assessment list does not provide concrete answers to address the raised questions, it encourages reflection on how Trustworthy AI can be operationalised, and on the potential steps that should be taken in this regard.

- *Relation to existing law and processes*

It is also important for AI practitioners to recognise that there are various existing laws mandating particular processes or prohibiting particular outcomes, which may overlap and coincide with some of the measures listed in the assessment list. For example, data protection law sets out a series of legal requirements that must be met by those engaged in the collection and processing of personal data. Yet, because Trustworthy AI also requires the ethical handling of data, internal procedures and policies aimed at securing compliance with data protection laws might also help to facilitate ethical data handling and can hence complement existing legal processes. Compliance with this assessment list is *not*, however, evidence of legal compliance, nor is it intended as guidance to ensure compliance with applicable laws.

Moreover, many AI practitioners already have existing assessment tools and software development processes in place to ensure compliance also with non-legal standards. The below assessment should not necessarily be carried out as a stand-alone exercise, but can be incorporated into such existing practices.

TRUSTWORTHY AI ASSESSMENT LIST (PILOT VERSION)

1. Human agency and oversight

Fundamental rights:

- ✓ Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?
- ✓ Does the AI system interact with decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)?
  - Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?
  - Did you consider whether the AI system should communicate to (end) users that a decision, content, advice or outcome is the result of an algorithmic decision?
  - In case of a chat bot or other conversational system, are the human end users made aware that they are interacting with a non-human agent?

Human agency:

- ✓ Is the AI system implemented in work and labour process? If so, did you consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control?
  - Does the AI system enhance or augment human capabilities?
  - Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?

Human oversight:

- ✓ Did you consider the appropriate level of human control for the particular AI system and use case?
  - Can you describe the level of human control or involvement?
  - Who is the "human in control" and what are the moments or tools for human intervention?
  - Did you put in place mechanisms and measures to ensure human control or oversight?
  - Did you take any measures to enable audit and to remedy issues related to governing AI autonomy?
- ✓ Is there is a self-learning or autonomous AI system or use case? If so, did you put in place more specific mechanisms of control and oversight?
  - Which detection and response mechanisms did you establish to assess whether something could go wrong?

- Did you ensure a stop button or procedure to safely abort an operation where needed? Does this procedure abort the process entirely, in part, or delegate control to a human?

2. Technical robustness and safety

Resilience to attack and security:

- ✓ Did you assess potential forms of attacks to which the AI system could be vulnerable?
  - Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks?
- ✓ Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?
- ✓ Did you verify how your system behaves in unexpected situations and environments?
- ✓ Did you consider to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or deploying the system)?

Fallback plan and general safety:

- ✓ Did you ensure that your system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)?
- ✓ Did you consider the level of risk raised by the AI system in this specific use case?
  - Did you put any process in place to measure and assess risks and safety?
  - Did you provide the necessary information in case of a risk for human physical integrity?
  - Did you consider an insurance policy to deal with potential damage from the AI system?
  - Did you identify potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse? Is there a plan to mitigate or manage these risks?
- ✓ Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity?
  - Did you consider the liability and consumer protection rules, and take them into account?
  - Did you consider the potential impact or safety risk to the environment or to animals?
  - Did your risk analysis include whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behaviour of the AI system?
- ✓ Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?
  - Did you define thresholds and did you put governance procedures in place to trigger alternative/fallback plans?
  - Did you define and test fallback plans?

Accuracy

- ✓ Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?
  - Did you assess how accuracy is measured and assured?
  - Did you put in place measures to ensure that the data used is comprehensive and up to date?
  - Did you put in place measures in place to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias?
- ✓ Did you verify what harm would be caused if the AI system makes inaccurate predictions?
- ✓ Did you put in place ways to measure whether your system is making an unacceptable amount of inaccurate predictions?
- ✓ Did you put in place a series of steps to increase the system's accuracy?

Reliability and reproducibility:

- ✓ Did you put in place a strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications?
  - Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility?
  - Did you put in place verification methods to measure and ensure different aspects of the system's reliability and reproducibility?
  - Did you put in place processes to describe when an AI system fails in certain types of settings?
  - Did you clearly document and operationalise these processes for the testing and verification of the reliability of AI systems?
  - Did you establish mechanisms of communication to assure (end-)users of the system's reliability?

3. Privacy and data governance

Respect for privacy and data Protection:

- ✓ Depending on the use case, did you establish a mechanism allowing others to flag issues related to privacy or data protection in the AI system's processes of data collection (for training and operation) and data processing?
- ✓ Did you assess the type and scope of data in your data sets (for example whether they contain personal data)?
- ✓ Did you consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?
- ✓ Did you build in mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable)?
- ✓ Did you take measures to enhance privacy, such as via encryption, anonymisation and aggregation?
- ✓ Where a Data Privacy Officer (DPO) exists, did you involve this person at an early stage in the process?

Quality and integrity of data:

- ✓ Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?
- ✓ Did you establish oversight mechanisms for data collection, storage, processing and use?
- ✓ Did you assess the extent to which you are in control of the quality of the external data sources used?
- ✓ Did you put in place processes to ensure the quality and integrity of your data? Did you consider other processes? How are you verifying that your data sets have not been compromised or hacked?

Access to data:

- ✓ What protocols, processes and procedures did you follow to manage and ensure proper data governance?
  - Did you assess who can access users' data, and under what circumstances?
  - Did you ensure that these persons are qualified and required to access the data, and that they have the necessary competences to understand the details of data protection policy?
  - Did you ensure an oversight mechanism to log when, where, how, by whom and for what purpose data was accessed?

4. Transparency

Traceability:

- ✓ Did you establish measures that can ensure traceability? This could entail documenting the following methods:
  - ➢ Methods used for designing and developing the algorithmic system:
    - Rule-based AI systems: the method of programming or how the model was built;
    - Learning-based AI systems; the method of training the algorithm, including which input data was gathered and selected, and how this occurred.

- Methods used to test and validate the algorithmic system:
  - Rule-based AI systems; the scenarios or cases used in order to test and validate;
  - Learning-based model: information about the data used to test and validate.
- Outcomes of the algorithmic system:
  - The outcomes of or decisions taken by the algorithm, as well as potential other decisions that would result from different cases (for example, for other subgroups of users).

Explainability:

✓ Did you assess:
- to what extent the decisions and hence the outcome made by the AI system can be understood?
- to what degree the system's decision influences the organisation's decision-making processes?
- why this particular system was deployed in this specific area?
- what the system's business model is (for example, how does it create value for the organisation)?

✓ Did you ensure an explanation as to why the system took a certain choice resulting in a certain outcome that all users can understand?

✓ Did you design the AI system with interpretability in mind from the start?
- Did you research and try to use the simplest and most interpretable model possible for the application in question?
- Did you assess whether you can analyse your training and testing data? Can you change and update this over time?
- Did you assess whether you can examine interpretability after the model's training and development, or whether you have access to the internal workflow of the model?

Communication:

✓ Did you communicate to (end-)users – through a disclaimer or any other means – that they are interacting with an AI system and not with another human? Did you label your AI system as such?

✓ Did you establish mechanisms to inform (end-)users on the reasons and criteria behind the AI system's outcomes?
- Did you communicate this clearly and intelligibly to the intended audience?
- Did you establish processes that consider users' feedback and use this to adapt the system?
- Did you communicate around potential or perceived risks, such as bias?
- Depending on the use case, did you consider communication and transparency towards other audiences, third parties or the general public?

✓ Did you clarify the purpose of the AI system and who or what may benefit from the product/service?
- Did you specify usage scenarios for the product and clearly communicate these to ensure that it is understandable and appropriate for the intended audience?
- Depending on the use case, did you think about human psychology and potential limitations, such as risk of confusion, confirmation bias or cognitive fatigue?

✓ Did you clearly communicate characteristics, limitations and potential shortcomings of the AI system?
- In case of the system's development: to whoever is deploying it into a product or service?
- In case of the system's deployment: to the (end-)user or consumer?

5. Diversity, non-discrimination and fairness

Unfair bias avoidance:

✓ 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 assess and acknowledge the possible limitations stemming from the composition of the used data sets?
- Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?
- Did you research and use available technical tools to improve your understanding of the data, model and performance?

- Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system?

✓ Depending on the use case, did you ensure a mechanism that allows others to flag 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 consider others, potentially indirectly affected by the AI system, in addition to the (end)-users?

✓ Did you assess whether there is any possible decision variability that can occur under the same conditions?
- If so, did you consider what the possible causes of this could be?
- In case of variability, did you establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights?

✓ Did you ensure an adequate working definition of "fairness" that you apply in designing AI systems?
- Is your definition commonly used? Did you consider other definitions before choosing this one?
- 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 systems? Did you consider other potential mechanisms?

Accessibility and universal design:

✓ Did you ensure that the AI system accommodates a wide range of individual preferences and abilities?
- Did you assess whether the AI system usable by those with special needs or disabilities or those at risk of exclusion? How was this designed into the system and how is it verified?
- Did you ensure that information about the AI system is accessible also to users of assistive technologies?
- Did you involve or consult this community during the development phase of the AI system?

✓ Did you take the impact of your AI system on the potential user audience into account?
- Did you assess whether the team involved in building the AI system is representative of your target user audience? Is it representative of the wider population, considering also of other groups who might tangentially be impacted?
- Did you assess whether there could be persons or groups who might be disproportionately affected by negative implications?
- Did you get feedback from other teams or groups that represent different backgrounds and experiences?

Stakeholder participation:

✓ Did you consider a mechanism to include the participation of different stakeholders in the AI system's development and use?

✓ Did you pave the way for the introduction of the AI system in your organisation by informing and involving impacted workers and their representatives in advance?

6. Societal and environmental well-being

Sustainable and environmentally friendly AI:

✓ Did you establish mechanisms to measure the environmental impact of the AI system's development, deployment and use (for example the type of energy used by the data centres)?

✓ Did you ensure measures to reduce the environmental impact of your AI system's life cycle?

Social impact:

✓ In case the AI system interacts directly with humans:
- Did you assess whether the AI system encourages humans to develop attachment and empathy towards the system?
- Did you ensure that the AI system clearly signals that its social interaction is simulated and that it

has no capacities of "understanding" and "feeling"?

- ✓ Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or de-skilling of the workforce? What steps have been taken to counteract such risks?

Society and democracy:

- ✓ Did you assess the broader societal impact of the AI system's use beyond the individual (end-)user, such as potentially indirectly affected stakeholders?

7. Accountability

Auditability:

- ✓ Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI system's processes and outcomes?
- ✓ Did you ensure, in applications affecting fundamental rights (including safety-critical applications) that the AI system can be audited independently?

Minimising and reporting negative Impact:

- ✓ Did you carry out a risk or impact assessment of the AI system, which takes into account different stakeholders that are (in)directly affected?
- ✓ Did you provide training and education to help developing accountability practices?
  - Which workers or branches of the team are involved? Does it go beyond the development phase?
  - Do these trainings also teach the potential legal framework applicable to the AI system?
  - Did you consider establishing an 'ethical AI review board' or a similar mechanism to discuss overall accountability and ethics practices, including potentially unclear grey areas?
- ✓ Did you foresee any kind of external guidance or put in place auditing processes to oversee ethics and accountability, in addition to internal initiatives?
- ✓ Did you establish processes for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks or biases in the AI system?

Documenting trade-offs:

- ✓ Did you establish a mechanism to identify relevant interests and values implicated by the AI system and potential trade-offs between them?
- ✓ How do you decide on such trade-offs? Did you ensure that the trade-off decision was documented?

Ability to redress:

- ✓ Did you establish an adequate set of mechanisms that allows for redress in case of the occurrence of any harm or adverse impact?
- ✓ Did you put mechanisms in place both to provide information to (end-)users/third parties about opportunities for redress?

We invite all stakeholders to pilot this Assessment List in practice and to provide feedback on its implementability, completeness, relevance for the specific AI application or domain, as well as overlap or complementarity with existing compliance or assessment processes. Based on this feedback, a revised version of the Trustworthy AI assessment list will be proposed to the Commission in early 2020

Key guidance derived from Chapter III:

- ✓ Adopt a Trustworthy AI **assessment list** when developing, deploying or using AI systems, and adapt it to the specific use case in which the system is being applied.
- ✓ Keep in mind that such assessment list will **never be exhaustive**. Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying requirements, evaluating solutions and ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders therein.