European Commission

Ethics guidelines for trustworthy AI

On 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. This followed the publication of the guidelines' first draft in December 2018 on which more than 500 comments were received through an open consultation.

According to the Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment

Download the Guidelines in your language below:

BG | CS | DE | DA | EL | EN | ES | ET | FI | FR | HR | HU | IT | LT | LV | MT | NL | PL | PT | RO | SK | SL | SV

The Guidelines put forward a set of 7 key requirements that AI systems should meet in order to be deemed trustworthy. A specific assessment list aims to help verify the application of each of the key

requirements:

- Human agency and oversight: AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches
- Technical Robustness and safety: AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.
- Privacy and data governance: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.
- Transparency: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.
- Diversity, non-discrimination and fairness: Unfair bias must be avoided, as it could could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.
- Societal and environmental well-being: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.
- Accountability: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured.

The AI HLEG has also prepared a document which elaborates on a Definition of Artificial Intelligence used for the purpose of the Guidelines.

Download the Definition of AI in your language below:

BG | CS | DE | DA | EL | EN | ES | ET | FI | FR | HR | HU | IT | LT | LV | MT | NL | PL | PT | RO | SK | SL | SV

Piloting Process

The document also provides an assessment list that operationalises the key requirements and offers guidance to implement them in practice. Starting from the 26th of June, this assessment list underwent a piloting process, to which all stakeholders were invited to test the assessment list and provide practical feedback on how it can be improved.

Feedback was received through different tracks:

- An open survey or "quantitative analysis" sent to all those who registered to the piloting
- In-depth interviews with a number of representative organisations to gather more detailed

feedback for different sectors
- Continuous possibility to upload feedback and best practices through the European AI Alliance

The piloting phase closed on 1 December 2019

Based on the feedback received, the AI HLEG presented the final Assessment List for Trustworthy AI (ALTAI) in July 2020. ALTAI is practical tool that translates the Ethics Guidelines into an accessible and dynamic (self-assessment) checklist. The checklist can be used by developers and deployers of AI who want to implement the key requirements in practice. This new list is available as a prototype web based tool and in PDF format.

See also

A European approach to artificial intelligence

Thèmes associés

Technologies numériques avancées Intelligence artificielle
