January 2019

# AI ethics guidelines

## Draft CEC feedback on the draft proposal of the High-Level Expert Group on Artificial Intelligence

*Document: "Draft Ethics guidelines for trustworthy AI" (click here)*

### Context

The European Union's High-Level Expert Group on Artificial Intelligence (AI) has recently published the first draft of its ethical guidelines on AI. The document is a starting point for an extensive reflection upon the ethical, social and economic implications AI has and could have in the future. Since the development of AI will be shaped jointly by management decisions, systems design and legal frameworks, the EU can now create a level playing field for implementing socially, economically and environmentally beneficial systems. Following the formulation of the ethical framework, the expert group will draft policy and investment recommendations to be submitted to the European Commission.

In the following paragraphs, CEC highlights, discusses and critically evaluates key aspects of the draft ethical guidelines on AI, following the structure of the document. As a responsible European social partner and association representing managers in Europe, CEC takes the impact of AI on decision-making, working and private life extremely seriously, therefore actively contributing to the debate.

## Rationale and framework

In the introductory part of the document, the rationale and framework of the proposed approach are presented. In short, AI shall contribute to increase human well-being and the common good by being human-centric and trustworthy. Trustworthy AI is defined as respecting "fundamental rights and applicable regulation, as well as core principles and values, ensuring an 'ethical purpose'", and by being "technically robust and reliable."

The first part of the document explains the "ethical purpose" AI should have. The second part defines a set of requirements for trustworthy AI – of both technical and non-technical nature. Finally, the document contains a practical assessment list for trustworthy AI.

## I. Respecting Fundamental Rights, Principles and Values - Ethical Purpose

The document defines the **"ethical purpose" of AI as respecting the rights, principles and values** as enshrined in the EU Treaties and in the Charter of Fundamental Rights of the European Union. Unfortunately, the delimitation between the concept of rights, principles and values appear rather vague and even tautological in their current formulation. The "rights-based approach" taken delivers insufficiently on an ethical case for these rights in proper terms.

Furthermore, the document is ambiguous over the term "ethical purpose", since AI systems shall on the one hand "comply with" values, principles and rights (p. 3) and on the other serve them as a purpose. The latter case implies that AI, and thus also organisations developing it, can only be ethical if they serve the purpose of advancing fundamental rights. At the same time, these rights and their underpinnings can evolve over time, making the need for a stronger ethical foundation of the guidelines even more important.

### Central question: what shall we do?

Shifting away from the questions of rights, it may be argued that the ground-breaking trait of AI lies in its unmeasurable **potential to create a utopian or dystopian society** from the contemporary point of view and compared to previous technologies. This brings up classical ethical questions about the "good life", as well as the Kantian questions about what the human being is, what the human can hope for, what it can know and what it should do. Since the human, at least seemingly, could soon know and hope (for) almost everything, the central question appears to be: what, if almost everything is indeed possible, should the human do? And who is the human in this position? Of course, these questions are closely related to the purpose of work both conceptually and factually as a historically defining feature of human life.

### Ethical principles and challenges

Later in the chapter, a set of five ethical principles is defined: beneficence, non-maleficence, autonomy, justice and explicability. Considering the powerful long-term potential of AI for delivering on some of the most pressing contemporary challenges, the **principle of "sustainability"** [1]could be added. AI could have a positive long-term effect to ensure a living basis for everyone (cf. SDGs) and to limit pressure on Earth's life-supporting systems. On the other hand, the development of AI itself, in terms of energy and raw material use, has to be examined critically.

Finally, the last part of the chapter discusses some ethical challenges posed by AI, including consent, transparency of AI systems, mass citizens' scoring, lethal autonomous weapon systems and potential long-term concerns. As far as **scoring systems** are concerned, employees should be protected from extensive and unnecessary surveillance and have the right to be forgotten at the

[1] Please find CEC's Sustainable Leadership Guidelines here: https://www.cec-managers.org/sustainableleadership/

end of their employment relation. When it comes to the speculative long-term concerns, CEC reaffirms its opposition to a techno-deterministic view, which would acknowledge the possibility of artificial consciousness or attributing rights to technical objects performing tasks, even if complex and seemingly humanoid. This view is contrary to a human-centred approach to AI and stands in contrast to humanistic, religious and evolutionary worldviews.

## II. Realising Trustworthy AI

The second chapter lays out the main requirements for AI's trustworthiness as well as the methods to meeting the requirements in the development, deployment and usage phases of AI. The ten requirements are as follows:

1. Accountability
2. Data Governance
3. Design for all
4. Governance of AI Autonomy (Human oversight)
5. Non-Discrimination
6. Respect for (& Enhancement of) Human Autonomy
7. Respect for Privacy
8. Robustness
9. Safety
10. Transparency

These conditions cover the elements identified in the first chapter, providing a globally satisfactory set of requirements that AI systems shall fulfil.

Remark on accountability: Complying with the  the human-centred approach CEC stands for, individuals shall remain at the heart of decision-making, the ultimate responsibility and liability for errors or biases in the system design shall lie in those in charge of the system. At all critical moments at least, a human evaluator and decision maker is needed, ideally with ethical knowledge and relevant skills.

Remark on safety: to effectively ensure safety, another requirement is needed beforehand - the precautionary principle. The precautionary principle[2] foresees that in the case activities can lead to morally unacceptable harm, even if uncertain, measures should be taken to avoid or diminish it. "Morally unacceptable harm" usually refers to harm to humans or the environment that is: "threatening to human life or health, or serious and effectively irreversible, or inequitable to present or future generations, or imposed without adequate consideration of the human rights of those affected"[3]. A sound risk analysis and its constant update are needed to assess the potential

[2] World Commission on the Ethics of Scientifc Knowledge and Technology: The Precautionary Principle (2005): http://unesdoc.unesco.org/images/0013/001395/139578e.pdf
[3] Ibid.

damages.

After having presented the requirements to achieve trustworthy AI, the guidelines continue with technical and non-technical methods for that purpose. These include, among others, for the technical part, ethics by design, testing, auditability and for the non-technical one, regulation, standardisation, stakeholder & social dialogue and education.

Remark on stakeholder and social dialogue: making use of the diversity among workers, managers and employers and other stakeholders can be a tool to flexibly adapt to developments of AI and labour market related implications. Social dialogue in particular can inform decision-making on the development of AI systems at company, sectoral, national and European level. Societies in the future will also require institutions to hold deeper and critical debates about the implication technology has on work and life. Ultimately, social dialogue may gain in importance to fit this requirement.

Remark on education and awareness to foster an ethical mind-set: throughout lifetime, prospective decision-makers and AI designers shall be equipped with the necessary knowledge and skills to deal with ethical questions. Being able to understand the technical, ethical and socio-economic implications of AI will prove increasingly important. Particularly a scenario of self-reinforcing algorithms, based on utility calculations, may prove both ethically and economically problematic – requiring critical and empathic humans. Such algorithms are ethically problematic, because they may – particularly if no measures for traceability are implemented – restrict the scope of potential decisions illegitimately, de-facto excluding contingent developments (e.g. showing results only based on previous decisions). They are economically problematic, because the necessary space for creativity that is needed to innovate, could be restricted through algorithms serving a utilitarian logic.

## III. Assessing Trustworthy AI

The final chapter contains a list operationalising the concepts contained in the previous chapters. This list serves as a non-exhaustive tool for compliance.