ETHICS GUIDELINES FOR AI

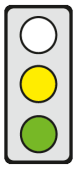
cep**PolicyBrief** No. 2019/16

Centrum für Europäische Politik

KEY ISSUES

Objective of the Communication: The Commission presents – non-binding – ethical principles and key requirements for artificial intelligence (AI) which must be followed by developers and users of AI across the EU.

Affected parties: AI developers, companies and individuals who use or are affected by AI.



Pro: (1) The guidelines are a useful first step towards an EU-wide concept of "ethically acceptable" AI.

(2) The "Trustworthy AI Assessment List" helps to provide guidance on how to implement the ethical requirements.

Contra: Concrete examples should be added to further facilitate implementation of the guidelines.

The most important passages in the text are indicated by a line in the margin.

CONTENT

Title

Communication COM(2019) 168 of 8 April 2019: **Building Trust in Human-Centric Artificial Intelligence,** referring to **Ethics Guidelines for Trustworthy AI** of 8 April 2019, set up by the "High-level Expert Group on Artificial Intelligence"

Note: Page references without further citation refer to the Communication COM(2019) 168, references with "GL" to the "Ethics Guidelines for Trustworthy AI" referred to in the Communication.

Brief Summary

- **General Background**
  - Artificial intelligence ("AI") systems are software systems designed by humans that [GL p. 36]
    - perceive their environment through collection of data,
    - interpret the collected data,
    - draw conclusions from this interpretation, or process the information derived from the data, and
    - decide the best actions to achieve a given complex goal.
  - AI has the potential to transform our world for the better: it can e.g. improve healthcare, reduce energy consumption, make cars safer and enable farmers to use natural resources more efficiently [p. 1].
  - Artificial intelligence ("AI") raises ethical and legal questions, inter alia because it enables machines to "learn" independently and to take automated decisions without human intervention. Such decisions may be tampered with by cyber-attackers or be wrong, e.g. drawn from datasets that are incomplete or biased and thus unrepresentative [p. 2].

- **Ethics Guidelines for "trustworthy AI"**
  - AI must serve humans with the aim of increasing human well-being. Therefore, AI must be "trustworthy" and "human-centric", i.e. "developed in a way that puts people at the centre" [p. 1, 2].
  - To ensure this, the Commission set up an "independent AI high-level expert group" ("AIHLEG") to develop "Ethics Guidelines for Trustworthy AI" [the "guidelines", p. 2 et seq., GL p. 4].
  - The guidelines
    - are non-binding and aim to establish an "ethical level playing field" within the EU and make the EU a global leader in "trustworthy AI" [p. 2, 9, GL p. 4],
    - are addressed to all stakeholders, e.g. AI developers and deployers – i.e. private or public AI users who offer products and services to others –, but also to end-users and others directly or indirectly affected by AI [GL p. 5, 14].
  - The guidelines stipulate that to be "trustworthy", AI must – throughout its entire lifecycle – be [p. 3, GL p. 5-10]
    - lawful, i.e. comply with the applicable law and in particular with legally binding fundamental rights,
    - robust, in particular from a technical perspective, e.g. resilient to cyberattacks, and
    - ethical, i.e. ensure adherence to ethical values even when laws are not up to speed with technical progress. Although non-binding, these ethical values help identify
      - what should be done rather than what can be done with AI, and
      - how AI may implicate fundamental rights and their underlying values [GL p. 10].
  - The guidelines do not deal with "lawful" AI but aim to offer guidance for the creation of "ethical" and "robust" AI. Nevertheless, fundamental rights are also relevant for "ethical AI": fundamental rights have underlying values and are bestowed on individuals by virtue of their moral status as human beings. [GL p. 6, 10]

Authors: Alessandro Gasparotti and Dr. Anja Hoffmann, LL.M. Eur. | hoffmann@cep.eu
cep | Kaiser-Joseph-Strasse 266 | 79098 Freiburg | Germany | Telephone +49 (0)761 38693-107 | www.cep.eu

– The guidelines [GL p. 9, 14 et seqq.]
  - outline the "foundations" of "trustworthy AI", which are grounded in fundamental rights and give rise to four "ethical principles", and
  - transform these theoretical principles into seven key requirements for "trustworthy AI".

► **Fundamental rights and ethical principles as the "foundations" of "trustworthy AI"**

– "Trustworthy AI" must be based on the fundamental rights and values enshrined in the EU Treaties, the EU Charter of Fundamental Rights and international law [p. 2, GL p. 9, 10]. Fundamental rights comprise inter alia the following rights, which are of particular importance for AI:
  - respect for human dignity implying that every human being possesses an "intrinsic worth",
  - freedom of the individual, e.g. freedom of expression and of assembly, the right to private life and privacy, and
  - equality, non-discrimination and solidarity.

– Four ethical principles must be respected by AI. These principles reflect fundamental rights [GL p. 11-13]:
  - respect for human autonomy: humans should keep full self-determination over themselves and oversight over AI; AI should not unjustifiably deceive or manipulate humans;
  - prevention of harm: AI should not cause harm or otherwise adversely affect humans;
  - fairness: benefits and costs of AI should be justly distributed within society, discrimination and unfair biases avoided and effective redress mechanisms against AI-driven decisions should be available; and
  - explicability: the purpose of AI should be communicated, processes transparent and decisions as explainable as possible, depending on the context and the severity of the consequences of a wrong decision.

– These principles may conflict, e.g. facial recognition technology can reduce crime (prevention of harm), while limiting privacy and individual liberty (i.e. human autonomy).

– Such trade-offs should be acknowledged and solved through "reasoned reflection". Human dignity cannot, however, be balanced against other rights [GL p. 13].

– The guidelines transform these ethical principles into a non-exhaustive list of seven key requirements which "trustworthy AI" should meet [p. 4-6, GL p. 15-20]:

► **Seven key requirements for the realisation of "trustworthy AI"**

– **Human agency and oversight:** To safeguard human autonomy and decision-making, AI should support human agency and allow for human oversight [p. 4, GL p. 15, 16]:
  - AI should help humans to make better, more informed choices and thus support human agency.
  - AI should allow for human oversight which can be achieved through governance mechanisms; e.g.
    - ensuring levels of human discretion and the ability to override AI decisions and to decide not to use AI in specific situations,
    - monitoring AI activity; the less oversight that is possible, the more extensively AI must be tested.
  - Public enforcers must have the ability to exercise oversight over the use of AI systems in line with their mandate.
  - Negative effects on fundamental rights should be assessed before the development of AI and must be reduced or justified.

– **Technical robustness and safety:** To prevent harm, AI must be technically robust and safe, i.e. [p. 5, GL p. 16, 17]
  - have a level of safety proportionate to the magnitude of the risk posed by an AI system,
  - be reliable, secure and resilient to attacks, e.g. hacking and manipulation,
  - create the same, i.e. reproducible results under the same conditions, so that its behaviour can be described,
  - be accurate and inform users about the likelihood of possible mistakes (e.g. limited accuracy), and
  - have safeguards that enable a fall-back plan in case of problems, e.g. involve a human operator.

– **Privacy and data governance:** To prevent harm to privacy, adequate data governance is needed [p. 5, GL p. 17]:
  - the quality, integrity and relevance of data fed into an AI system must be ensured to avoid biases and mistakes,
  - humans must have full control over their data collected for or by AI, and
  - this data must not be used unlawfully.

– **Transparency:** To support "explicable" AI, its elements – data, systems, and business models – must be transparent; in particular [p. 5, GL p. 18]:
  - AI systems must be traceable, e.g. by documenting their decisions and the underlying process (including the data).
  - Algorithmic decision-making processes must be as explainable as possible and their explainability must be weighed against a potential reduction in accuracy and conditional upon whether the AI in question has "a significant impact on people's lives".
  - Users must be aware that they are interacting with an AI system, be informed about the system's limitations and be allowed to request human interaction "where needed to ensure compliance with fundamental rights".

– **Diversity, non-discrimination and fairness:** To be fair, AI must be designed to allow everyone equal access to the product or service. Affected stakeholders should be involved in design processes. "Unfair" biases e.g. in data sets should be avoided, as they could lead to discrimination. [p. 6, GL p. 18]

– **Societal and environmental well-being:** To be fair and prevent harm, AI should be sustainable, ecologically and societally friendly; its effects on the environment, humans, society and democracy should be monitored [p. 6].
– **Accountability:** To be fair, AI should be designed in a way that it can be audited – without the need to disclose intellectual property related or other proprietary information –, especially when its use affects fundamental rights. Negative impacts should be reported, and minimised and adequate redress mechanisms foreseen. [p. 6, GL p. 20]
– The requirements can be implemented by [GL p. 20-23]
  - "technical methods", e.g. by implementing procedures that AI has to follow or must not follow, and "ethics by design", i.e. enforcing compliance with ethical norms from the beginning of the AI design process, and/or
  - "non-technical methods" of governance, such as regulation, codes of conduct, standardisation and certification.

► **Assessment list, piloting phase and international consensus-building**
– To ensure that the guidelines will be implemented in practice, the AIHLEG operationalised the requirements into a "Trustworthy AI Assessment List" which is meant to guide primarily AI developers and deployers to achieve "trustworthy AI" [GL p. 25-31].
– During a piloting phase until December 2019, stakeholders can test the list and provide feedback. The AIHLEG will update the guidelines in early 2020. [p. 7, GL p. 24]

Policy Context

In 2018, the Commission published an AI strategy [COM(2018) 237] and a "Coordinated Plan" [COM(2018) 795], which inter alia aim to ensure appropriate legal and ethical rules for AI [cf. cep**PolicyBrief** No. 2019/13, see also cep**PolicyBriefs** No. 2019/10 and No 2019/12]. In 2017, the European Parliament proposed an ethical conduct code [EP resolution] and, in 2019, demanded an ethical framework for "human-centric AI" [EP resolution]. The Council highlighted the importance of implementing ethics guidelines for AI within the EU and at the global level [Conclusions of 02/2019]. In May 2019, the OECD published its AI ethics guidelines, endorsed by the G20 [cf. cep**Input** No. 2019/07].

Options for Influencing the Political Process

Directorates General:	DG for Communications Networks, Content & Technology
Committees of the European Parliament:	Industry, Research and Energy (leading)
Federal Germany Ministries:	Interior, Building & Community, Justice & Consumer Protection (leading), Data Ethics Commission of the German Government
Committees of the German Bundestag:	Education, Research and Technology Assessment (leading); Enquete Commission "Artificial Intelligence", chair: Daniela Kolbe (SDP).

ASSESSMENT

Economic Assessment

The Commission's aim to foster "trustworthy AI" may facilitate the acceptance of this technology. However, **the guidelines are too general and vague to be implemented directly**. Moreover, they do not increase legal certainty because they do not help AI developers to comply with applicable laws. It is thus questionable whether the guidelines alone will establish an EU-wide ethical level playing field for AI and thus make the EU a global leader in trustworthy AI. **Nonetheless the** seven **key requirements provide a comprehensive frame for the further development of more precise** – e.g. sector specific – **guidelines**.
In addition, where ethical principles conflict during the development and use of AI, the public should also be informed about how the developer solved the ethical trade-off. Such information does not harm companies nor does it place a heavy burden on them. The guidelines should offer solutions or examples on how such ethical trade-offs can be solved.
The seven requirements may be assessed as follows:
(1) It is appropriate that AI must be subject to human oversight and ensure levels of human discretion e.g. to override AI decisions that are biased or have been tampered with. However, no definition of the appropriate level of discretion is provided.
(2) It is appropriate that AI deploys a level of safety proportionate to the magnitude of its risk. Thus, to use the same standards for an AI-controlled nuclear reactor as for AI that provides users with music suggestions would be disproportionate. However, by failing to outline any standard, the guidelines encourage companies to downplay the possible risks of their AI in order to meet their own minimal standards.
(3) The quality, integrity, and relevance of the data fed into an AI system is key to reducing the risk of unreliable results. Competitive markets tend to ensure this result. Nevertheless, examples should be included to provide guidance in the handling of data.
(4) In order to foster transparency, AI decisions have to be as traceable and explainable as possible. Well-informed users are in a better position to assess how far they want to rely on an AI. In particular, users should be able to view

and correct information relating to them, such as their age or interests, that is collected by AI and used in its decision-making.
(5) To ensure non-discrimination, it should be clarified when a bias is “unfair” . It is unlikely that it will be possible to involve all affected stakeholders in design processes. However, manufacturers have a vested interest in comprehensive inclusion, as feedback improves the quality of AI.
(6) It is unclear what “societal well-being” means. For example, the “Trustworthy AI Assessment List” states that, to be societally friendly, AI should counteract the risk of job losses. It is unclear whether, by substituting HR managers and performing fairer, e.g. non-discriminatory, interviews, AI can be seen as increasing societal well-being given that some human resources managers would lose their jobs.
(7) To guarantee the accountability of AI without reducing incentives to innovation, the requirement of auditability should – as far as possible – avoid the disclosure of proprietary information. Beyond this, in order to increase legal certainty for both companies and consumers, adequate redress mechanisms are essential.

The “Trustworthy AI Assessment List” helps to provide guidance on how to implement the ethical requirements. However, **concrete examples should be added to further facilitate implementation of the guidelines.** In addition to the – useful – real-world testing of the list by AI developers and operators, comprehensive and transparent public discourse about AI should be actively encouraged, in which ethics bodies, universities, parliaments and the public must be more actively included. At the same time, protection – possibly going beyond the guidelines – against future risks and the possible establishment of ethical limits on the use of AI in the EU should be discussed in detail.

Legal Assessment

Competence

Unproblematic. These are not legal but “ethical“ guidelines and in addition they are non-binding.

Subsidiarity

Unproblematic (see above).

Proportionality with regard to Member States

Depends on the design of follow-up measures.

Compatibility with EU law in other respects

“Trustworthy” AI must observe the law and in particular fundamental rights. However, since the guidelines contain “ethical” rather than legal requirements applicable to AI, there is no doubt about their compatibility with EU law. Instead, the question – requiring urgent discussion – is whether and under what circumstances the increasing use of AI is ethically acceptable or even necessary.

The guidelines are a useful first step towards an EU-wide concept on “ethically acceptable” AI. Ethical perspectives, values and moral attitudes vary in the EU despite common basic values, thus there is a risk that the ethical requirements applicable to AI will become fragmented. The four ethical principles that have been selected – human autonomy, prevention of harm, fairness and the explicability of AI – are an appropriate starting point. They are rightly derived from the fundamental rights and values that apply in the EU, and in particular from the guarantee of human dignity which is the primary fundamental value [Art. 2, sentence 1 TEU] and results from the constitutional traditions common to the Member States [Meyer-Borowski, EU Charter of Fundamental Rights, Art. 1 para. 26]. Thus, in questions of interpretation, there is a link to constitutional judicature, and major discrepancies between ethics and the law can be avoided. Since the canon of EU values places humans at the centre of the EU project [Calliess/Ruffert, Art. 2 EUV, para. 11], the EU’s decision to opt for “human-centric” AI is logical.

Impact on German law

Dependent on the design of follow-up measures.

Conclusion

The guidelines are too general and vague to be implemented directly. Nonetheless the key requirements provide a comprehensive frame for the further development of more precise guidelines. The guidelines are therefore a useful first step towards an EU-wide concept of “ethically acceptable” AI. The “Trustworthy AI Assessment List” helps to provide guidance on how to implement the ethical requirements. However, concrete examples should be added to further facilitate implementation of the guidelines.