Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

Numbered questions based on the format <KeyRequirement>.<Question>.<SubQuestion>

ID	Question
REQUIREMENT #1 Human Agency and Oversight	
Human Agency and Autonomy	
1.1	Is the AI system designed to interact, guide or take decisions by human end-users that affect humans or society?
1.1.1	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?
1.1.2	Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision?
1.2	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?
1.2.1	Are end-users or subjects informed that they are interacting with an AI system?
1.3	Could the AI system affect human autonomy by generating over-reliance by end-users?
1.3.1	Did you put in place procedures to avoid that end-users over-rely on the AI system?
1.4	Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way?
1.4.1	Did you put in place any procedure to avoid that the AI system inadvertently affects human autonomy?
1.5	Does the AI system simulate social interaction with or between end-users or subjects?
1.6	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:
1.6.1	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?
1.6.2	Did you take measures to minimise the risk of addiction?
1.6.3	Did you take measures to mitigate the risk of manipulation?
Human Oversight	
1.7	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.
1.8	Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?
1.9	Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?
1.10	Did you ensure a 'stop button' or procedure to safely abort an operation when needed?
1.11	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	
Resilience to Attack and Security	
2.1	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?
2.2	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?
2.3	How exposed is the AI system to cyber-attacks?
2.3.1	Did you assess potential forms of attacks to which the AI system could be vulnerable?
2.3.2	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)
2.4	Did you put measures in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle?
2.5	Did you red-team/pentest the system?
2.6	Did you inform end-users of the duration of security coverage and updates?
2.6.1	What length is the expected timeframe within which you provide security updates for the AI system?
General Safety	
2.7	Did you define risks, risk metrics and risk levels of the AI system in each specific use case?
2.7.1	Did you put in place a process to continuously measure and assess risks?
2.7.2	Did you inform end-users and subjects of existing or potential risks?
2.8	Did you identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible consequences?
2.8.1	Did you assess the risk of possible malicious use, misuse or inappropriate use of the AI system?
2.8.2	Did you define safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system?
2.9	Did you assess the dependency of a critical AI system's decisions on its stable and reliable behaviour?
2.9.1	Did you align the reliability/testing requirements to the appropriate levels of stability and reliability?
2.10	Did you plan fault tolerance via, e.g. a duplicated system or another parallel system (AI-based or 'conventional')?
2.11	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	
2.12	Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?
2.13	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?
2.14	Did you put in place a series of steps to monitor, and document the AI system's accuracy?
2.15	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?
2.16	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?
Reliability, Fall-back plans and Reproducibility	
2.17	Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?

2.17.1	Did you put in place a well-defined process to monitor if the AI system is meeting the intended goals?
2.17.2	Did you test whether specific contexts or conditions need to be taken into account to ensure reproducibility?
2.18	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?
2.18.1	Did you clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system?
2.19	Did you define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them?
2.20	Did you put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score?
2.21	Is your AI system using (online) continual learning?
2.21.1	Did you consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function?

REQUIREMENT #3 Privacy and Data Governance	
Privacy	
3.1	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?
3.2	Depending on the use case, did you establish mechanisms that allow flagging issues related to privacy concerning the AI system?
Data Governance	
3.3	Is your AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)?
3.4	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);
3.4.1	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?
3.4.2	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?
3.5	Did you consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data?
3.6	Did you align the AI system with relevant standards (e.g. ISO, IEEE) or widely adopted protocols for (daily) data management and governance?

REQUIREMENT #4 Transparency	
Traceability	
4.1	Did you put in place measures that address the traceability of the AI system during its entire lifecycle?
4.1.1	Did you put in place measures to continuously assess the quality of the input data to the AI system?
4.1.2	Can you trace back which data was used by the AI system to make a certain decision(s) or recommendation(s)?
4.1.3	Can you trace back which AI model or rules led to the decision(s) or recommendation(s) of the AI system?
4.1.4	Did you put in place measures to continuously assess the quality of the output(s) of the AI system?
4.1.5	Did you put adequate logging practices in place to record the decision(s) or recommendation(s) of the AI system?
Explainability	
4.2	Did you explain the decision(s) of the AI system to the users?
4.3	Do you continuously survey the users if they understand the decision(s) of the AI system?
Communication	
4.4	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?
4.5	Did you establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?
4.5.1	Did you communicate the benefits of the AI system to users?
4.5.2	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?
4.5.3	Did you provide appropriate training material and disclaimers to users on how to adequately use the AI system?

REQUIREMENT #5 Diversity, Non-discrimination and Fairness	
Avoidance of unfair Bias	
5.1	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?
5.2	Did you consider diversity and representativeness of end-users and/or subjects in the data?
5.2.1	Did you test for specific target groups or problematic use cases?
5.2.2	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?
5.2.3	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)?
5.2.4	Where relevant, did you consider diversity and representativeness of end-users and or subjects in the data?
5.3	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?
5.4	Did you ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system?
5.4.1	Did you establish clear steps and ways of communicating on how and to whom such issues can be raised?
5.4.2	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?
5.5	Is your definition of fairness commonly used and implemented in any phase of the process of setting up the AI system?
5.5.1	Did you consider other definitions of fairness before choosing this one?
5.5.2	Did you consult with the impacted communities about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities?
5.5.3	Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness?
5.5.4	Did you establish mechanisms to ensure fairness in your AI system?
Accessibility and Universal Design	
5.6	Did you ensure that the AI system corresponds to the variety of preferences and abilities in society?
5.7	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?
5.7.1	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)?
5.7.2	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?
5.8	Did you ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable?
5.9	Did you take the impact of the AI system on the potential end-users and/or subjects into account?
5.9.1	Did you assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects?
5.9.2	Did you assess whether there could be groups who might be disproportionately affected by the outcomes of the AI system?
5.9.3	Did you assess the risk of the possible unfairness of the system onto the end-user's or subject's communities?
Stakeholder Participation	
5.10	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	
Environmental Well-being	
6.1	Are there potential negative impacts of the AI system on the environment?
6.1.1	Which potential impact(s) do you identify?
6.2	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)?
6.2.1	Did you define measures to reduce the environmental impact of the AI system throughout its lifecycle?
Impact on Work and Skills	
6.3	Does the AI system impact human work and work arrangements?
6.4	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?
6.5	Did you adopt measures to ensure that the impacts of the AI system on human work are well understood?
6.5.1	Did you ensure that workers understand how the AI system operates, which capabilities it has and which it does not have?
6.6	Could the AI system create the risk of de-skilling of the workforce?
6.6.1	Did you take measures to counteract de-skilling risks?
6.7	Does the system promote or require new (digital) skills?
6.7.1	Did you provide training opportunities and materials for re- and up-skilling?
Impact on Society at large or Democracy	
6.8	Could the AI system have a negative impact on society at large or democracy?
6.8.1	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?
6.8.2	Did you take action to minimize potential societal harm of the AI system?
6.8.3	Did you take measures that ensure that the AI system does not negatively impact democracy?

REQUIREMENT #7 Accountability	
Auditability	
7.1	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)?
7.2	Did you ensure that the AI system can be audited by independent third parties?
Risk Management	
7.3	Did you foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures?
7.3.1	Does the involvement of these third parties go beyond the development phase?
7.4	Did you organise risk training and, if so, does this also inform about the potential legal framework applicable to the AI system?
7.5	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?
7.6	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)?
7.6.1	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?
7.6.2	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?
7.7	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?
7.7.1	Does this process foster revision of the risk management process?
7.8	For applications that can adversely affect individuals, have redress by design mechanisms been put in place?