position themselves as global leaders in cutting-edge AI worthy of our individual and collective trust.

Accessibility: Extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use (which includes direct use or use supported by assistive technologies).

Accountability: This term refers to the idea that one is responsible for their action – and as a corollary their consequences – and must be able to explain their aims, motivations, and reasons. Accountability has several dimensions. Accountability is sometimes required by law. For example, the General Data Protection Regulation (GDPR) requires organisations that process personal data to ensure security measures are in place to prevent data

breaches and report if these fail. But accountability might also express an ethical standard, and fall short of legal consequences. Some tech firms that do not invest in facial recognition technology in spite of the absence of a ban or technological moratorium might do so out of ethical accountability considerations.

Accuracy: The goal of an AI model is to learn patterns that generalize well for unseen data. It is important to check if a trained AI model is performing well on unseen examples that have not been used for training the model. To do this, the model is used to predict the answer on the test dataset and then the predicted target is compared to the actual answer. The concept of accuracy is used to evaluate the predictive capability of the AI model. Informally, accuracy is the fraction of predictions the model got right. A number of metrics are used in machine learning (ML) to measure the predictive accuracy of a model. The choice of the accuracy metric to be used depends on the ML task.

Reliability: AI reliability -- An AI system is said to be reliable if it behaves as expected, even for novel inputs on which it has not been trained or tested earlier.

Auditability: Auditability refers to the ability of an AI system to undergo the assessment of the system's algorithms, data and design processes. This does not necessarily imply that information about business models and Intellectual Property related to the AI system must always be openly available. Ensuring traceability and logging mechanisms from the early design phase of the AI system can help enable the system's auditability.

Governance: Data governance -- Data governance is a term used on both a macro and a micro level. On the macro level, data governance refers to the governing of cross-border data flows by countries, and hence is more precisely called international data governance. On the micro level, data governance is a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data, and data controls are implemented that support business objectives. The key focus areas of data governance include data availability, usability, consistency, integrity, and sharing. It also regards establishing processes to ensure effective data management throughout the enterprise such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be used by the entire organization.

Explainability: Feature of an AI system that is intelligible to non-experts. An AI system is intelligible if its functionality and operations can be explained non technically to a person not skilled in the art.

Fairness: Fairness refers to a variety of ideas known as equity, impartiality, egalitarianism, non-discrimination and justice. Fairness embodies an ideal of equal treatment between individuals or between groups of individuals. This is what is generally referred to as 'substantive' fairness. But fairness also encompasses a procedural perspective, that is the ability to seek and obtain relief when individual rights and freedoms are violated.

Fault Tolerance: Fault tolerance is the property that enables a system to continue operating properly in the event of the failure of (or one or more faults within) some of its components. If its operating quality decreases at all, the decrease is proportional to the severity of the failure, as compared to a naively designed system, in which even a small failure can cause total breakdown. Fault tolerance is particularly sought after in high-availability or safetycritical systems. Redundancy or duplication is the provision of additional functional capabilities that would be unnecessary in a fault-free environment. This can consist of backup components that automatically 'kick in' if one component fails.

Oversight: Human oversight, human-in-the-loop, human-on-the-loop, human-in-command --Human oversight helps ensure that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. Human-in-the-loop refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. Humanon-the-loop refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. Human-in-command refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system. Moreover, it must be ensured that public enforcers have the ability to exercise oversight in line with their mandate. Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system's application area and potential risk. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.

Interpretability: Interpretability refers to the concept of comprehensibility, explainability, or understandability. When an element of an AI system is interpretable, this means that it is possible at least for an external observer to understand it and find its meaning.

Continual Learning: Online continual learning -- The ability to continually learn over time by accommodating new knowledge while retaining previously learned experiences is referred to as continual or lifelong learning. Learning continually is crucial for agents and robots operating in changing environments and required to acquire, fine-tune, adapt, and transfer increasingly complex representations of knowledge. Such a continuous learning task has represented a long-standing challenge for machine learning and neural networks and, consequently, for the development of artificial intelligence (AI) systems. The main issue of computational models regarding lifelong learning is that they are prone to catastrophic forgetting or catastrophic interference, i.e., training a model with new information interferes with previously learned knowledge.

Reproducibility: Reproducibility refers to the closeness between the results of two actions, such as two scientific experiments, that are given the same input and use the methodology, as described in a corresponding scientific evidence (such as a scientific publication). A related concept is replication, which is the ability to independently achieve non-identical conclusions that are at least similar, when differences in sampling, research procedures and data analysis methods may exist. Reproducibility and replicability together are among the main tools of the scientific method.

Robustness: Robustness AI -- Robustness of an AI system encompasses both its technical robustness (appropriate in a given context, such as the application domain or life cycle phase) and as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. Robustness is the third of the three components necessary for achieving Trustworthy AI.

Traceability: Ability to track the journey of a data input through all stages of sampling, labelling, processing and decision making.

Trustworthiness: Trustworthy AI --Trustworthy AI has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Trustworthy AI concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors that are part of the AI system's life cycle.

1. Agency & Oversight

Support human agency and human decision-making.

Human Agency and Oversight — AI systems should support human agency and human decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both: act as enablers for a democratic, flourishing and equitable society by supporting the user's agency; and uphold fundamental rights, which should be underpinned by human oversight. In this section AI systems are assessed in terms of their respect for human agency and autonomy as well as human oversight. Glossary: AI System; Autonomous AI System; End User; Human-in-Command; Human-inthe-Loop; Human-on-the-Loop; Self-learning AI System; Subject; User.

1.1. Agency & Autonomy

Deal with the effect AI systems can have on human behaviour.

Human Agency and Autonomy — This subsection deals with the effect AI systems can have on human behaviour in the broadest sense. It deals with the effect of AI systems that are aimed at guiding, influencing or supporting humans in decision making processes, for example, algorithmic decision support systems, risk analysis/prediction systems (recommender systems, predictive policing, financial risk analysis, etc.). It also deals with the effect on human perception and expectation when confronted with AI systems that 'act' like humans. Finally, it deals with the effect of AI systems on human affection, trust and (in)dependence.

Performance Indicators

1.1.1 Decision Making

Description	Type	Yes/No	Start Date	End Date
Determine if AI systems are designed to interact, guide or take decisions.	Target			
	Actual			

Is the AI system designed to interact, guide or take decisions by human end-users that affect humans or society?

1.1.1.1 Decisions, Content, Advice & Outcomes

Description	Type	Yes/No	Start Date	End Date
Determine whether AI systems may confuse end-users or subjects on whether decisions, content, advice or outcomes result from algorithms.	Target			
	Actual			

Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision?

1.1.1.2 Awareness

Description	Type	Status	Start Date	End Date
Make end-users or other subjects aware when decisions, content, advice or outcomes result from algorithms.	Target			
	Actual			

Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision?

1.1.2 Interaction

Description	Type	Status	Start Date	End Date
Avoid confusing end-users or subjects on whether they are interacting with humans or AI systems.	Target			
	Actual			

Could the AI system generate confusion for some or all end-users or subjects on whether they are interacting with a human or AI system?

1.1.2.1 Notice

Description	Type	Status	Start Date	End Date
Inform end-users or subjects when they are interacting with AI systems.	Target			
	Actual			

Are end-users or subjects informed that they are interacting with an AI system?

1.1.3 Autonomy

Description	Type	Yes/No	Start Date	End Date
Determine whether the system may generate over-reliance by end-users.	Target			
	Actual			

Could the AI system affect human autonomy by generating over-reliance by end-users?

1.1.3.1 Reliance

Description	Type	Status	Start Date	End Date
Avoid over-reliance on AI systems.	Target			
	Actual			

Did you put in place procedures to avoid that end-users over-rely on the AI system?

1.1.4 Inteference

Description	Type	Status	Start Date	End Date
Avoid interfering with the end-users' decision-making processes in unintended and undesirable ways.	Target			
	Actual			

Could the AI system affect human autonomy by interfering with the end-user’s decision-making process in any other unintended and undesirable way?

1.1.4.1 Autonomy

Description	Type	Status	Start Date	End Date
Avoid inadvertent effects on human autonomy.	Target			
	Actual			

Did you put in place any procedure to avoid that the AI system inadvertently affects human autonomy?

1.1.5 Interaction

Description	Type	Yes/No	Start Date	End Date
Determine whether the system simulates social interaction with or between end-users or subjects.	Target			
	Actual			

Does the AI system simulate social interaction with or between end-users or subjects?

1.1.6 Risks

Description	Type	Yes/No	Start Date	End Date
Determine whether AI systems risk creating human attachment, stimulating addictive behaviour, or manipulating user behaviour.	Target			
	Actual			

Does the AI system risk creating human attachment, stimulating addictive behaviour, or manipulating user behaviour? Depending on which risks are possible or likely, please answer the questions below:

1.1.6.1 Attachment

Description	Type	Status	Start Date	End Date
Deal with possible negative consequences associated with disproportionate attachment to the system.	Target			
	Actual			

Did you take measures to deal with possible negative consequences for end-users or subjects in case they develop a disproportionate attachment to the AI System?

1.1.6.2 Addiction

Description	Type	Status	Start Date	End Date
Minimise the risk of addiction.	Target			
	Actual			

Did you take measures to minimise the risk of addiction?

1.1.6.3 Manipulation

Description	Type	Status	Start Date	End Date
Mitigate the risk of manipulation.	Target			
	Actual			

Did you take measures to mitigate the risk of manipulation?

1.2. Oversight

Assess oversight measures through governance mechanisms.

Human Oversight — This subsection helps to self-assess necessary oversight measures through governance mechanisms such as human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approaches. Human-in-the-loop refers to the capability for human intervention in every decision cycle of the system. Human-on-the-loop refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. Human-in-command refers to the capability to oversee the overall

activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the AI system in any particular situation. The latter can include the decision not to use an AI system in a particular situation to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by an AI system.

Performance Indicators

1.2.1 Determinations

Description	Type	Status	Start Date	End Date
Make the appropriate determinations regarding AI systems.	Target			
	Actual			

Please determine whether the AI system (choose as many as appropriate):

1.2.1 Learning & Autonomy

Description	Type	Yes/No	Start Date	End Date
Determine whether AI systems are self-learning or autonomous.	Target			
	Actual			

Is a self-learning or autonomous system

1.2.2 Humans-in-the-Loop

Description	Type	Yes/No	Start Date	End Date
Determine whether AI systems are overseen by Humans-in-the-Loop.	Target			
	Actual			

Is overseen by a Human-in-the-Loop

1.2.3 Humans-on-the-Loop

Description	Type	Yes/No	Start Date	End Date
Determine whether AI systems are overseen by Humans-on-the-Loop.	Target			
	Actual			

Is overseen by a Human-on-the-Loop

1.2.4 Humans-in-Command

Description	Type	Yes/No	Start Date	End Date
Determine whether AI systems are overseen by a Humans-in-Command.	Target			
	Actual			

Is overseen by a Human-in-Command

1.3. Training

Train humans on how to exercise oversight.

Performance Indicators

1.3.1 Training

Description	Type	Status	Start Date	End Date
Train the humans on how to exercise oversight.	Target			
	Actual			

Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?

1.4. Adverse Effects

Establish detection and response mechanisms for undesirable effects.

Performance Indicators

1.4.1 Detection & Response

Description	Type	Status	Start Date	End Date
Establish detection and response mechanisms for adverse effects of the AI system.	Target			
	Actual			

Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?

1.5. Cessation

Ensure a 'stop button' or procedure to safely abort operations.

Performance Indicators

1.5.1 Abortion

Description	Type	Yes/No	Start Date	End Date
Establish procedures to safely abort operations when needed.	Target			
	Actual			

Did you ensure a 'stop button' or procedure to safely abort an operation when needed?

1.6. Oversight & Controls

Implement oversight and control measures to reflect the self-learning and autonomous nature of AI systems.

Performance Indicators

1.6.1 Oversight & Control

Description	Type	Yes/No	Start Date	End Date
Implement oversight and control measures to reflect the self-learning or autonomous nature of the AI system.	Target			
	Actual			

Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?

2. Robustness & Safety

Develop AI reliable systems with a preventative approach to risk.

Technical Robustness and Safety — A crucial requirement for achieving Trustworthy AI systems is their dependability (the ability to deliver services that can justifiably be trusted) and resilience (robustness when facing changes). Technical robustness requires that AI systems are developed with a preventative approach to risks and that they behave reliably and as intended while minimising unintentional and unexpected harm as well as preventing it where possible. This should also apply in the event of potential changes in their operating environment or the presence of other agents (human or artificial) that may interact with the AI system in an adversarial manner. The questions in this section address four main issues: 1) security; 2) safety; 3) accuracy; and 4) reliability, fall-back plans and reproducibility. Glossary: Accuracy; AI Bias; AI System; AI Reliability; AI Reproducibility; (Low) Confidence Score; Continual Learning; Data Poisoning; Model Evasion; Model Inversion; Pen Test; Redteam.

2.1. Resilience & Security

Make AI systems secure and resilient to attacks.

Resilience to Attack and Security

Performance Indicators

2.1.1 Bad Effects

Description	Type	Status	Start Date	End Date
Determine whether the system could have adversarial, critical or damaging effects.	Target			
	Actual			

Could the AI system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?

2.1.2 Cybersecurity

Description	Type	Status	Start Date	End Date
Certify the system for cybersecurity.	Target			
	Actual			

Is the AI system certified for cybersecurity (e.g. the certification scheme created by the Cybersecurity Act in Europe) or is it compliant with specific security standards?

2.1.3 Attack Exposure

Description	Type	Status	Start Date	End Date
Assess exposure of the system to cyber-attacks.	Target			
	Actual			

How exposed is the AI system to cyber-attacks?

2.1.3.1 Attack Assessment

Description	Type	Status	Start Date	End Date
Assess the vulnerabilities of the system.	Target			
	Actual			

Did you assess potential forms of attacks to which the AI system could be vulnerable?

2.1.3.2 Vulnerability Assessment

Description	Type	Status	Start Date	End Date
Consider the different types of vulnerabilities and potential entry points for attacks.	Target			
	Actual			

Did you consider different types of vulnerabilities and potential entry points for attacks such as:

2.1.3.2.1 Data Poisoning

Description	Type	Status	Start Date	End Date
Consider the risk that the data could be poisoned.	Target			
	Actual			

Data poisoning (i.e. manipulation of training data)

2.1.3.2.2 Model Evasion

Description	Type	Status	Start Date	End Date
Consider the risk the model could be evaded.	Target			
	Actual			

Model evasion (i.e. classifying the data according to the attacker's will)

2.1.3.2.3 Model Inversion

Description	Type	Status	Start Date	End Date
Consider the risk of model inversion.	Target			
	Actual			

Model inversion (i.e. infer the model parameters)

2.1.4 Integrity, Robustness & Security

Description	Type	Status	Start Date	End Date
Ensure the integrity, robustness and security of the system.	Target			
	Actual			

Did you put measures in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle?

2.1.5 Testing

Description	Type	Status	Start Date	End Date
Red-team/pentest the system.	Target			
	Actual			

Did you red-team/pentest the system?

2.1.6 Information

Description	Type	Status	Start Date	End Date
Inform end-users of the duration of security coverage and updates.	Target			
	Actual			

Did you inform end-users of the duration of security coverage and updates?

2.1.6.1 Timeframe

Description	Type	Start Date	End Date	Days
Provide security updates for the AI system within the specified timeframe.	Target			
	Actual			

What length is the expected timeframe within which you provide security updates for the AI system?

2.2. Safety

Ensure that AI systems are safe.

General Safety

Performance Indicators

2.2.1 Definitions

Description	Type	Status	Start Date	End Date
Define risks, risk metrics and risk levels of the system for each use case.	Target			
	Actual			

Did you define risks, risk metrics and risk levels of the AI system in each specific use case?

2.2.1.1 Metrics & Assessment

Description	Type	Status	Start Date	End Date
Continuously measure and assess risks.	Target			
	Actual			

Did you put in place a process to continuously measure and assess risks?

2.2.1.2 Information

Description	Type	Status	Start Date	End Date
Inform end-users and subjects of risks.	Target			
	Actual			

Did you inform end-users and subjects of existing or potential risks?

2.2.2 Threats & Consequences

Description	Type	Status	Start Date	End Date
Identify the threats and consequences.	Target			
	Actual			

Did you identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible consequences?

2.2.2.1 Misuse

Description	Type	Status	Start Date	End Date
Assess the risk of malicious use, misuse or inappropriate use of the system.	Target			
	Actual			

Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system?

2.2.2.2 Safety

Description	Type	Status	Start Date	End Date
Define safety criticality levels of the possible consequences of faults or misuse of the system.	Target			
	Actual			

Did you define safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system?

2.2.3 Dependencies

Description	Type	Status	Start Date	End Date
Assess the dependency of a critical decisions on the system's stability and reliability.	Target			
	Actual			

Did you assess the dependency of a critical AI system’s decisions on its stable and reliable behaviour?

2.2.3.1 Requirements

Description	Type	Status	Start Date	End Date
Align the reliability/testing requirements to the appropriate levels of stability and reliability.	Target			
	Actual			

Did you align the reliability/testing requirements to the appropriate levels of stability and reliability?

2.2.4 Fault Tolerance

Description	Type	Status	Start Date	End Date
Plan for fault tolerance.	Target			
	Actual			

Did you plan fault tolerance via, e.g. a duplicated system or another parallel system (AI-based or 'conventional')?

2.2.5 Mechanism

Description	Type	Status	Start Date	End Date
Evaluate when the AI system has been changed to merit a new review of its technical robustness and safety.	Target			
	Actual			

Did you develop a mechanism to evaluate when the AI system has been changed to merit a new review of its technical robustness and safety?

2.3. Accuracy

Ensure the accuracy of AI systems.

Accuracy

Performance Indicators

2.3.1 Bad Consequences

Description	Type	Yes/No	Start Date	End Date
Determine whether a low level of accuracy of the system could result in critical, adversarial or damaging consequences.	Target			
	Actual			

Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?

2.3.2 Data Quality

Description	Type	Status	Start Date	End Date
Ensure the data used to develop the system is up-to-date, of high quality, complete and representative.	Target			
	Actual			

Did you put in place measures to ensure that the data (including training data) used to develop the AI system is up-to-date, of high quality, complete and representative of the environment the system will be deployed in?

2.3.3 Accuracy

Description	Type	Status	Start Date	End Date
Monitor and document the system's accuracy.	Target			
	Actual			

Did you put in place a series of steps to monitor, and document the AI system's accuracy?

2.3.4 Invalidation

Description	Type	Status	Start Date	End Date
Consider whether the system's operation could invalidate the data or assumptions on which it was trained.	Target			
	Actual			

Did you consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects?

2.3.5 Communication

Description	Type	Status	Start Date	End Date
Ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated.	Target			
	Actual			

Did you put processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated?

2.4. Reliability, Fall-Back & Reproducibility

Plan for the reliability, fall-back, and reproducibility of AI systems.

Reliability, Fall-back plans and Reproducibility

Performance Indicators

2.4.1 Bad Consequences

Description	Type	Yes/No	Start Date	End Date
Determine whether the system could cause critical, adversarial, or damaging consequences.	Target			
	Actual			

Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?

2.4.1.1 Goal Monitoring

Description	Type	Status	Start Date	End Date
Monitor if the system is meeting the intended goals.	Target			
	Actual			

Did you put in place a well-defined process to monitor if the AI system is meeting the intended goals?

2.4.1.2 Reproducibility

Description	Type	Status	Start Date	End Date
Test whether specific contexts or conditions must be taken into account to ensure reproducibility.	Target			
	Actual			

Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility?

2.4.2 Verification & Validation

Description	Type	Status	Start Date	End Date
Verify, validate, and document the system's reliability and reproducibility.	Target			
	Actual			

Did you put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the AI system's reliability and reproducibility?

2.4.2.1 Documentation & Operationalisation

Description	Type	Status	Start Date	End Date
Document and operationalise processes for the testing and verification of the reliability and reproducibility of the system.	Target			
	Actual			

Did you clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system?

2.4.3 Fallback

Description	Type	Status	Start Date	End Date
Define plans to address AI system errors and institute procedures to trigger them.	Target			
	Actual			

Did you define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them?

2.4.4 Confidence

Description	Type	Status	Start Date	End Date
Institute a procedure for handling cases when the system yields results with low confidence scores.	Target			
	Actual			

Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score?

2.4.5 Learning

Description	Type	Yes/No	Start Date	End Date
Determine if the system is using (online) continual learning	Target			
	Actual			

Is your AI system using (online) continual learning?

2.4.5.1 Bad Consequences

Description	Type	Status	Start Date	End Date
Consider negative consequences from the system learning novel or unusual methods to score well on its objective function.	Target			
	Actual			

Did you consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function?

3. Privacy & Data Governance

Ensure the quality and integrity of data used in AI systems and protect privacy.

Privacy and Data Governance — Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy. Glossary: Aggregation and Anonymisation; AI System; Data Governance; Data Protection Impact Assessment (DPIA); Data Protection Officer (DPO); Encryption; Lifecycle; Pseudonymisation; Standards; Use Case.

3.1. Privacy & Protection

Assess the impacts of AI systems on privacy and data protection.

Privacy — This subsection helps to self-assess the impact of the AI system's impact on privacy and data protection, which are fundamental rights that are closely related to each other and to the fundamental right to the integrity of the person, which covers the respect for a person's mental and physical integrity.

Performance Indicators

3.1.1 Rights & Morality

Description	Type	Status	Start Date	End Date
Consider the impact of AI systems on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection.	Target			
	Actual			

Did you consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection?

3.1.2 Flagging

Description	Type	Status	Start Date	End Date
Establish mechanisms to flag issues related to privacy in AI systems.	Target			
	Actual			

Depending on the use case, did you establish mechanisms that allow flagging issues related to privacy concerning the AI system?

3.2. Data Protection

Protect data in AI systems.

Data Governance — This subsection helps to self-assess the adherence of the AI system('s use) to various elements concerning data protection.

Performance Indicators

3.2.1 Personal Data

Description	Type	Status	Start Date	End Date
Identify usage and processing of personal data.	Target			
	Actual			

Is your AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)?

3.2.2 GDPR

Description	Type	Status	Start Date	End Date
Put in place measures under the General Data Protection Regulation.	Target			
	Actual			

Did you put in place any of the following measures some of which are mandatory under the General Data Protection Regulation (GDPR), or a non-European equivalent?

- Data Protection Impact Assessment (DPIA);
- Designate a Data Protection Officer (DPO) and include them at an early state in the development, procurement or use phase of the AI system;
- Oversight mechanisms for data processing (including limiting access to qualified personnel, mechanisms for logging data access and making modifications);
- Measures to achieve privacy-by-design and default (e.g. encryption, pseudonymisation, aggregation, anonymisation);
- Data minimisation, in particular personal data (including special categories of data).

3.2.2.1 Withdrawals, Objections & Forgetting

Description	Type	Status	Start Date	End Date
Implement the right to withdraw consent, the right to object and the right to be forgotten.	Target			
	Actual			

Did you implement the right to withdraw consent, the right to object and the right to be forgotten into the development of the AI system?

3.2.2.2 System Life Cycles

Description	Type	Status	Start Date	End Date
Consider the privacy and data protection implications over AI system life cycles.	Target			
	Actual			

Did you consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's life cycle?

3.2.3 Non-Personal Data

Description	Type	Status	Start Date	End Date
Consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data.	Target			
	Actual			

Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data?

3.2.4 Standards & Protocols

Description	Type	Status	Start Date	End Date
Align AI systems with relevant standards and protocols.	Target			
	Actual			

Did you align the AI system with relevant standards (e.g. ISO25, IEEE26) or widely adopted protocols for (daily) data management and governance?

4. Transparency

Ensure transparency of AI systems.

Transparency — A crucial component of achieving Trustworthy AI is transparency which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system. Glossary: AI System; End-User; Explicability; Lifecycle; Subject; Traceability; Workflow of the Model.

4.1. Traceability

Traceability — This subsection helps to self-assess whether the processes of the development of the AI system, i.e. the data and processes that yield the AI system's decisions, is properly documented to allow for traceability, increase transparency and, ultimately, build trust in AI in society.

Performance Indicators

4.1.1 Traceability Metrics

Description	Type	Status	Start Date	End Date
Address the traceability of the system during its entire lifecycle.	Target			
	Actual			

Did you put in place measures that address the traceability of the AI system during its entire lifecycle?

4.1.1.1 Data Quality

Description	Type	Metrics	Start Date	End Date
Assess the quality of the input data.	Target			
	Actual			

Did you put in place measures to continuously assess the quality of the input data to the AI system?

4.1.1.2 Data Tracing

Description	Type	Status	Start Date	End Date
Trace which data was used by the AI system.	Target			
	Actual			

Can you trace back which data was used by the AI system to make a certain decision(s) or recommendation(s)?

4.1.1.3 Decision Tracing

Description	Type	Status	Start Date	End Date
Trace which model or rules led to the decision(s) or recommendation(s).	Target			
	Actual			

Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system?

4.1.1.4 Metrics

Description	Type	Status	Start Date	End Date
Implement measures to assess the quality of the output(s) of the system.	Target			
	Actual			

Did you put in place measures to continuously assess the quality of the output(s) of the AI system?

4.1.1.5 Logging

Description	Type	Status	Start Date	End Date
Put logging practices in place to record the decision(s) or recommendation(s) of the system.	Target			
	Actual			

Did you put adequate logging practices in place to record the decision(s) or recommendation(s) of the AI system?

4.2. Explainability

Assess the explainability of AI systems.

Explainability — This subsection helps to self-assess the explainability of the AI system. The questions refer to the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes. Explainability is crucial for building and maintaining users' trust in AI systems. AI driven decisions – to the extent possible – must be explained to and understood by those directly and indirectly affected, in order to allow for contesting of such decisions. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'blackboxes' and require special attention. In those circumstances, other explainability measures (e.g. traceability, auditability and transparent communication on the AI system's capabilities) may be required, provided that the AI system as a whole respects fundamental rights. The degree to which explainability is needed depends on the context and the severity of the consequences of erroneous or otherwise inaccurate output to human life.

Performance Indicators

4.2.1 Explanation

Description	Type	Status	Start Date	End Date
Explain the decision(s) of the system to its users.	Target			
	Actual			

Did you explain the decision(s) of the AI system to the users?

4.2.2 Surveys

Description	Type	Status	Start Date	End Date
Survey the users to learn if they understand the decision(s) of the system.	Target			
	Actual			

Do you continuously survey the users if they understand the decision(s) of the AI system?

4.3. Communication

Communicate AI system capabilities and limitations to the users.

Communication — This subsection helps to self-assess whether the AI system's capabilities and limitations have been communicated to the users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy as well as its limitations.

Performance Indicators

4.3.1 Communication

Description	Type	Status	Start Date	End Date
Communicate to users when they are interacting with AI systems.	Target			
	Actual			

In cases of interactive AI systems (e.g., chatbots, robo-lawyers), do you communicate to users that they are interacting with an AI system instead of a human?

4.3.2 Mechanisms

Description	Type	Status	Start Date	End Date
Inform users about the purpose, criteria and limitations of the decision(s) generated by the system.	Target			
	Actual			

Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?

4.3.2.1 Benefit Communication

Description	Type	Status	Start Date	End Date
Inform users of the benefits of the system.	Target			
	Actual			

Did you communicate the benefits of the AI system to users?

4.3.2.2 Limitation & Risk Communication

Description	Type	Status	Start Date	End Date
Inform users of the technical limitations and potential risks of the system.	Target			
	Actual			

Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/ or error rates?

4.3.2.3 Training & Disclaimers

Description	Type	Status	Start Date	End Date
Provide training material and disclaimers to users on how to adequately use the system.	Target			
	Actual			

Did you provide appropriate training material and disclaimers to users on how to adequately use the AI system?

5. Diversity, Non-Discrimination & Fairness

Enable inclusion and diversity throughout the life cycles of AI systems.

Diversity, Non-discrimination and Fairness — In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness, and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a nontransparent market. Identifiable and discriminatory bias should be removed in the collection phase where possible. AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance. Glossary: AI Bias; AI System; AI Designer; AI Developer; Accessibility; Assistive Technology; End-User; Fairness; Subject; Universal Design; Use Case.

5.1. Bias

Avoid unfair bias.

Avoidance of Unfair Bias

Performance Indicators

5.1.1 Strategy & Procedures

Description	Type	Status	Start Date	End Date
Establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the system.	Target			
	Actual			

Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?

5.1.2 Diversity & Representativeness

Description	Type	Status	Start Date	End Date
Consider diversity and representativeness of end-users and/or subjects in the data.	Target			
	Actual			

Did you consider diversity and representativeness of end-users and/or subjects in the data?

5.1.2.1 Testing

Description	Type	Status	Start Date	End Date
Test for specific target groups or problematic use cases.	Target			
	Actual			

Did you test for specific target groups or problematic use cases?

5.1.2.2 Research & Tools

Description	Type	Status	Start Date	End Date
Research and use publicly available technical tools to improve your understanding of the data, model and performance.	Target			
	Actual			

Did you research and use publicly available technical tools, that are state-ofthe-art, to improve your understanding of the data, model and performance?

5.1.2.3 Testing & Monitoring

Description	Type	Status	Start Date	End Date
Test and monitor for potential biases during the entire lifecycle of the system.	Target			
	Actual			

Did you assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)?

5.1.2.4 Diversity & Representativeness

Description	Type	Status	Start Date	End Date
Consider diversity and representativeness of end-users and or subjects in the data.	Target			
	Actual			

Where relevant, did you consider diversity and representativeness of end-users and or subjects in the data?

5.1.3 Education & Awareness

Description	Type	Status	Start Date	End Date
Help AI designers and AI developers be more aware of the possible bias they can inject in the system.	Target			
	Actual			

Did you put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system?

5.1.4 Issue Flagging

Description	Type	Status	Start Date	End Date
Flag issues related to bias, discrimination or poor performance of the system.	Target			
	Actual			

Did you ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system?

5.1.4.1 Issue Communication

Description	Type	Status	Start Date	End Date
Establish steps and means of communicating on how and to whom issues can be raised.	Target			
	Actual			

Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?

5.1.4.2 Subject Identification

Description	Type	Status	Start Date	End Date
Identify the subjects who could be affected by the system.	Target			
	Actual			

Did you identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end-)users and/or subjects?

5.1.5 Fairness Definition

Description	Type	Yes/No	Start Date	End Date
Use your definition of fairness when setting up the system.	Target			
	Actual			

Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI system?

5.1.5 Other Definitions

Description	Type	Yes/No	Start Date	End Date
Consider other definitions of fairness.	Target			
	Actual			

Did you consider other definitions of fairness before choosing this one?

5.1.5.2 Consultation

Description	Type	Status	Start Date	End Date
Consult with the impacted communities about the definition of fairness.	Target			
	Actual			

Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities?

5.1.5.3 Metrics & Analysis

Description	Type	Status	Start Date	End Date
Apply quantitative metrics and analysis to test the definition of fairness.	Target			
	Actual			

Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?

5.1.5.4 Mechanisms

Description	Type	Statu	Start Date	End Date
Establish mechanisms to ensure fairness.	Target			
	Actual			

Did you establish mechanisms to ensure fairness in your AI system?

5.2. Accessibility & Universal Design

Designed systems to allow all people to use AI products or services, regardless of their age, gender, abilities or characteristics.

Accessibility and Universal Design — Particularly in business-to-consumer domains, AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance. AI systems should not have a one-size-fits-all approach and should consider Universal Design principles addressing the widest possible range of users, following relevant accessibility standards. This will enable equitable access and active participation of all people in existing and emerging computer-mediated human activities and with regard to assistive technologies.

Performance Indicators

5.2.1 Preferences & Abilities

Description	Type	Status	Start Date	End Date
Ensure the system corresponds to the preferences and abilities in society.	Target			
	Actual			

Did you ensure that the AI system corresponds to the variety of preferences and abilities in society?

5.2.2 Usability

Description	Type	Status	Start Date	End Date
Ensure the system is usable by those with special needs or disabilities or at risk of exclusion.	Target			
	Actual			

Did you assess whether the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion?

5.2.2.1 Accessibility & Usability

Description	Type	Status	Start Date	End Date
Ensure that information about the system is accessible and usable to users of assistive technologies.	Target			
	Actual			

Did you ensure that information about, and the AI system's user interface of, the AI system is accessible and usable also to users of assistive technologies (such as screen readers)?

5.2.2.2 Consultation

Description	Type	Status	Start Date	End Date
Involve or consult with end-users or subjects in need for assistive technology.	Target			
	Actual			

Did you involve or consult with end-users or subjects in need for assistive technology during the planning and development phase of the AI system?

5.2.3 Universal Design

Description	Type	Status	Start Date	End Date
Ensure that Universal Design principles are taken into account during the planning and development process.	Target			
	Actual			

Did you ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable?

5.2.4 Impact Analysis

Description	Type	Status	Start Date	End Date
Take into account the impact of the system on the potential end-users and/or subjects.	Target			
	Actual			

Did you take the impact of the AI system on the potential end-users and/or subjects into account?

5.2.4.1 Engagement

Description	Type	Status	Start Date	End Date
Assess whether the team building the system engaged with the end-users and/or subjects.	Target			
	Actual			

Did you assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects?

5.2.4.2 Disproportionate Impacts

Description	Type	Status	Start Date	End Date
Assess whether some groups may be disproportionately affected by the system.	Target			
	Actual			

Did you assess whether there could be groups who might be disproportionately affected by the outcomes of the AI system?

5.2.4.4 Assessment

Description	Type	Status	Start Date	End Date
Assess the risk of the possible unfairness of the system.	Target			
	Actual			

Did you assess the risk of the possible unfairness of the system onto the end-user's or subject's communities?

5.3. Stakeholder Participation

Consult stakeholders who may directly or indirectly be affected by AI systems.

Stakeholder Participation — In order to develop Trustworthy AI, it is advisable to consult stakeholders who may directly or indirectly be affected by the AI system throughout its life cycle. It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organisations.

Performance Indicators

5.3.1 Mechanisms

Description	Type	Status	Start Date	End Date
Include the widest range of possible stakeholders in the system's design and development.	Target			
	Actual			

Did you consider a mechanism to include the participation of the widest range of possible stakeholders in the AI system's design and development?

6. Well-Being

Consider other sentient beings and the environment as stakeholders throughout the AI system's life cycle.

Societal and Environmental Well-being — In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's life cycle. Ubiquitous exposure to social AI systems in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or negatively impact our social relationships and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could equally affect peoples' physical and mental well-being. The effects of AI systems must therefore be carefully monitored and considered. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, for instance the Sustainable Development Goals. Overall, AI should be used to benefit all human beings, including future generations. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. AI systems must not undermine democratic processes, human deliberation or democratic voting systems or pose a systemic threat to society at large.

6.1. Environment

Self-assess the impacts of AI systems on the environment.

Environmental Well-being — This subsection helps to self-assess the (potential) positive and negative impacts of the AI system on the environment. AI systems, even if they promise to help tackle some of the most pressing societal concerns, e.g. climate change, must work in the most environmentally friendly way possible. The AI system’s development, deployment and use process, as well as its entire supply chain, should be assessed in this regard (e.g. via a critical examination of the resource usage and energy consumption during training, opting for less net negative choices). Measures to secure the environmental friendliness of an AI system’s entire supply chain should be encouraged.

Performance Indicators

6.1.1 Negative Impacts

Description	Type	Yes/No	Start Date	End Date
Determine if there are potential negative impacts on the environment.	Target			
	Actual			

Are there potential negative impacts of the AI system on the environment?

6.1.1.1 Impacts

Description	Type	Identification	Start Date	End Date
Identify the potential impacts.	Target			
	Actual			

Which potential impact(s) do you identify?

6.1.2 Mechanisms

Description	Type	Status	Start Date	End Date
Establish mechanisms to evaluate environmental impacts.	Target			
	Actual			

Where possible, did you establish mechanisms to evaluate the environmental impact of the AI system's development, deployment and/or use (for example, the amount of energy used and carbon emissions)?

6.1.2.1 Measures

Description	Type	Status	Start Date	End Date
Define measures to reduce environmental impacts.	Target			
	Actual			

Did you define measures to reduce the environmental impact of the AI system throughout its lifecycle?

6.2. Work & Skills

Assess the impact of AI systems and their use on workers, the relationship between workers and employers, and on skills.

Impact on Work and Skills — AI systems may fundamentally alter the work sphere. They should support humans in the working environment, and aim for the creation of meaningful work. This subsection helps self-assess the impact of the AI system and its use in a working environment on workers, the relationship between workers and employers, and on skills.

Stakeholder(s):

Workers

Employers

Performance Indicators

6.2.1 Impact

Description	Type	Yes/No	Start Date	End Date
Determine whether the system impacts human work and work arrangements.	Target			
	Actual			

Does the AI system impact human work and work arrangements?

6.2.2 Information & Consultation

Description	Type	Status	Start Date	End Date
Inform and consult with impacted workers and their representatives.	Target			
	Actual			

Did you pave the way for the introduction of the AI system in your organisation by informing and consulting with impacted workers and their representatives (trade unions, (European) work councils) in advance?

6.2.3 Measures

Description	Type	Status	Start Date	End Date
Ensure that the impacts on human work are well understood.	Target			
	Actual			

Did you adopt measures to ensure that the impacts of the AI system on human work are well understood?

6.2.3.1 Understanding

Description	Type	Status	Start Date	End Date
Ensure workers understand how the system operates.	Target			
	Actual			

Did you ensure that workers understand how the AI system operates, which capabilities it has and which it does not have?

6.2.4 De-Skilling Risk

Description	Type	Yes/No	Start Date	End Date
Determine if the system could create the risk of de-skilling of the workforce.	Target			
	Actual			

Could the AI system create the risk of de-skilling of the workforce?

6.2.4.1 Measures

Description	Type	Status	Start Date	End Date
Counteract de-skilling risks.	Target			
	Actual			

Did you take measures to counteract de-skilling risks?

6.2.5 New Skills

Description	Type	Yes/No	Start Date	End Date
Determine whether the system promotes or requires new (digital) skills.	Target			
	Actual			

Does the system promote or require new (digital) skills?

6.2.5.1 Training

Description	Type	Status	Start Date	End Date
Provide training opportunities and materials for re- and up-skilling.	Target			
	Actual			

Did you provide training opportunities and materials for re- and up-skilling?

6.3. Society & Democracy

Assess the impact of AI systems from a societal perspective.

Impact on Society at large or Democracy — This subsection helps to self-assess the impact of an AI system from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI

systems should be given careful consideration, particularly in situations relating to the democratic processes, including not only political decision-making but also electoral contexts (e.g. when AI systems amplify fake news, segregate the electorate, facilitate totalitarian behaviour, etc.).

Performance Indicators

6.3.1 Impact

Description	Type	Yes/No	Start Date	End Date
Determine if the system may have a negative impact on society at large or democracy.	Target			
	Actual			

Could the AI system have a negative impact on society at large or democracy?

6.3.1.1 Assessment

Description	Type	Status	Start Date	End Date
Assess the societal impact of the system's use beyond the (end-)user and subject.	Target			
	Actual			

Did you assess the societal impact of the AI system's use beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large?

6.3.1.2 Action

Description	Type	Status	Start Date	End Date
Minimize potential societal harm.	Target			
	Actual			

Did you take action to minimize potential societal harm of the AI system?

6.3.1.3 Measures

Description	Type	Status	Start Date	End Date
Ensure the system does not negatively impact democracy.	Target			
	Actual			

Did you take measures that ensure that the AI system does not negatively impact democracy?

7. Accountability

Put in place mechanisms to ensure responsibility for the development, deployment and/or use of AI systems.

Accountability — The principle of accountability necessitates that mechanisms be put in place to ensure responsibility for the development, deployment and/or use of AI systems. This topic is closely related to risk management, identifying and mitigating risks in a transparent way that can be explained to and audited by third parties. When unjust or adverse impacts occur, accessible mechanisms for accountability should be in place that ensure an adequate possibility of redress. Glossary: Accountability; AI Ethics Review Board; Redress by Design.

7.1. Auditability

Evaluate AI systems.

Auditability — This subsection helps to self-assess the existing or necessary level that would be required for an evaluation of the AI system by internal and external auditors. The possibility to conduct evaluations as well as to access records on said evaluations can contribute to Trustworthy AI. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available.

Performance Indicators

7.1.1 Auditability

Description	Type	Status	Start Date	End Date
Establish mechanisms that facilitate AI system auditability.	Target			
	Actual			

Did you establish mechanisms that facilitate the AI system’s auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system’s processes, outcomes, positive and negative impact)?

7.1.2 Auditing

Description	Type	Status	Start Date	End Date
Ensure that AI systems can be audited.	Target			
	Actual			

Did you ensure that the AI system can be audited by independent third parties?

7.2. Risk Management

Report on and respond to actions or decisions that contribute to the AI system's outcome.

Risk Management — Both the ability to report on actions or decisions that contribute to the AI system's outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, documenting and minimising the potential negative impacts of AI systems is especially crucial for those (in)directly affected. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system. When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs. Such trade-offs should be addressed in a

rational and methodological manner within the state of the art. This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to safety and ethical principles, including fundamental rights. Any decision about which trade-off to make should be well reasoned and properly documented. When adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress.

Performance Indicators

7.2.1 Auditing & Accountability

Description	Type	Status	Start Date	End Date
Determine whether to apply external guidance or third-party auditing processes to oversee ethical concerns and accountability measures.	Target			
	Actual			

Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures?

7.2.1 Third Parties

Description	Type	Status	Start Date	End Date
Consider whether the involvement of third parties goes beyond the development phase.	Target			
	Actual			

Does the involvement of these third parties go beyond the development phase?

7.2.2 Training

Description	Type	Status	Start Date	End Date
Organise risk training.	Target			
	Actual			

Did you organise risk training and, if so, does this also inform about the potential legal framework applicable to the AI system?

7.2.3 Accountability & Ethics

Description	Type	Status	Start Date	End Date
Establish AI ethics review boards.	Target			
	Actual			

Did you consider establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas?

7.2.4 Monitoring & Assessment

Description	Type	Status	Start Date	End Date
Discuss and continuously monitor and assess adherence to the ALTAI.	Target			
	Actual			

Did you establish a process to discuss and continuously monitor and assess the AI system's adherence to this Assessment List for Trustworthy AI (ALTAI)?

7.2.4.1 Conflicts

Description	Type	Status	Start Date	End Date
Identify and document conflicts between the requirements and ethical principles.	Target			
	Actual			

Does this process include identification and documentation of conflicts between the 6 aforementioned requirements or between different ethical principles and explanation of the 'trade-off' decisions made?

7.2.4.2 Training

Description	Type	Status	Start Date	End Date
Provide training to those involved in ensuring adherence to the ALTAI process.	Target			
	Actual			

Did you provide appropriate training to those involved in such a process and does this also cover the legal framework applicable to the AI system?

7.2.5 Third Parties

Description	Type	Status	Start Date	End Date
Establish processes for third parties to report potential vulnerabilities, risks or biases.	Target			
	Actual			

Did you establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?

7.2.5.1 Revisions

Description	Type	Status	Start Date	End Date
Foster revision of risk management processes.	Target			
	Actual			

Does this process foster revision of the risk management process?

7.2.6 Redress

Description	Type	Status	Start Date	End Date
Put redress-by-design mechanisms in place.	Target			
	Actual			

For applications that can adversely affect individuals, have redress by design mechanisms been put in place?

Administrative Information

Start Date:
End Date:

Publication Date: **2020-07-26**
Source: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

Submitter:

Given Name: Owen
Surname: Ambur
Email: Owen.Ambur@verizon.net
Phone:

_b16039dc-ceae-11ea-b252-37280b83ea00