*CEPS Task Force on*

# *Artificial Intelligence and Cybersecurity*

## *Technology, Governance and Policy Challenges*

---

## *Task Force Evaluation of the HLEG Trustworthy AI Assessment List (Pilot Version)*

*22 January 2020*

Rapporteurs: Lorenzo Pupillo
Afonso Ferreira
Stefano Fantin

# List of Task Force Members and Invited Speakers[1]

**Coordinator and rapporteur:** Lorenzo Pupillo, CEPS
**Rapporteurs:** Afonso Ferreira, CNRS - Stefano Fantin, KU Leuven
**Research Assistant:** Carolina Polito, CEPS

**Advisory Board**
Joanna Bryson, Hertie School for Governance, Berlin
Jean-Marc Rickli, Geneva Centre for Security Policy
Marc Ph. Stoecklin, Security Department, IBM Research Center, Zurich
Mariarosaria Taddeo, Digital Ethics Lab, University of Oxford

**Companies and European Organisations**
Confederation of Danish Industry, Andreas Brunsgaard
Deloitte, Massimo Felici
ETNO, Paolo Grassia
F-Secure, Matti Aksela
FTI Consulting, William Dazy
ICANN, Elena Plexida
JP Morgan, Renata Shepard
McAfee, Chris Hutchins
Microsoft, Florian Pennings
Palo Alto Networks, Sebastian Gerlach
Raiffsen Bank International AG , Martin Koeb
SAP, Corinna Schultze
VECTRA, Matt Walmesley
VISA, Pedro Simoes
Wavestone, Gerome Billois
Zurich Insurance, Marc Radice

**European Institutions, Agencies, Intergovernmental and International Organizations**
Council of the European Union, General Secretariat, Monika Kopcheva*
EDA, Mario Beccia**
ENISA, Apostolos Malatras***

[1] The names and affiliations on this list do not necessarily equate to the person's (or the organization's) endorsement of the content of this report

European Commission, Nineta Polemi

European Central Bank, Klaus Lober**

European Parliament, Adam Bowering

European Investment Bank, Harald Gruber

Financial Conduct Authority, Tim Machaiah

Hybrid CoE - The European Centre of Excellence for Countering Hybrid Threats in Helsinki/Finland, Josef Schroefl**

NATO, Michal Polakow**

OECD, Laurent Bernart

** **In her own capacity***

*** **Observer***

**** **ENISA participates as an observer and did not take part in this evaluation***

**Academics-Think Thanks**

Centre for Economics and Foreign Policy Studies (EDAM), Usaal Sahbaz

Centre for Russia, Europe, Asia Studies (CREAS), Theresa Fallon

DeepIn Research Network/I-COM, Antonio Manganelli

TNO, Alex Sanger

University of Amsterdam, Federica Russo

**Civil Society**

Homo Digitalis, Eleftherios Chelioudakis

Humanity of Things Agency, Marisa Monteiro

**Invited Speakers**

Marcus Comiter, Harvard Kennedy School, Belfer Center

David Clark, MIT Computer Science & Artificial Intelligence Laboratory

Miguel Gonzalez-Sancho, European Commission

Martin Dion, Kudelski

Jouni Kallunki, F-Secure

Andres Ojamaa, Guardtime

Andrea Renda, CEPS

Rob Spiger, Microsoft

## Introduction[2]

The Centre for European Policy Studies launched a Task Force on Artificial Intelligence (AI) and Cybersecurity in September 2019. The goal of this Task Force is to bring attention to the market, technical, ethical and governance challenges posed by the intersection of AI and cybersecurity, focusing both on AI for cybersecurity but also cybersecurity for AI. The Task Force is multi-stakeholder by design and composed of academics, industry players from various sectors, policymakers and civil society.

The Task Force is currently discussing issues such as the state and evolution of the application of AI in cybersecurity and cybersecurity for AI; the debate on the role that AI could play in the dynamics between cyber attackers and defenders; the increasing need for sharing information on threats and how to deal with the vulnerabilities of AI-enabled systems; options for policy experimentation; and possible EU policy measures to ease the adoption of AI in cybersecurity in Europe.

As part of such activities, this report aims at assessing the High-Level Expert Group (HLEG) on AI Ethics Guidelines for Trustworthy AI, presented on April 8, 2019. In particular, this report analyses and makes suggestions on the Trustworthy AI Assessment List (Pilot version), a non-exhaustive list aimed at helping the public and the private sector in operationalising Trustworthy AI. The list is composed of 131 items that are supposed to guide AI designers and developers throughout the process of design, development, and deployment of AI, although not intended as guidance to ensure compliance with the applicable laws. The list is in its piloting phase and is currently undergoing a revision that will be finalised in early 2020.

This report would like to contribute to this revision by addressing in particular the interplay between AI and cybersecurity. This evaluation has been made according to specific criteria: whether and how the items of the Assessment List refer to existing legislation (e.g. GDPR, EU Charter of Fundamental Rights); whether they refer to moral principles (but not laws); whether they consider that AI attacks are fundamentally different from traditional cyberattacks; whether they are compatible with different risk levels; whether they are flexible enough in terms of clear/easy measurement, implementation by AI developers and SMEs; and overall, whether they are likely to create obstacles for the industry.

The HLEG is a diverse group, with more than 50 members representing different stakeholders, such as think tanks, academia, EU Agencies, civil society, and industry, who were given the difficult task of producing a simple checklist for a complex issue. The public engagement exercise looks successful overall in that more than 450 stakeholders have signed in and are contributing to the process.

The next sections of this report present the items listed by the HLEG followed by the analysis and suggestions raised by the Task Force (see list of the members of the Task Force in Annex 1).

[2] The Rapporteurs of the Task Force would like to thank all Task Force Members who kindly and actively participated in the drafting of this document by providing meaningful and insightful comments. In particular, the rapporteurs would like to thank Matti Aksela, Marcus Comiter, Eleftherios Chelioudakis and Marisa Monteiro. The views presented in this report do not necessarily represent the opinions of all participants of the Task Force or their organizations nor do they explicitly represent the view of any individual participant. The views expressed in this report are those of the authors writing in a personal capacity and do not necessarily reflect those of CEPS or any other institutions with which they are associated.

## 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?

### Analysis

Compliance with the provisions of the EU Charter of Fundamental Rights is a mandatory legal obligation for any public and private stakeholder within the EU. Thus, respecting the rights to privacy, data protection, freedom of expression and information, rights of the child, non-discrimination, etc. should not fall under a voluntary framework.[3]

This issue seems particularly pressing as developers and deployers of AI systems, for which the list is mainly intended, might not possess a thorough understanding of the differences between legal requirements and ethical considerations. Hence, these professionals could falsely believe that complying with the EU Charter is a matter of moral choice rather than a legal obligation. According to the recent case-law of the Court of Justice of the European Union (CJEU), namely the Bauer case, the EU Charter has a direct horizontal effect on the relationships between private parties that are governed by EU law. As such, any dispute between private parties is not immune to legally binding measures relative to fundamental rights.

Moreover, the use of the term 'human autonomy' might also seem inaccurate under this section. Human autonomy is interrelated with the EU Charter of Fundamental Rights since it has strong

[3] The AI HLEG states that the EU Charter is legally binding at page 10 of the Ethics Guidelines document. However, the AI HLEG neither makes the same statement within the AI Assessment List, nor provides a reference to the page 10. In this way, AI developers/deployers that read only the AI Assessment list and not the full document will miss the point. Thus, the addition of a short statement or reference to page 10 within the AI Assessment list could have an added value in creating more awareness and responsibility.

connections with human integrity and dignity. It can be argued that human autonomy is partly achieved thanks to the protection of the freedoms and the rights described under the EU Charter's provisions. However, the term 'human autonomy' is not used in any Article of the EU Charter's provisions, while it also encompasses broader conceptual variations that go beyond human rights law.

The concept of autonomy functions differently in a variety of contexts, and it plays various roles in the theoretical accounts of persons, conceptions of moral obligation and responsibility, justification of social policies and in numerous aspects of political theory. Thus, linking the concept to the provision related to fundamental rights does not seem to be fit for purpose and could create confusion for the reader of the Assessment List. The provisions of the Charter are specific. Therefore, any terms that are not defined under these provisions, such as human autonomy, should not be examined under the "Fundamental rights" section.

Nevertheless, the interaction of AI systems with the autonomy of end-users during the decision-making process is an ethical issue of utmost importance. Therefore, the inclusion of questions of autonomy in the Trustworthy AI Assessment List should be endorsed overall. However, such questions should not be placed within the Fundamental Rights section. To avoid the term being interpreted vaguely by the AI developers and deployers, the use of a working example of what constitutes an interference with user autonomy would be appropriate as it could offer guidance to stakeholders and help them provide better feedback.

On the other hand, the last two questions of this section, do seem to account for fundamental rights concerns in a proper manner. Both of these questions enjoy a good connection with Article 8 of the EU Charter of Fundamental Rights, i.e. the protection of personal data, and aim to further enhance the right of the data subject to be informed about the existence of automated decision-making, under Articles 13, 14 and 15 of the GDPR.

Furthermore, questions on the respect of fundamental rights should only apply to use cases that have been identified as having the potential to cause harm. In those limited cases, the questions should be made more specific, to make sure deployers of AI/ML systems are considering the same factors when establishing compliance processes related to the respect of fundamental rights.

When it is a question of algorithmic decisions, it is important to keep the scope appropriately open. While we understand the particular concerns/interests with Artificial Intelligence i.e. Machine Learning and Deep Learning solutions, we tend to ignore that these are a combination of algorithms and data. Algorithms and data feature on a spectrum ranging from plain 'if-then-else' or 'loop' statements to sophisticated neural networks, a handful of data points to the 'whole internet'. Cathy O'Neil clearly shows how devastating and biased the most 'simple' paper form and Excel worksheet can be.[4] Therefore, AI/ML may pose particular challenges regarding 'Blackbox' features (deep learning/neural networks) but there are a plethora of high-profile examples where the algorithms were simple enough ('pen and paper') and their implications have been ignored with consequences.

Regarding the point "Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?" it may be helpful to consider expanding or

[4] Cathy O'Neil (2016), *Weapons of Math Destruction*, Crown, New York.

enumerating the definition of "unintended" to ensure it is fully considered by users of the Assessment List. For example, unintended interactions could stem from traditional limitations of AI (e.g., limited by the training data used to craft the system), from failures of AI (e.g., a potentially incorrect AI system or glitch in the AI system), or from purposeful attacks on the AI system (e.g., from manipulated input into the system).[5] By expanding the definition of unintended, users of the Assessment List can better understand the full scope of unintended behaviour that may emerge from the AI system.

### 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?

### Analysis

The questions presented relative to human agency, in spite of being high level, nevertheless attempt to address real ethical challenges related to Human-Computer Interaction, and issues related to human enhancement through AI systems. Also, they seek to obtain feedback on procedures and checks and balances that AI developers and deployers have in place, aiming to prevent overconfidence in AI systems among end-users.[6] However, it is important to understand that the safeguards mentioned under the Human Agency item – if also based on AI – can themselves fail in pernicious ways. For instance, one such safeguard is for the AI system to generate a confidence level in the AI system's output. These confidence levels are often used to signal to the end-user the confidence of the AI system in its own decision. The intended purpose of this is to prevent overconfidence in the system, as the user will be informed when the system is confident in its output or when the system is processing input that leads it to be less confident about its decision (e.g., stemming from previously unseen data, noisy input, etc.). As a result, users may come to rely on these safeguards. However, these mechanisms can also be subverted by an adversary: it is possible to craft input so that an AI system will fail with high confidence. As a result, the safeguards themselves also need to be audited for failures, especially for cases in which their failure will be correlated with failures of the underlying AI system they are trying to protect.

[5] See Marcus Comiter (2019), Attacking Artificial Intelligence, Harvard Kennedy School-Belfer Center, pages 17-27.
[6] In applying AI to cybersecurity in the upcoming decade, it is not realistic to think that AI will replace humans. AI can be used to automate parts of the work of SOC analysts, but a lot of human expertise will still be required.

### 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 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?

### Analysis

The questions addressed in regard to human oversight seem to be suitably located. They complement the legal provisions on the right to human intervention provided under Article 22(3) GDPR and aim to investigate further under which measures such an intervention is guaranteed and how it is exercised in practice. Furthermore, they positively seek to investigate related internal audits and possible remedial routes.

The question related to ensuring human control calls for a level of expertise by humans to exercise 'true' control or oversight. In particular, it is necessary to protect against what can be called the 'half-automation problem'. By this is meant the phenomenon in which, when tasks are highly but not fully automated, the human operator tends to rely on the AI system as if it were fully automated. Examples of this have occurred multiple times with semi-autonomous vehicles: these vehicles are not fully automated and require full human attention and the ability for humans to intervene at split-second notice. However, use patterns show that humans do not maintain this level of control despite requirements to do so, instead tending to fall asleep or watch Netflix. This should be explicitly considered, as broad directives to maintain human control may not be followed in practice.

Particularly relevant is also the reference to the need for stop buttons or 'kill switches' to maintain human control over AI processes, especially in the field of an automated response. But are kill switches an option in all cases? Do humans in control have to take back full control? In some cases, like the self-driving vehicles or auto-pilots on modern aircraft, the issue of when to return control is an open issue that still needs to be answered.

*Suggestions for the "Human agency and oversight" requirement*

1. The "Fundamental rights" section should be removed from "Human Agency and Oversight", and placed under a separate requirement, while clearly indicating that compliance with the provisions of the EU Charter is a legal obligation and not a moral choice.

   Multiple examples (identifying non problematic and problematic use cases by category, possibly healthcare, manufacturing, SDV or closely related) of what constitutes a limitation on end-user's autonomy based on the feedback acquired during the pilot phase, should be provided so that the AI developers and deployers can obtain a better and more specific idea about what this term encompasses in practice.

2. We suggest to make explicit the need to train the teams concerned. Indeed, AI and self-learning applications have specific behaviours and teams should be trained to deal with them.

## 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)?

## Analysis

AI-dependent solution engineering can benefit by being considered holistically in the context of threats, risks, mitigations, and testing. A solution manufacturer can address some of the holistic picture based on anticipated customer usage scenarios, but adopters of AI-based solutions also need to consider their environment, dependencies, and operationalisation of solutions, whether based on AI or not.

This section provides many valuable examples of individual questions for organisations to consider based on the current state of the art. It could be further enhanced by placing individual artificial intelligence considerations into a broader, more comprehensive cybersecurity risk-based context that uses risk analysis to guide engineers to prioritise activities to achieve the most impactful security contribution to the overall solution and its anticipated usage environment.

Therefore, it would make sense to address the need for doing a risk assessment in this section. This is an input for defining the relevant cybersecurity measures. The risk assessment should assess:

- The AI application cybersecurity sensitivity (in terms of confidentiality, integrity, availability, and traceability)
- The need for specific measures related to the type of data that feeds the AI (speech, images, videos, texts do not have the same cybersecurity requirements)
- The potential privacy issues related to AI training and use
- The risks related to outsourcing (data used, model development, hosting)

Although the HLEG report covers many relevant items, the Assessment List misses a mention of adversarial attack methods against machine learning models, such as input attack, model evasion, stealing, and poisoning, but also attacks against the supply chain that may allow attackers to poison, *trojanize*, or backdoor models at their source. It is not possible to build models that are completely resilient to adversarial attacks since it is not possible to test every potential input to a model. Therefore, emphasis should be put not only on model resilience but on monitoring systems that utilise machine-learning models to detect inputs indicative of adversarial attacks. The report should also mention that traditional cybersecurity protection measures should also be employed in public-facing systems that serve machine-learning-driven applications.

As AI-based systems become more commonplace, the incentive will arise for adversaries to learn how to attack them. To remain competitive, companies or organisations may dangerously ignore safety concerns, downplay already identified risks or abandon robustness guidelines to push the boundaries of their work, or to ship a product ahead of a competitor. This trend towards low quality, fast time-to-market is already prevalent in the Internet of Things industry and is considered highly problematic by most cybersecurity practitioners. Similar recklessness in the AI space could be equally negatively impactful. As such, AI researchers and engineers will need to be aware of the sorts of ethical issues they may encounter in their work and understand how to respond to them.

The “Resilience to attack and security” section touches on the dual-use potential of artificial intelligence. We agree that this is a valid concern that would require a much broader assessment.

*Suggestions for the "Resilience to attack and security" section*

1. To help with being more exhaustive and specific we suggest using in addition the following wording:

   "Did you consider different types of attack methods (data pollution, application fooling or privacy attacks), the different components that could be targeted (training data, model, physical infrastructure) and the different nature of attacks (opportunist attack, massive cyber-attacks, cyber-espionage)?"

2. We need to be very careful with the word "resilience" as many things are included behind this concept today in cybersecurity. Here it seems that we address protection more than detection and reaction. In order to avoid confusion within the cybersecurity community, it may be better to use the word "protection" against attacks.

3. As the research in this domain is constantly evolving, it would make sense to add the need to make sure the latest trends have been taken into consideration with an additional sub-question:

   "Did you consider state-of-the-art vulnerabilities and measures while assessing and designing protection mechanisms?"

4. To keep the scope of the guidelines quite open, for each of the guidelines section an as exhaustive list as possible of current examples should be provided (e.g. annex) recognising that the field is evolving fast and the language/guidelines need to accommodate nascent and as yet unknown technologies.

5. Section 2 also mentions considering if a solution could be "dual-use", however the document does not clarify what "dual-use" means specifically. More explanation of how consumers of the publication are expected to use and apply this guidance would be beneficial.

6. We suggest promoting ad hoc initiatives for "a fundamental cyber hygiene programme" with specific requirements for AI/ML.

## 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?

## Analysis

We feel that the "Fallback plan and general safety" section of the Trustworthy Assessment List should also be focused more on risk assessment.[7] Widespread adoption of AI will mean more applications making automated decisions for us. It is essential that the consequences of moving those decisions from humans to machines are carefully evaluated. Risk assessment is a fundamental step in this process, and the document should provide more advice on how to do this as well as explore how this is done in existing applications. While the section as a whole contains valuable recommendations, we feel it should put greater stress on the need for dynamic monitoring and anomalous behaviour detection to be applied to the output of decision-making models. The section covers fallback plans in depth. However, a fallback plan can only be activated when it has been determined that the system in question is no longer functioning within specified parameters. As such, detecting this is as important as knowing what to do in the case of a failure.[8]

Indeed, the true danger of adversarial attacks is that they are often difficult if not impossible to detect. To the AI system, the input is processed as normal, and to a human, the input may appear completely normal. In particular, there are adversarial or input attacks that can be completely invisible to the human eye. As such, a fallback plan must be preceded by detection mechanisms that would alert that adversarial attacks are occurring (e.g., manual audits, etc.), after which a fallback plan may be activated. In order to make sure a fallback plan can sufficiently address all of these varied forms of adversarial attacks, it may be helpful to include an ad hoc taxonomy of adversarial attacks or a directive such as "...even those not able to be perceived by the human eye".[9]

---

[7] A risk-based approach could be further elaborated in a similar structure to that of the German Data ethics Commission's recommendations. Concrete risk levels need to be developed in line with user cases/different AI applications.

[8] The detection of an issue in complex systems involving for instance machine learning, deep learning and deep reinforcement learning is not trivial. In the case of decision-making systems that operate on a continuous/real-time or near continuous/real-time basis, the detection becomes truly challenging, especially *if real-time monitoring objectives must be met*. The following approach comes to mind:

- Another system is deployed to control/monitor the decisions made by the primary system. It would either trigger the fallback automatically or request the involvement of a human operator (human on the loop). The monitoring system itself will by definition have to exhibit an appropriate degree of sophistication and therefore complexity; otherwise, a simple decision-making system would have been deployed in the first place. The involvement of a second complex system raises the question of the detection of issues in the monitoring process, as the monitoring system could itself be biased or flawed.

*If the real-time issue detection constraint is lifted*, the following approaches come to mind:

- A monitoring system that processes the logs and validates input/outcome expectations could be deployed. Most probably, the amount of data and detection of off-patterns will require machine or deep learning techniques. It would either trigger the fallback automatically or request the involvement of a human operator (human **on** the loop). In all cases, the fallback plan would have to account for the time delay between the issue detection and the continued operation of the primary system.
- One or multiple human operators are kept **in** the loop at all times. Whether or not the human operator is assisted by a monitoring system, the primary system must operate at a pace that is commensurate with human operator checks/interventions. Here too, the fallback plan would have to account for the time delay between the issue detection and the continued operation of the primary system.

Note that the technical expertise of human operators needs to match the complexity of the deployed systems.
In general, rigid "checks and balances" may defeat the purpose of deploying decision-making systems in the first place.

[9] See Comiter (2019).

*Suggestions for the "Fallback plan and general safety" section*

1. We suggest including an ad hoc taxonomy of adversarial attacks
2. We should require copies of previous states of the application to be kept and a capacity to restore them to be maintained. In case the application is subject to training data pollution, returning to a standard functioning could be hard (impossibility to know when the data started to be polluted, difficulty to re-train the model from the beginning…)

### 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?

## Analysis

The "Accuracy" section of the Trustworthy Assessment List mentions bias as a side note. We feel that bias is an extremely important topic and should be covered in more depth. The process of training a machine learning model often involves a tight feedback loop that can cause unforeseeable drifts and biases. Furthermore, the data used to train models may contain statistical, societal, or human biases. Imbalances in datasets are incredibly difficult to find and fix, given that they arise from social and organisational reasons, in addition to technical reasons. Detection and remediation of biases in machine learning models is currently a hot research topic in technical circles, and discussion of how biased models may affect society and human rights is an equally hot topic for experts in humanities and social sciences.

The "Accuracy" section should also consider how the impact of model selection and design may affect its robustness. Different methodologies provide different trade-offs in terms of a model's robustness,

explainability, and traceability. Furthermore, the document should consider other uses of machine learning outside of predictive applications.

We welcome the fact that the "Accuracy" section adequately covers risk and impact assessment recommendations in this area.

*Suggestions for the "Accuracy" section*

1. It may be a right idea to add a specific question on the trade-off between accuracy and other criteria like robustness:

   "Was the trade-off between accuracy and robustness discussed between business teams, development teams and cybersecurity teams to find the best compromise?"

2. It is important to stress the need for uncorrupted data. It is essential to protect the data acquisition chain to make sure that no one can modify the data – from the moment it is acquired to the point when it is provided as inputs to the algorithm to make the predictions. Therefore, we suggest adding here:

   "Did you put in place measures to ensure that the data used is comprehensive, up to date and uncorrupted?"

### 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?

## Analysis

Reliability and reproducibility in cybersecurity are closely related to ensuring robustness. If the application is sensitive to variations in environment it is more likely to be fooled by evasion attacks. Before testing it (which is highlighted here), it should be taken into account in the development process.

For some AI/ML solutions, reproducibility may only be possible with full log information – time, market, situation dependencies. The question then is how long the logs can reasonably be kept – SDV, autopilot, etc. Are we therefore rather talking about algorithmic outcome comparability?

While the "Reliability and Reproducibility" section of the Trustworthy Assessment List contains recommendations on monitoring machine-learning-based applications, it would be preferable for it to go into more detail and, as such, provide concrete examples on how to conduct this exercise. It would also be insightful for the list to provide recommendations on monitoring for security purposes. The quality of any decision made by an AI solution significantly depends on the quality and quantity of the data used. An absence of large sets of high quality data is, in general, one of the major obstacles to the application of AI solutions. We appreciate that it is hard to verify whether an AI-based system is behaving correctly because it is not possible to test all conceivable combinations of inputs. The section should then mention logging, which is of great use when attempting to achieve reproducibility. A robust logging strategy also brings benefits such as auditability and traceability, which can be core requirements for critical systems.

However, it should be noted that over-emphasis of logging may, in some cases, lead to adverse results. In a situation where a complex model is trained online using high volumes of input data, an in-depth logging strategy may require petabytes of data to be stored – for every event one would need to store the input data, input metadata, all model parameters, results, and actions taken. This may be justifiable for life-critical applications such as automated airplanes or surgical robots, autonomous weapon systems or facial recognition for surveillance purposes, but for most non-critical applications it may risk unfair competition between big and small players, whereby the latter may simply not be able to afford to abide by such regulations due to costs of implementation.

*Suggestions for the "Reliability and Reproducibility" section*

1. We suggest adding the following question:

   "Did you define reliability and reproducibility requirements to take into account in AI development projects?"

2. We also suggest to use techniques like Randomisation, Adversarial Training, Noise Prevention, Defensive Distillation, Ensemble Learning as well in order to enhance AI application reliability and reproducibility.

*Suggestions for the "Technical robustness and safety" requirement*

Overall, the report on Ethics Guidelines for Trustworthy AI contains many valuable conclusions. The guidelines presented in the Trustworthy Assessment List section of this document touch on significant topics and address many relevant concerns. The list itself stays at high-level and could be refined with some additional granularity. One area that is largely missing is the need for basic security controls on systems running machine-learning-based services. We recommend the checklist to include more recommendations, such that it provides potential directions towards the resolution or mitigation of the concerns raised. We would welcome more detail in a few key areas, including recommendations on monitoring and logging, more in-depth discussion of bias and feedback loops, and mention of AI-specific attacks (and mitigation strategies including monitoring of both inputs and outputs).

Key recommendations

To further enhance the European Commission's document, we recommend more emphasis on valuable topics such as:

1. Importance of continuous monitoring
2. Awareness of AI-specific attacks
3. Importance/relevance of traditional cyber security measures
4. Awareness of ethical issues one may encounter
5. Risk and impact assessment
6. Emphasis on possible effects of bias
7. Consideration of machine learning outside of the prediction space
8. Awareness of the trade-offs in model design (accuracy, robustness, explainability, etc)
9. Promoting fundamental cyber hygiene programs with specific requirements for AI/ML
10. Putting in place measure to ensure that the data used is comprehensive, up to date, and uncorrupted.
11. Enhancing AI application reliability and reproducibility

With the added focus, the document would be of major value for machine learning practitioners aiming to create not only effective, but robust, secure and ethical AI solutions.

## 3. Privacy and data governance

### 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?

## Analysis

One of the issues that need to be underlined concerning this section is the fact that the clear legal obligations that arise from the GDPR's provisions are described under this requirement with a different wording from the one used in the GDPR's text, and thus presented as new ethical requirements. This could create confusion for AI developers and deployers between legal obligations that they must follow and extra requirements that they can apply voluntarily. Consequently, this may lead to a lack of transparency for some AI solutions and practical challenges in relation to GDPR, especially when there is a requirement to clarify to data subjects how their personal data is used and give them a meaningful explanation of the assumptions and drivers behind a fully AI-driven automated decision with a significant impact on their data.

In this respect, the question "Did you consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?", is legally coined as follows: "Did you manage to develop the AI system or train the model while complying with the legal principle of data minimisation (Article 5 GDPR) and respecting the legal requirement of data protection by design and by default (Article 25 GDPR)?" Even though the processing of special categories of personal data is in principle prohibited and allowed only on specific legal bases (Article 9 & Article 22.4 GDPR), the aforementioned question creates the false impression that processing of sensitive data for developing and training AI systems is something that the AI developers and deployers can themselves decide upon. This lack of conformity with the GDPR language should be avoided in the new version of the list. Otherwise, AI developers and deployers will probably misinterpret legal obligations arising from the GDPR´s text as voluntary ethical choices that they can adopt or not.

Finally, the questions posed in this section do not seem to reflect ongoing ethical debates about privacy, such as those on group privacy that are quite GDPR-oriented. We believe that this section could serve as a great opportunity to ask the AI developers/deployers questions regarding their practices on group profiling and predictive analytics as well as to further discuss approaches and methods that can be used to address the social and ethical problems posed by such group profiling activities.

*Suggestions for the "Privacy and data protection" section*

This section should be more specific about the following privacy-related issues:

1. Did you assess data protection issues while outsourcing the solution? (did you assess shared training risks, etc.?)
2. Did you assess the risks to being able to retrieve data from the trained model?
3. Did you raise developers' awareness about data protection? (not to publish training data on a public forum when trying to debug, to favour in-house model development framework hosting instead of using cloud solutions when data desensitisation is not possible.

Anonymisation is very hard, if not impossible, even with techniques such as Multi-Party Computation. We suggest replacing the word "anonymisation" with "pseudonymisation".

## 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?

## Analysis

In this section, terms that are clearly defined under data protection law, such as "data quality" are presented rather vaguely. According to Article 5 of the Convention 108 of the Council of Europe (1981 – which laid down the basic foundations for future EU data protection laws, including GDPR), data

quality is associated with the data protection principles of lawfulness, fairness, purpose limitation, data minimisation, accuracy, and storage limitation. The questions of the list do not cover the full spectrum of the notion of data quality, as rolled out in Convention 108, since they focus mostly on standardisation, data accuracy, and data security. As such, the questions could, again, create confusion for AI developers and deployers.

The data must be protected during the whole acquisition chain and solution as blockchain could be considered to guarantee data integrity.

*Suggestions for the "Quality and integrity of data" section*

A dedicated question about Big Data infrastructure security may be interesting here:

> "Did you apply Big Data security best practices in both your training and production environments?"

*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?

## Analysis

The section "Access to data" appears to miss the opportunity to raise important questions related to open data governance, open data policies for making datasets available to the public or to other stakeholders, interoperability and data portability, and challenges related to the use of public sector data. Such questions could fuel the existing debate on data silos and offer the European Commission important insights on this topic for future policy initiatives.

*Suggestions for the "Privacy and data governance" section*

Despite the negative issues underlined above, this requirement serves an important role: the answers/feedback from the stakeholders involved will bring to the European Commission important insights about the state of play of data protection compliance in the EU in the context of AI development and deployment. Nevertheless, we would like to suggest the following:

1. To use the terminology and wording of the GDPR when referring to legal obligations arising from its provisions.
2. To distinguish between mandatory legal obligations that AI developers/deployers must follow, and voluntary ethical requirements that AI developers/deployers might choose to adopt. Colour codes could help in such a distinction (for example legal obligations could be presented in the text with a blue colour, and ethical requirements with a green colour).
3. To include questions related to ethical issues on privacy, such as group privacy.
4. To include questions on open data governance, open data policies for making datasets available to the public or to other stakeholders, interoperability and data portability, and challenges related to the use of public sector data
5. To add the following subpoints:
    a. "Do you have processes and procedures implemented to detect unauthorised access to the data or unauthorised modification of the data?".
    b. Access to the data used to train the AI system makes formulating attacks on AI systems significantly easier and is therefore likely to be a first step in an AI attack pipeline. As a result, data security should not be limited only to traditional access control, but also involve a heightened alertness to unauthorised access or hacking of data systems.
6. Further, we recommend increasing the protection of data used to train AI systems beyond just access control. In order to protect data from being used to formulate attacks on AI systems, data must be protected as are passwords with high levels of encryption. This will not always be possible or feasible, but where it is, it will go a long way to protecting AI systems in the face of inevitable cyberattacks and hacking. This system works well because it increases the overall resiliency of the system, as a hack of encrypted data will not necessarily endanger the AI system built on this data.

## 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).

### Analysis

In the questions under the "Traceability" section focused on the methods used to design and test the AI systems, the HLEG differentiates between Rules-based and Learning-based systems recognising in this way their inherent differences in functionality and development.

*Suggestions for the "Traceability" section*

In addition to the decision output of the model, retaining an audit trail of the parameters that influenced the decisions the algorithms make should also be part of the traceability requirement.

### 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?

## Analysis

Under the "Explainability" section, stakeholders are called upon to reflect whether there is an added value in using AI systems for their business, while the concept of interpretability by design lies at the core of this section. Such questions further enhance the legal requirements arising from Article 22 GDPR, as well as Recital 71 GDPR, where the only clear reference on explainability exists ("to obtain an explanation of the decision reached after such assessment and to challenge the decision").

*Suggestions for the "Explainability" section*

1. We suggest adding the following sub-question to the explanation question in order to keep this 'one-time' explanation as proof for possible further investigations:

   "Did you ensure these explanations are kept as proof long enough to meet a posteriori investigation needs?"

2. We suggest adding the wording in brackets:
   a. to what degree the system's decision influences [people?] and the organisation's decision-making processes?
   b. what the system's business model is (for example, how does it create value for [people?] and the organisation)?

### 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?

## Analysis

The "Communication" section further enhances the legal obligations that arise from Article 13(2)(f), 14 and 15 GDPR. Thus, the feedback acquired from these questions will probably help the European Commission in identifying and tracking weaknesses and shortcomings in the current practices of the AI developers/deployers to move forward with its related legislative proposal. We recall that, as European Commission President-elect, Ursula von der Leyen promised ad hoc actions to address the human and ethical implications of Artificial Intelligence (AI) during her first 100 days in office.

*Suggestions for the Transparency section*

The questions raised under this requirement stay at a high level but are nevertheless helpful. They will provide important feedback on stakeholder practices related to the transparency of AI systems. Moreover, they further enhance and complement the legal requirements that arise from the GDPR's provisions. The HLEG is moving in uncharted waters and is taking the first step towards a better assessment of the transparency of AI systems. Here are our suggestions:

1. To avoid repeating the same questions under different requirements. For example the question "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?" under this requirement has a similar meaning with questions under the "Human agency and oversight" requirement, i.e. "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?". Such repetitions, if further used, could result in interviewee fatigue.
2. The degree of transparency should be different depending on the defined risk level of the AI application. The AI HLEG should work on different categories of required transparency.
3. As currently drafted, the assessment list cannot be operationalised in a reasonable amount of time and with limited resources, making the effort rather challenging for small companies, start-ups and regulators or whoever is designated to supervise and audit such operationalisation. Indeed:
    a. The assessment list includes questions that are too open-ended and not sufficiently contextualised. The lack of examples and scenarios risks preventing the user from obtaining a proper understanding of the purpose and scope of the questions.
    b. The assessment list appears largely in the wrong order for practical use and is not aligned with the way a typical AI product is developed and deployed to the market.
    c. Should the checklist be deployed on the market as it is, it would be too burdensome for most of the businesses investing in AI-based technologies, and especially for companies with limited resources and investment capacity.

Therefore, the assessment list should:

- Be streamlined to avoid too prescriptive and redundant questions, which would prevent companies, especially if small and with limited resource, from successfully operationalising the checklist.
- Be more aligned with standard development processes of AI-based technologies. Provide guidance on the specific design and deployments contexts in which the questions would assume relevance.
- Promote more agile and flexible processes incentivising the use of documentation models while at the same time guaranteeing flexible processes, formats and tools.

## 5. Diversity, non-discrimination, and fairness

### 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 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?

## Analysis

The questions focus on the strategies and procedures used by the stakeholders to mitigate machine, algorithms and data bias. Attention is paid to issues related to the representativeness of different parts of the population in the datasets. Such representativeness is important for public stakeholders who might use such datasets to assist in their decision-making processes. However, there is no specific reference to these stakeholders in the text. Some of the questions attempt to ensure a working definition of fairness and to explain the methodology used for the selection of such a definition. They also reflect and address the lack of conformity that exists with regards to the meaning of the notion of fairness, and, thus, go in the right direction.

*Suggestions for the "Avoidance of unfair bias" section*

1. We suggest adding a sub-question about the relationship between possible decision variabilities and the way to detect and set fallback mechanisms, such as:

   "Did you establish the detection mechanisms, and did you think about fallback plans for identified cases of decision variability?"

2. The Commission could consider introducing mechanisms and tools for comprehensive human review and oversight of algorithmic processes. This could be done via independent authorised third parties.

3. We suggest adding the following words in brackets:

   "Did you consider the diversity and representativeness of users [including proxies thereof] in the data?"

4. We suggest that the HLEG specifies the EU definition of fairness. What are the expectations of AI/ML solutions operating, say, globally? Develop and train and test for specific markets?

5. Since some methods to address diversity, non-discrimination and fairness pose security risks, it is important to mention the need to design new methods to allow for audits of systems without compromising security, such as restricting audits to a trusted third party rather than publishing openly. Therefore, we suggest adding the following sub-question:

   "Have you considered how processes and tests to ensure diversity, non-discrimination and fairness affect the security of the AI system?".

### 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?

## Analysis

Under this section, stakeholders are called upon to reflect whether the developers of their AI systems take into consideration the necessities of people with special needs or disabilities or those at risk of exclusion when developing AI systems. Moreover, more questions are asked regarding the representation of different societal groups in the team, and the consequences following the lack of such representation for accessibility and inclusiveness. However, there are no questions related to access issues arising from financial criteria or technological illiteracy, which could create exclusion as well.

### 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 organization by informing and involving impacted workers and their representatives in advance?

## Analysis

This section introduces questions related to the use of participatory design for the development of AI systems. We believe that it is positive to touch upon this topic, but we would like to see some

examples of what participatory design is and what the differences with other models of design are, such as value-sensitive design. Such examples could put all the stakeholders on the same page and ensure better responses. Indeed, if we take cybersecurity as our focus here, the past has shown that it is really hard to integrate cybersecurity measures once the system has already been developed, and even designed. Cybersecurity stakeholders should be involved from the project framing phase in order to make sure the final product will be as vulnerability-proofed as possible.
Adopting, designing and developing AI Systems require participating stakeholders to embrace a more scientific mind-set, including being comfortable with a trial and error journey, understanding and accepting risks and tests that will fail; and continuously testing the feasibility of the AI System.

*Suggestions for the "Diversity, non-discrimination and fairness" section*

1. To provide boxes with good practices and/or examples of successful stakeholder participation in the development of AI systems.
2. To address the lack of accessibility to AI systems arising from financial criteria or technological illiteracy, as well.

## 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?

## Analysis

Generally, the broader issue regarding societal well-being is very important, because not only the impact of automation in our jobs but even the impact that new technologies have, for instance, on mental health or health in general have to be understood.

The questions posed in the Assessment List, though, focus on a very specific and important topic, the environmental impact of AI systems and development, and focus particularly on the energy used by data centres.[10]

[10] Besides energy consumption, another issue that appears to arise is high carbon dioxide emissions. In a recent paper, Strubell, Ganesh and McCallum (2019) show that neural networks are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud computing time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware (neural architectures emit the carbon dioxide equivalent of nearly five times the lifetime emissions of an average American car). See Emma Strubell, Ananya Ganesh, Andrew McCallum (2019), **Energy and Policy Considerations for Deep Learning in NLP.** https://arxiv.org/abs/1906.02243.

### 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?

## Analysis

The questions regarding social impact deal with the attachment and empathy towards the system and with how to ensure that the AI system clearly signals that its social interaction is simulated, and it has no capacities of understanding and feeling. Therefore, these kinds of recommendations need to offer a measurable variable to avoid workarounds. In this regard, is mentioning that the system is a bot/chatbot enough to prove the interaction with the system is simulated? If that is the case, then the policy needs to be reworded to inform the user they are not speaking to a human being.

The Assessment List should pose the question in a way that is easier for the developers to understand and come up with a clear answer. How people from different fields communicate with each other is a key point for reaching understanding. In the same way, with respect to the job loss and de-skilling of workforce risks, the need for a measurable variable in the short-term has been pointed out. Without providing a way to measure this variable there will be no way for developers to provide an answer.

### 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?

## Analysis

Regarding this section, the main issue seems to be that unless measurable variables are provided, this recommendation will be very easy to bypass. Like every software system, an AI-powered system is just a piece of code and the developer can describe it in whichever way they want. Thus, when dealing with practical applications in the field, assessing the broader societal impact seems to be particularly challenging but an important element of the overall assessment.

## 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?

### Analysis

Accountability will likely be one of the most complex themes in AI. In legal terms, it is hard to establish how to defend someone who developed an AI system that caused major harm. In more general terms, auditability is pivotal because there is no transparency and accountability if a system cannot be audited. If we are not able to perceive and understand the system, then we cannot trust it. AI Systems can make auditability and traceability challenging, and the speed at which they typically evolve may result in large-scale errors, in a very short timeframe. Thus, auditability of any system is in a close relationship with trust. In this context, it is important to identify who will be responsible and will have the skills for these audits. Today, deep learning, the closest to AI that we have nowadays, is not traceable, and logging is not possible because of the way it works. It is a black box. Given that there could be no trust without audit, this is a major point of discussion. If we manage to audit, then we will have a basis for trust. Not because we understand the system, but because we will know it is possible to understand it. This should be pointed out as a main short-term goal.

### Minimising and reporting negative impacts and Documenting trade-offs

- 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?
- 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?

### Analysis

The answers to these questions are strongly dependent on a company's size. In the case of small start-ups, carrying out an impact assessment of the AI system is often not trivial. Not because they are not sensitive to the topic, but because of their business model. Normally, once a start-up receives the funding for a project, it focuses as much as possible on developing the product and waits for the next investment to arrive. Start-ups do not deploy economic effort to perform these kinds of assessments in the beginning, because often it is a matter of their project growing or not. Maybe the way in which such projects are financed should be reconsidered, in order to help start-ups conduct these kinds of assessments.

The same is true for training and education to help in developing accountability practices. While some AI developers have, indeed, the financial resources for providing such services for the whole team, most of the small players lack such a possibility. The same could be said with respect to the external guidance for overseeing ethics and accountability. In this regard, it is noticeable that the process of thinking about ethical notions is already in place. Most start-ups are proactively trying to understand what they are building. The problem is that the ethical debate around AI is still too complex to be timely and adequately addressed by small start-ups.

Finally, some AI developers make use of open-source legal templates. Indeed, affording a legal adviser is often not possible for these communities. Thus, again, these questions need to be better targeted.

*Suggestions for the "Minimising and reporting negative impacts and Documenting trade-offs" section*

We suggest creating an entity entrusted with such assessments by the state before putting AI-products on the market in order to help the SMEs

### 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?

### Analysis

As cyber-resilience is currently a hot topic (stating that we can be attacked and making sure that most critical activities are able to restart quickly), this part "ability to redress" is key. The two questions are very important even though they arrive very late in the document and seem very short compared to some other paragraphs.

In the section on the ability to redress, it is important to define the entity that would set precise and measurable criteria that are needed in order to guarantee adequate information. Currently, there are software systems on the market that can and are doing more harm than any existing AI. Thus, without policies for the accountability of all software including AI, it will be impossible to evaluate all software systems on the market whether they are using AI or not.

It is very complex to understand the policies, rules, and regulations, how they are made, and how to apply them even in the national and EU systems. Hence, the focus should be on cooperating with developers in order to help them achieve proper legal and ethical compliance. In this regard, communities are making a lot of effort by themselves, but one thing is what we need to have in place, and another thing is what we have, and the gap between the two is still quite big.

# ABOUT CEPS

Founded in Brussels in 1983, CEPS is widely recognised as the most experienced and authoritative think tank operating in the European Union today. CEPS acts as a leading forum for debate on EU affairs, distinguished by its strong in-house research capacity and complemented by an extensive network of partner institutes throughout the world.

## Goals

- Carry out state-of-the-art policy research leading to innovative solutions to the challenges facing Europe today
- Maintain the highest standards of academic excellence and unqualified independence
- Act as a forum for discussion among all stakeholders in the European policy process
- Provide a regular flow of authoritative publications offering policy analysis and recommendations

## Assets

- Multidisciplinary, multinational & multicultural research team of knowledgeable analysts
- Participation in several research networks, comprising other highly reputable research institutes from throughout Europe, to complement and consolidate CEPS' research expertise and to extend its outreach
- An extensive membership base of some 132 Corporate Members and 118 Institutional Members, which provide expertise and practical experience and act as a sounding board for the feasibility of CEPS policy proposals

## Programme Structure

### In-house Research Programmes

Economic and Finance
Regulation
Rights
Europe in the World
Energy, Resources and Climate Change
Institutions

### Independent Research Institutes managed by CEPS

European Capital Markets Institute (ECMI)
European Credit Research Institute (ECRI)
Energy Climate House (ECH)

### Research Networks organised by CEPS

European Network of Economic Policy Research Institutes (ENEPRI)
European Policy Institutes Network (EPIN)