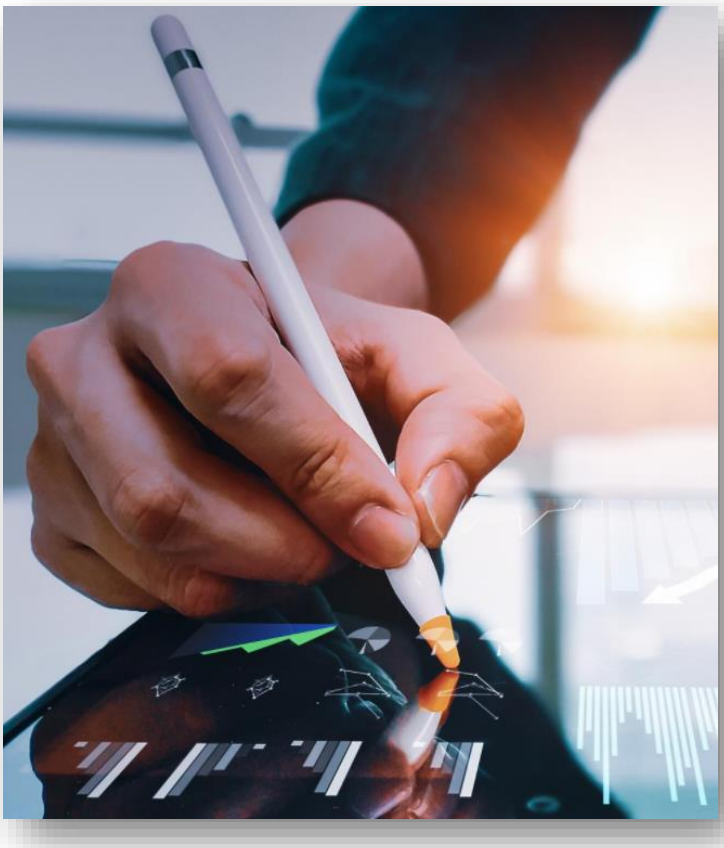
AI & Partners

Assessment List for Trustworthy AI under the EU AI Act

The assessment tool for AI systems' compliance with Trustworthy AI requirements



Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

Disclaimer

The contents of this document are for information purposes only. AI & Partners assumes no liability or responsibility for any inaccurate or incomplete information, nor for any actions taken in reliance thereon. The published material is distributed without warranty of any kind, either express or implied, and the responsibility for the interpretation and use of the material lies with the reader. In no event shall, AI & Partners be liable for damages arising from its use.

AI & Partners takes no responsibility for the content of any external website referenced in this publication or for any defamatory, offensive or misleading information which might be contained on these third-party websites. Any links to external websites do not constitute an endorsement by AI & Partners, and are only provided as a convenience. It is the responsibility of the reader to evaluate the content and usefulness of information obtained from other sites.

The views, thoughts and opinions expressed in the content of this publication belong solely to the authors and do not necessarily reflect the views or policies of AI & Partners, their partners, their consultants, nor does it imply any endorsement. Therefore, AI & Partners carries no responsibility for the opinions expressed in this publication.

AI & Partners does not endorse or recommend any product, process, or service. Therefore, mention of any products, processes, or services in this document cannot be construed as an endorsement or recommendation by AI & Partners.

The contents of this document may be quoted or reproduced, provided that the source of information is acknowledged. AI & Partners would like to receive a copy of the document in which this publication is used or quoted.

EU AI Act – Advisory (Scoping) | Consultancy | Compliance Software
+31 6 57285579 and +44(0)75 35994 132
s.musch@ai-and-partners.com and m.borrelli@ai-and-partners.com
https://www.ai-and-partners.com/
https://twitter.com/AI_and_Partners @AI_and_Partners
https://www.linkedin.com/company/ai-&-partners/

Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

Overview

What

This document contains a redacted version of a recommended Assessment List for Trustworthy AI ("ALTAI"). ALTAI was developed with reference to accepted standards under the proposed European Union Artificial Intelligence Act (the "EU AI Act") , set up by the European Commission ("EC") to help assess whether the AI system that is being developed, deployed, procured or used, complies with the EC's proposed six general principles applicable to all AI systems. ALTAI establishes a high-level framework that promotes a coherent human-centric European approach to ethical and trustworthy Artificial Intelligence, which is fully in line with the EU AI Act:

- Human Agency and Oversight.
- Technical Robustness and Safety.
- Privacy and Data Governance.
- Transparency.
- Diversity, Non-discrimination and Fairness.
- Societal and Environmental Well-being.

Who

This document is relevant to any individual or team in an enterprise who is interested in learning more about how to conduct an assessment of AI systems and wants to understand the recommended approach. This document is of interest to:

- AI designers and AI-developers of the AI system.
- Front-end staff that will use or work with the AI system.
- Legal/compliance officers.

Why

This document provides a basis evaluation process for Trustworthy AI self-evaluation. Businesses can draw elements relevant to the particular AI system or add elements to it as they see fit, taking into consideration the sector they operate in.

Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

Table of Contents

Disclaimer 1
Overview 2
What 2
Who 2
Why 2
Table of Contents 3
Principles 5
Human Agency and Oversight 5
Human Autonomy 5
Human Oversight 6
Technical Robustness and Safety 6
General Safety 7
Accuracy 7
Reliability, Fall-Back Plans and Reproducibility 7
Privacy and Data Governance 7
Transparency 8
Traceability 8
Explainability 9
Communication 9
Diversity, Non-Discrimination and Fairness 10
Avoidance of Unfair Bias 10
Accessibility and Universal Design 10
Stakeholder Participation 11

EU AI Act – Advisory (Scoping) | Consultancy | Compliance Software
+31 6 57285579 and +44(0)75 35994 132
s.musch@ai-and-partners.com and m.borrelli@ai-and-partners.com
https://www.ai-and-partners.com/
https://twitter.com/AI_and_Partners @AI_and_Partners
https://www.linkedin.com/company/ai-&-partners/

AI & Partners

Societal and Environmental Well-Being 11

Environmental Well-Being 12

Impact on Work and Skills 12

Example 13

Result 13

Recommendation 13

Fundamental Rights 14

Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

Principles

Human Agency and Oversight

AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and upholding fundamental rights, which should be underpinned by human oversight. In this section, we are asking you to assess the AI system in terms of the respect for human agency, as well as human oversight.

Human 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 ('subjects') or society?

- Yes ☐
- To some extent ☐
- No ☐
- Don't know ☐

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

- Yes ☐
- No ☐

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

- Yes ☐
- No ☐

Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

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 describe whether the AI system (tick 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☐
- Other☐
- Don't know☐

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.

Is the AI system certified for cybersecurity or is it compliant with security standards?

- Yes ☐
- No ☐
- Don't know ☐

Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

General Safety

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?

- Yes ☐
- No ☐
- Don't know ☐

Accuracy

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

- Yes ☐
- No ☐
- Don't know ☐

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?

- Yes ☐
- No ☐
- Don't know ☐

Is your AI system using online continual learning?

- Yes ☐
- No ☐
- Don't know ☐

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.

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?

- Yes ☐
- No ☐
- Don't know ☐

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. Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured.

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.

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 systemís decisions, is properly documented to allow for traceability, increase transparency and, ultimately, build trust in AI in society.

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

- Yes ☐
- To some extent☐
- No ☐
- Don't know ☐

Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

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 usersí trust in AI systems. AI driven decisions ñ to the extent possible ñ must be explained and understood to 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 ëblack boxes' and require special attention. In those circumstances, other explainability measures (e.g. traceability, auditability and transparent communication on the AI system 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 lives.

Did you explain the decision of the AI system to the users?

- Yes ☐
- No ☐
- Don't know ☐

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?

- Yes ☐
- To some extent☐
- No ☐
- Don't know ☐

Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

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.

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?

- Yes ☐
- No ☐

Accessibility and Universal Design

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

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

- Yes ☐
- No ☐

Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

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?

- Yes ☐
- No ☐

Societal and Environmental Well-Being

In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's life cycle. Ubiquitous exposure to social AI systems in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or negatively impact our social relationships and attachment.

While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could equally affect peoples' physical and mental well-being. The effects of AI systems must therefore be carefully monitored and considered.

Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, for instance the Sustainable Development Goals.

Overall, AI should be used to benefit all human beings, including future generations. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. AI systems must not undermine democratic processes, human deliberation or democratic voting systems or pose a systemic threat to society at large.

EU AI Act – Advisory (Scoping) | Consultancy | Compliance Software
+31 6 57285579 and +44(0)75 35994 132
s.musch@ai-and-partners.com and m.borrelli@ai-and-partners.com
https://www.ai-and-partners.com/
https://twitter.com/AI_and_Partners @AI_and_Partners
https://www.linkedin.com/company/ai-&-partners/

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?

- Yes ☐
- No ☐

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

- Yes ☐
- No ☐

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?

- Yes ☐
- To some extent☐
- No ☐
- Don't know☐

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

- Yes ☐
- No ☐

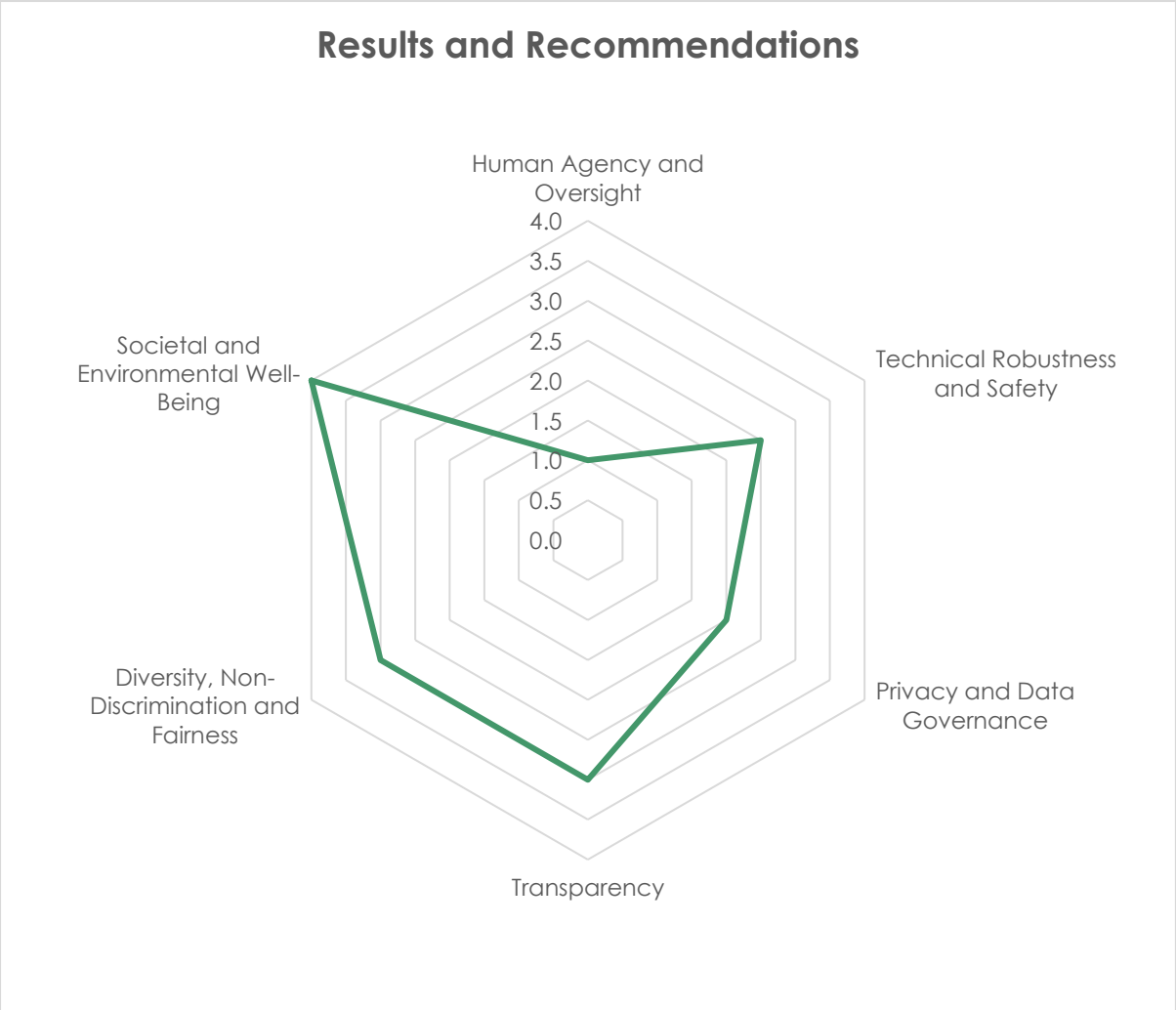
Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

Example

Result

Output of the assessment in a standardised report to suit management needs.



Recommendation

Human Agency and Oversight

Ensure that the end-users or subjects are adequately informed that they are interacting with an AI system.

EU AI Act – Advisory (Scoping) | Consultancy | Compliance Software
+31 6 57285579 and +44(0)75 35994 132
s.musch@ai-and-partners.com and m.borrelli@ai-and-partners.com
https://www.ai-and-partners.com/
https://twitter.com/AI_and_Partners @AI_and_Partners
https://www.linkedin.com/company/ai-&-partners/

Assessment List for Trustworthy AI under the EU AI Act

AI & Partners

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") its protocols and the European Social Charter.

Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, colour or ethnic?

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

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

Does the AI system protect the right to privacy, including personal data relating to individuals in line with GDPR?

- 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 to ensure the protection of personal data with respect to the development, deployment and use phases of the AI system?

Does the AI system respect the freedom of expression or assembly?

- Could the AI-system potentially limit a person's freedom to openly express an opinion, partake in a peaceful demonstration or join a union?