INDEPENDENT

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

SET UP BY THE EUROPEAN COMMISSION

THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI)

for self assessment

Table of Contents

Introduction 3
How to use this Assessment List for Trustworthy AI (ALTAI) 4
REQUIREMENT #1 Human Agency and Oversight 7
Human Agency and Autonomy 7
Human Oversight 8
REQUIREMENT #2 Technical Robustness and Safety 9
Resilience to Attack and Security 9
General Safety 9
Accuracy 10
Reliability, Fall-back plans and Reproducibility 10
REQUIREMENT #3 Privacy and Data Governance 12
Privacy 12
Data Governance 12
REQUIREMENT #4 Transparency 14
Traceability 14
Explainability 14
Communication 15
REQUIREMENT #5 Diversity, Non-discrimination and Fairness 15
Avoidance of Unfair Bias 16
Accessibility and Universal Design 17
Stakeholder Participation 18
REQUIREMENT #6 Societal and Environmental Well-being 18
Environmental Well-being 19
Impact on Work and Skills 19
Impact on Society at large or Democracy 20
REQUIREMENT #7 Accountability 21
Auditability 21
Risk Management 21
Glossary 23

This document was written by the High-Level Expert Group on AI (AI HLEG). It is the third deliverable of the AI HLEG and follows the publication of the group's deliverable of the Ethics Guidelines for Trustworthy AI, published on the 8th of April 2019. The members of the AI HLEG named in this document have contributed to the formulation of the content throughout the running of their mandate. The work was informed by the piloting phase of the original assessment list contained in the Ethics Guidelines for Trustworthy AI, conducted by the European Commission from the 26th of June 2019 to the 1st of December 2019. They support the broad direction of the Assessment List for Trustworthy AI put forward in this document, although they do not necessarily agree with every single statement therein.

The High-Level Expert Group on AI is an independent expert group that was set up by the European Commission in June 2018.

Disclaimer

This Assessment List (ALTAI) is a self-assessment tool. The individual or collective members of the High Level Expert Group on AI do not offer any guarantee as to the compliance of an AI system assessed by using ALTAI with the 7 requirements for Trustworthy AI. Under no circumstances are the individual or collective members of the High Level Expert Group on AI liable for any direct, indirect, incidental, special or consequential damages or lost profits that result directly or indirectly from the use of or reliance on (the results of using) ALTAI.

Contact Charlotte Stix - AI HLEG Coordinator
E-mail CNECT-HLG-AI@ec.europa.eu

European Commission
B-1049 Brussels

Document made public on the 17th of July 2020.

Book: ISBN 978-92-76-20009-3 doi:10.2759/791819 KK-02-20-479-EN-C
PDF: ISBN 978-92-76-20008-6 doi:10.2759/002360 KK-02-20-479-EN-N

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information. The contents of this publication are the sole responsibility of the High-Level Expert Group on Artificial Intelligence (AI HLEG). Although Commission staff facilitated the preparation thereof, the views expressed in this document reflect the opinion of the AI HLEG only and may not in any circumstances be regarded as reflecting an official position of the European Commission.

More information on the High-Level Expert Group on Artificial Intelligence is available online[1].

The reuse policy of European Commission documents is regulated by Decision 2011/833/EU (OJ L 330, 14.12.2011, p.39). For any use or reproduction of photos or other material that is not under the EU copyright, permission must be sought directly from the copyright holders.

[1] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

Introduction

In 2019 the High-Level Expert Group on Artificial Intelligence (AI HLEG),[2] set up by the European Commission, published the Ethics Guidelines for Trustworthy Artificial Intelligence.[3] The third chapter of those Guidelines contained an Assessment List to help assess whether the AI system that is being developed, deployed, procured or used, adheres to the seven requirements of Trustworthy Artificial Intelligence (AI), as specified in our Ethics Guidelines for Trustworthy AI:

1. Human Agency and Oversight;
2. Technical Robustness and Safety;
3. Privacy and Data Governance;
4. Transparency;
5. Diversity, Non-discrimination and Fairness;
6. Societal and Environmental Well-being;
7. Accountability.

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. It builds on the one outlined in the Ethics Guidelines for Trustworthy AI and was developed over a period of two years, from June 2018 to June 2020. In that period this Assessment List for Trustworthy AI (ALTAI) also benefited from a piloting phase (second half of 2019).[4] Through that piloting phase, the AI HLEG received valuable feedback through fifty in-depth interviews with selected companies; input through an open work stream on the AI Alliance[5] to provide best practices; and, via two publicly accessible questionnaires for technical and non-technical stakeholders.[6]

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,[7] the Charter of Fundamental Rights (the Charter)[8], and international human rights Law.[9] Please consult the text box below on fundamental rights to familiarise yourself with the concept and with the content of a Fundamental Rights Impact Assessment.

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. 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. It is intended to help organisations identify how proposed AI systems might

[2] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.
[3] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[4] https://ec.europa.eu/futurium/en/ethics-guidelines-trustworthy-ai/register-piloting-process-0.
[5] https://ec.europa.eu/futurium/en/eu-ai-alliance/BestPractices.
[6] https://ec.europa.eu/futurium/register-piloting-process.
[7] https://europa.eu/european-union/law/treaties_en
[8] https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en
[9] https://www.un.org/en/sections/universal-declaration/foundation-international-human-rights-law/index.html.

generate risks, and to identify whether and what kind of active measures may need to be taken to avoid and minimise those risks. 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. 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). It encourages the involvement of all relevant stakeholders. 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. This could be achieved through internal guidelines, governance processes, etc.

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.[10] 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. 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.

This document is the offline version of this Assessment List for Trustworthy AI (ALTAI). An online interactive version of this Assessment List for Trustworthy AI (ALTAI) is available.[11]

How to use this Assessment List for Trustworthy AI (ALTAI)

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 7 requirements and related questions. Among the stakeholders you may find for example the following:

- AI designers and AI developers of the AI system;
- data scientists;
- procurement officers or specialists;
- front-end staff that will use or work with the AI system;
- legal/compliance officers;
- management.

If you do not know how to address a question and find no useful help on the AI Alliance page,[12] it is advised to seek outside counsel or assistance. For each requirement, this Assessment List for Trustworthy AI (ALTAI) provides introductory guidance and relevant definitions in the Glossary. The online version of this Assessment List for Trustworthy AI (ALTAI) contains additional explanatory notes for many of the questions.[13]

[10] The three components of Trustworthy AI, as defined in the Ethics Guidelines for Trustworthy AI.
[11] https://ec.europa.eu/newsroom/dae/item.cfm?item_id=682761
[12] https://ec.europa.eu/newsroom/dae/item.cfm?item_id=682761
[13] https://ec.europa.eu/newsroom/dae/item.cfm?item_id=682761

Fundamental Rights

Fundamental rights encompass rights such as human dignity and non-discrimination, as well as rights in relation to data protection and privacy, to name just some examples. Prior to self-assessing an AI system with this Assessment List, a fundamental rights impact assessment (FRIA) should be performed.

A FRIA could include questions such as the following – drawing on specific articles in the Charter and the European Convention on Human Rights (ECHR)[14] its protocols and the European Social Charter.[15]

1. Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?

Have you put in place processes to test and monitor for potential negative discrimination (bias) during the development, deployment and use phases of the AI system?

Have you put in place processes to address and rectify for potential negative discrimination (bias) in the AI system?

2. Does the AI system respect the rights of the child, for example with respect to child protection
and taking the child's best interests into account?

Have you put in place processes to address and rectify for potential harm to children by the AI system?

Have you put in place processes to test and monitor for potential harm to children during the development, deployment and use phases of the AI system?

[14] https://www.echr.coe.int/Documents/Convention_ENG.pdf.
[15] https://www.coe.int/en/web/european-social-charter.

3. Does the AI system protect personal data relating to individuals in line with GDPR?[16]

Have you put in place processes to assess in detail the need for a data protection impact assessment, including an assessment of the necessity and proportionality of the processing operations in relation to their purpose, with respect to the development, deployment and use phases of the AI system?

Have you put in place measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the protection of personal data with respect to the development, deployment and use phases of the AI system?

See the section on Privacy and Data Governance in this Assessment List, and available guidance from the European Data Protection Supervisor.[17]

4. Does the AI system respect the freedom of expression and information and/or freedom of assembly and association?

Have you put in place processes to test and monitor for potential infringement on freedom of expression and information, and/or freedom of assembly and association, during the development, deployment and use phases of the AI system?

Have you put in place processes to address and rectify for potential infringement on freedom of expression and information, and/or freedom of assembly and association, in the AI system?

[16] https://gdpr.eu.

[17] https://edps.europa.eu/data-protection/notre-rôle-en-tant-que-contrôleur/data-protection-impact-assessment-dpia_en, and https://edps.europa.eu/sites/edp/files/publication/19-07-17_accountability_on_the_ground_part_ii_en.pdf.

REQUIREMENT #1 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-in-the-Loop; Human-on-the-Loop; Self-learning AI System; Subject; User.

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.

- Is the AI system designed to interact, guide or take decisions by human end-users that affect humans[18] or society?
    - 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?
    - Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision?
- 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?
    - Are end-users or subjects informed that they are interacting with an AI system?
- Could the AI system affect human autonomy by generating over-reliance by end-users?
    - Did you put in place procedures to avoid that end-users over-rely on the AI system?
- Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way?
    - Did you put in place any procedure to avoid that the AI system inadvertently affects human autonomy?
- Does the AI system simulate social interaction with or between end-users or subjects?

[18] Henceforward referred to as 'subjects'. The definition of subjects is available in the glossary.

- 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:
    - 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?
    - Did you take measures to minimise the risk of addiction?
    - Did you take measures to mitigate the risk of manipulation?

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.

- Please determine whether the AI system (choose as many as appropriate):
    - Is a self-learning or autonomous system;
    - Is overseen by a *Human*-in-the-Loop;
    - Is overseen by a *Human*-on-the-Loop;
    - Is overseen by a *Human*-in-Command.
- Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?
- Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?
- Did you ensure a 'stop button' or procedure to safely abort an operation when needed?
- Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?

REQUIREMENT #2 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; Red-team.

Resilience to Attack and Security

- 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?
- Is the AI system certified for cybersecurity (e.g. the certification scheme created by the Cybersecurity Act in Europe)[19] or is it compliant with specific security standards?
- How exposed is the AI system to cyber-attacks?
    - Did you assess potential forms of attacks to which the AI system could be vulnerable?
    - Did you consider different types of vulnerabilities and potential entry points for attacks such as:
        - Data poisoning (i.e. manipulation of training data);
        - Model evasion (i.e. classifying the data according to the attacker's will);
        - Model inversion (i.e. infer the model parameters)
- Did you put measures in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle?
- Did you red-team/pentest the system?
- Did you inform end-users of the duration of security coverage and updates?
    - What length is the expected timeframe within which you provide security updates for the AI system?

[19] https://ec.europa.eu/digital-single-market/en/eu-cybersecurity-act.

General Safety

- Did you define risks, risk metrics and risk levels of the AI system in each specific use case?
    - Did you put in place a process to continuously measure and assess risks?
    - Did you inform end-users and subjects of existing or potential risks?
- Did you identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible consequences?
    - Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system?
    - Did you define safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system?
- Did you assess the dependency of a critical AI system's decisions on its stable and reliable behaviour?
    - Did you align the reliability/testing requirements to the appropriate levels of stability and reliability?
- Did you plan fault tolerance via, e.g. a duplicated system or another parallel system (AI-based or 'conventional')?
- 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?

Accuracy

- Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?
- 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?
- Did you put in place a series of steps to monitor, and document the AI system's accuracy?[20]
- 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?
- 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?[21]

[20] Accuracy is only one performance metric and it might not be the most appropriate depending on the application. Monitoring false positives, false negatives, F1 score can help to determine if accuracy is actually reflecting the system's performance.

[21] Confidence of system users will depend on how much their expectation of system performance fits with its actual performance. Communicating accuracy metrics is therefore key.

Reliability, Fall-back plans and Reproducibility

- Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?
    - Did you put in place a well-defined process to monitor if the AI system is meeting the intended goals?[22]
    - Did you test whether specific contexts or conditions need to be taken into account
      to ensure reproducibility?
- 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?
    - Did you clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system?
- Did you define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them?
- Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score?
- Is your AI system using (online) continual learning?
    - Did you consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function?

[22] Performance metrics are an abstraction of the actual system behavior. Monitoring domain or application specific parameters by a supervisory mechanism is a way to verify that the system operates as intended.

REQUIREMENT #3 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.

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.

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

Data Governance

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

- Is your AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)?
- 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)[23];
    - Designate a Data Protection Officer (DPO)[24] 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);

[23] https://gdpr.eu/data-protection-impact-assessment-template/.
[24] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-officers/.

        - Data minimisation, in particular personal data (including special categories of data);
    - 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?
    - 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?
- Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data?
- Did you align the AI system with relevant standards (e.g. ISO[25], IEEE[26]) or widely adopted protocols for (daily) data management and governance?

[25] https://www.iso.org/committee/6794475.html.
[26] https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

REQUIREMENT #4 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.

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.

- Did you put in place measures that address the traceability of the AI system during its entire lifecycle?
    - Did you put in place measures to continuously assess the quality of the input data to the AI system?[27]
    - Can you trace back which data was used by the AI system to make a certain decision(s) or recommendation(s)?
    - Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system?
    - Did you put in place measures to continuously assess the quality of the output(s) of the AI system?[28]
    - Did you put adequate logging practices in place to record the decision(s) or recommendation(s) of the AI system?

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

[27] This could take the form of a standard automated quality assessment of data input: quantifying missing values, gaps in the data; exploring breaks in the data supply; detecting when data is insufficient for a task; detecting when the input data is erroneous, incorrect, inaccurate or mismatched in format – e.g. sensor is not working properly or health records are not recorded properly. A concrete example is sensor calibration: the process which aims to check and ultimately improve sensor performance by removing missing or otherwise inaccurate values (called structural errors) in sensor outputs.

[28] This could take the form of a standard automated quality assessment of AI output: e.g. predictions scores are within expected ranges; anomaly detection in output and reassign input data leading to the anomaly detected.

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.

- Did you explain the decision(s) of the AI system to the users?[29]
- Do you continuously survey the users if they understand the decision(s) of the AI system?

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.

- 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?
- Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?
    - Did you communicate the benefits of the AI system to users?
    - 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?
    - Did you provide appropriate training material and disclaimers to users on how to adequately use the AI system?

29 This depends on the organisation, if developers are involved directly with user interactions through workshops etc they could be addressed by this question, if they are not directly involved the organisation needs to make sure users understand the AI system and highlight any misunderstandings to the developing team.

REQUIREMENT #5 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 non-transparent 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.

Avoidance of Unfair Bias

- 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?
- Did you consider diversity and representativeness of end-users and/or subjects in the data?
    - Did you test for specific target groups or problematic use cases?
    - Did you research and use publicly available technical tools, that are state-of-the-art, to improve your understanding of the data, model and performance?
    - 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)?
    - Where relevant, did you consider diversity and representativeness of end-users and or subjects in the data?
- 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?
- Did you ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system?
    - Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?
    - 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?

- Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI system?
    - Did you consider other definitions of fairness before choosing this one?
    - Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities?
    - Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?
    - Did you establish mechanisms to ensure fairness in your AI system?

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[30] addressing the widest possible range of users, following relevant accessibility standards.[31] 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.

- Did you ensure that the AI system corresponds to the variety of preferences and abilities in society?
- 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?
    - 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)?
    - 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?
- Did you ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable?
- Did you take the impact of the AI system on the potential end-users and/or subjects into account?
    - Did you assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects?
    - Did you assess whether there could be groups who might be disproportionately affected by the outcomes of the AI system?
    - Did you assess the risk of the possible unfairness of the system onto the end-user's or subject's communities?

[30] https://www.cen.eu/news/brief-news/Pages/NEWS-2019-014.aspx.
[31] https://www.iso.org/standard/58625.html; https://www.iso.org/standard/33987.html; https://www.iso.org/obp/ui/#iso:std:iso:9241:-171:ed-1:v1:en; http://mandate376.standards.eu/standard.

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.

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

REQUIREMENT #6 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.[32] 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.

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.

- Are there potential negative impacts of the AI system on the environment?
    - Which potential impact(s) do you identify?
- 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)?
    - Did you define measures to reduce the environmental impact of the AI system throughout its lifecycle?

[32] https://sustainabledevelopment.un.org/?menu=1300.

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.

- Does the AI system impact human work and work arrangements?
- 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?
- Did you adopt measures to ensure that the impacts of the AI system on human work are well understood?
    - Did you ensure that workers understand how the AI system operates, which capabilities it has and which it does not have?
- Could the AI system create the risk of de-skilling of the workforce?
    - Did you take measures to counteract de-skilling risks?
- Does the system promote or require new (digital) skills?
    - Did you provide training opportunities and materials for re- and up-skilling?

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.).

- Could the AI system have a negative impact on society at large or democracy?
    - 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?
    - Did you take action to minimize potential societal harm of the AI system?
    - Did you take measures that ensure that the AI system does not negatively impact democracy?

REQUIREMENT #7 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.

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.

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

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.

- Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures?
    - Does the involvement of these third parties go beyond the development phase?
- Did you organise risk training and, if so, does this also inform about the potential legal framework applicable to the AI system?
- 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?
- 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)?
    - 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?
    - 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?
- 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?
    - Does this process foster revision of the risk management process?
- For applications that can adversely affect individuals, have redress by design mechanisms been put in place?

Glossary

This glossary is informed by the glossary that accompanied the Ethics Guidelines for Trustworthy AI.[33]

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.

AI bias: AI (or algorithmic) bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as favouring one arbitrary group of users over others. Bias can emerge due to many factors, including but not limited to the design of the algorithm or the unintended or unanticipated use or decisions relating to the way data is coded, collected, selected or used to train the algorithm. Bias can enter into algorithmic systems as a result of pre-existing cultural, social, or institutional expectations; because of technical limitations of their design; or by being used in unanticipated contexts or by audiences who are not considered in the software's initial design. AI bias is found across platforms, including but not limited to search engine results and social media platforms, and can have impacts ranging from inadvertent privacy violations to reinforcing social biases of race, gender, sexuality, and ethnicity.

AI designer: 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

[33] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

by designing platforms to support data collection and annotation, ensuring that data collection respects some properties (such as safety and fairness).

AI developer: 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.

AI Ethics Review Board: 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.[34]

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.

AI system: Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).

A separate document prepared by the AI HLEG and elaborating on the definition of AI used for the purpose of this document is titled "A definition of AI: Main capabilities and scientific disciplines".[35]

AI system environment: This denotes everything in the world which surrounds the AI system, but which is not a part of the system itself. More technically, an environment can be described as a situation in which the system operates. AI systems get information from their environment via sensors that collect data and modify the environment via suitable actuators.

34 More information can be found here: https://www.accenture.com/_acnmedia/PDF-107/Accenture-AI-And-Data-Ethics-Committee-Report-11.pdf#zoom=50.

35 https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

Depending on whether the environment is in the physical or virtual world, actuators can be hardware, such as robotic arms, or software, such as programs that make changes in some digital structure.

Assistive Technology: Software or hardware that is added to or incorporated within an ICT system to increase accessibility. Often it is specifically designed to assist people with disabilities in carrying out daily activities. Assistive technology includes wheelchairs, reading machines, devices for grasping, etc. In the area of Web Accessibility, common software-based assistive technologies include screen readers, screen magnifiers, speech synthesizers, and voice input software that operate in conjunction with graphical desktop browsers (among other user agents). Hardware assistive technologies include alternative keyboards and pointing devices.

Audit: An audit is an independent examination of some required properties of an entity, be it a company, a product, or a piece of software. Audits provide third-party assurance to various stakeholders that the subject matter is free from material misstatement. The term is most frequently applied to audits of the financial information relating to a legal person, but can be applied to anything else.

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.

Autonomous AI systems: An autonomous AI system is an AI system that performs behaviors or tasks with a high degree of autonomy, that is, without external influence.

Confidence score: Much of AI involves estimating some quantity, such as the probability that the output is a correct answer to the given input. Confidence scores, or confidence intervals, are a way of quantifying the uncertainty of such an estimate. A low confidence score associated with the output of an AI system means that the system is not too sure that the specific output is correct.

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.

Data poisoning: Data poisoning occurs when an adversarial actor attacks an AI system, and is able to inject bad data into the AI model's training set, thus making the AI system learn something that it should not learn. Examples show that in some cases these data poisoning attacks on neural nets can be very effective, causing a significant drop in accuracy even with very little data poisoning. Other kinds of poisoning attacks do not aim to change the behavior of the AI system, but rather they insert a backdoor, which is a data that the model's designer is not aware of, but that the attacker can leverage to get the AI system to do what they want.

Data Protection Impact Assessment (DPIA): Evaluation of the effects that the processing of personal data might have on individuals to whom the data relates. A DPIA is necessary in all cases in which the technology creates a high risk of violation of the rights and freedoms of individuals. The law requires a DPIA in case of automated processing, including profiling (i), processing of personal data revealing sensitive information like racial of ethnic origin, political opinions, religious or philosophical beliefs (ii), processing of personal data relating to criminal convictions and offences (iii) and systematic monitoring of a publicly accessible area on a large scale (iv).

Data Protection Officer (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.

Encryption, Pseudonymisation, Aggregation, and Anonymisation: Pseudonymisation refers to the idea that it is not possible to attribute personal data to a specific data subject without additional information. By contrast to pseudonymisation, anonymisation consists in preventing any identification of individuals from personal data. The link between an individual and personal data is definitively erased. Encryption is the procedure whereby clear text information is disguised by using especially a hash key. Encrypted results are unintelligible data for persons who do not have the encryption key. Aggregation is a process whereby data is gathered and expressed in a summary form, especially for statistical analysis.

End-user: 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.

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 safety-critical 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.

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. 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 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.

Lifecycle: The lifecycle of an AI system includes several interdependent phases ranging from its design and development (including sub-phases such as requirement analysis, data collection, training, testing, integration), installation, deployment, operation, maintenance, and disposal. Given the complexity of AI (and in general information) systems, several models and methodologies have been defined to manage this complexity, especially during the design and development phases, such as waterfall, spiral, agile software development, rapid prototyping, and incremental.

Model Evasion: Evasion is one of the most common attacks on machine learning models (ML) performed during production. It refers to designing an input, which seems normal for a human but is wrongly classified by ML models. A typical example is to change some pixels in a picture before uploading, so that the image recognition system fails to classify the result.

Model Inversion: Model inversion refers to a kind of attack to AI models, in which the access to a model is abused to infer information about the training data. So, model inversion turns the usual path from training data into a machine-learned model from a one-way one to a two-way one, permitting the training data to be estimated from the model with varying degrees of accuracy. Such attacks raise serious concerns given that training data usually contain privacy-sensitive information.

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,[36] 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.

Pen test: A penetration test, colloquially known as a pen test, pentest or ethical hacking, is an authorized simulated cyberattack on a computer system, performed to evaluate the security of the system. The test is performed to identify both weaknesses (also referred to as vulnerabilities), including the potential for unauthorised parties to gain access to the system's features and data, as well as strengths, enabling a full risk assessment to be completed.

Red team: 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.

Redress by design: Redress by design relates to the idea of establishing, from the design phase, mechanisms to ensure redundancy, alternative systems, alternative procedures, etc. in order to be able to effectively detect, audit, rectify the wrong decisions taken by a perfectly functioning system and, if possible, improve the system.

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.

[36] https://www.sciencedirect.com/topics/neuroscience/neural-networks.

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.

Self-learning AI system: Self-learning (or self-supervised learning) AI systems recognize patterns in the training data in an autonomous way, without the need for supervision.

Standards: Standards are norms designed by industry and/or Governments that set product or services' specifications. They are a key part of our society as they ensure quality and safety in both products and services in international trade. Businesses can be seen to benefit from standards as they can help cut costs by improved systems and procedures put in place. Standards are internationally agreed by experts and they usually represent what the experts think is the best way of doing something. It could be about making a product, managing a process, delivering a service or supplying materials – standards cover a huge range of activities. Standards are released by international organizations, such as ISO (International Organisation for Standardisation), IEEE (The Institute of Electrical and Electronics Engineers) Standard Association, and NIST (National Institute of Standards and Technology).

Subject: A subject is a person or a group of persons affected by the AI system (such as the recipient of benefits where the decision to grant or reject benefits is underpinned by an AI-system, or the general public for facial recognition).

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

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.[37] 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.

Universal Design: Terms such as "Design for All", "Universal Design", "accessible design", "barrier-free design", "inclusive design" and "transgenerational design" are often used interchangeably with the same meaning. These concepts have been developed by different stakeholders working to deliver high levels of accessibility. A parallel development of human-centred design emerged within ergonomics focusing on usability. These related concepts are expressed in the human rights perspective of the Design for All approach. The Design for All approach focuses on user involvement and experiences during the design and development process to achieve accessibility and usability. It should be applied from the

[37] This is informed by the Ethics Guidelines for Trustworthy AI, accessible here: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

earliest possible time, and throughout all stages in the life of products and services which are intended for mainstream use. A Design for All approach also focuses on user requirements and interoperability between products and services across the end-to-end chain of use to reach inclusive and non-stigmatizing solutions.

Use case: A use case is a specific situation in which a product or service could potentially be used. For example, self-driving cars or care robots are use cases for AI.

User: A user is a person that uses, supports or maintains the product, such as system administrators, database administrators, information technology experts, software professionals and computer technicians.

Workflow of the model: The workflow of an AI model shows the phases needed to build the model and their interdependencies. Typical phases are: Data collection and preparation, Model development, Model training, Model accuracy evaluation, Hyperparameters' tuning, Model usage, Model maintenance, Model versioning. These stages are usually iterative: one may need to reevaluate and go back to a previous step at any point in the process.

This Document was prepared by the High-Level Expert Group on AI the members of which are as follow:

Pekka Ala-Pietilä, Chair of the AI HLEG
Huhtamaki, Sanoma

Wilhelm Bauer
Fraunhofer

Urs Bergmann
Zalando

Mária Bieliková
Slovak University of Technology in Bratislava

Nozha Boujemaa
INRIA (06/18-02/19)

Cecilia Bonefeld-Dahl
DIGITALEUROPE

Yann Bonnet
ANSSI

Loubna Bouarfa
OKRA

Stéphan Brunessaux
Airbus

Raja Chatila
IEEE Initiative Ethics of Intelligent/Autonomous Systems & Sorbonne University

Mark Coeckelbergh
University of Vienna

Virginia Dignum
Umea University

Luciano Floridi
University of Oxford

Jean-François Gagné
Element AI

Chiara Giovannini
ANEC

Joanna Goodey
Fundamental Rights Agency

Sami Haddadin
MSRM, TUM

Gry Hasselbalch
The thinkdotank DataEthics & University of Copenhagen

Fredrik Heintz
Linköping University

Fanny Hidvegi
Access Now

Eric Hilgendorf
University of Würzburg

Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen

Mari-Noëlle Jégo-Laveissière
Orange

Leo Kärkkäinen
Nokia Bell Labs

Sabine Theresia Köszegi
TU Wien

Robert Kroplewski
Solicitor & Advisor to Polish Government

Elisabeth Ling
RELX

Pierre Lucas
Orgalim, Europe's technology Industries

Raoul Mallart
Sigfox (06/18-12/18)

Ieva Martinkenaite
Telenor

Thomas Metzinger
JGU Mainz & European University Association

Catelijne Muller
ALLAI Netherlands & EESC

Markus Noga
SAP

Barry O'Sullivan, Vice-Chair of the AI HLEG,
University College Cork

Ursula Pachl
BEUC

Lorena Jaume Palasi
AlgorithmWatch (06/18-10/18)

Nicolas Petit
University of Liège

Christoph Peylo
Bosch

Iris Plöger
BDI

Stefano Quintarelli
Garden Ventures

Andrea Renda
College of Europe Faculty & CEPS

Francesca Rossi
IBM

Cristina San José
European Banking Federation

Isabelle Schömann
ETUC (06/20-06/20)

George Sharkov
Digital SME Alliance

Philipp Slusallek
German Research Centre for AI (DFKI)

Françoise Soulié Fogelman
AI Consultant

Saskia Steinacker
Bayer

Reinhard Stolle
BMW (06/18-09/18)

Jaan Tallinn
Ambient Sound Investment

Thierry Tingaud
STMicroelectronics

Jakob Uszkoreit
Google

Thiébaut Weber
ETUC (06/18-08/19)

Aimee Van Wynsberghe
TU Delft

Cecile Wendling
AXA

Karen Yeung
The University of Birmingham

All AI HLEG members contributed to this deliverable. Catelijne Muller and Andrea Renda acted as co-leads. The following AI HLEG members contributed to the in-depth revision of specific key requirements (in alphabetical order):

Human Agency and Oversight
Joanna Goodey, Fredrik Heintz

Technical Robustness and Safety
Yann Bonnet, Raja Chatila, Pierre Lucas, Andrea Renda, George Sharkov, Jaan Tallinn, Cecile Wendling

Privacy and Data Governance
Cecilia Bonefeld-Dahl, Fanny Hidvegi

Transparency
Mária Bieliková, Ieva Martinkenaite, Ursula Pachl

Diversity, Non-discrimination and Fairness
Klaus Höckner, Francesca Rossi

Societal and Environmental Well-being
Mark Coeckelbergh, Virginia Dignum, Sabine Theresia Köszegi

Accountability
Robert Kroplewski, Christoph Peylo, Stefano Quintarelli

Glossary
Nicolas Petit, Francesca Rossi

The revisions were directly informed by the piloting process[38] of the Assessment List conducted in the second half of 2019. Feedback was received through a questionnaire accessible on the AI Alliance for technical and non-technical stakeholders, an open submission stream on the AI Alliance for written feedback and best practices, and fifty in-depth interviews with selected organisations from across the European Union.

Pekka Ala-Pietilä is the Chair of the AI HLEG. Barry O'Sullivan is the Vice-Chair. Both contributed to the work on this document.

Charlotte Stix coordinates the AI HLEG and provided editorial support.

38 More information about the process is available here: https://ec.europa.eu/futurium/en/ethics-guidelines-trustworthy-ai/pilot-assessment-list-ethics-guidelines-trustworthy-ai.

Book: ISBN 978-92-76-20009-3 doi:10.2759/791819 KK-02-20-479-EN-C
PDF: ISBN 978-92-76-20008-6 doi:10.2759/002360 KK-02-20-479-EN-N