Trustworthy AI - Ethics Guidelines of the EU Commission

Ethics and artificial intelligence? Do they go together? Hasn't it already been decided that the market alone determines what is feasible and permitted? No, because with the use of AI in industry, the question of ethics has again come to the fore.

In the last two to three decades we have become accustomed to the fact that digitalization and with it artificial intelligence have expanded into more and more areas of the consumer world without any state taking care of the question of whether this is right and proper, whether ethics and morals of our society, the freedom of the individual and the democratic constitutional state are even considered. Often enough this has not been the case.

It was only the Internet of Things and the incipient use of AI in industry and other areas of the economy and society that brought the subject of ethics onto the agenda, first and foremost in Europe, where the General Data Protection Regulation (GDPR) has already driven a stake into the ground for more responsible handling of personal data.

The first Ethics Committee for Artificial Intelligence was established by the Federal Ministry of Transport and Digital Infrastructure (BMVI) on the topic of "Automated and Networked Driving" and presented its report in June 2017. In April 2019, a High-Level Expert Group on Artificial Intelligence (AI HLEG), convened by the EU Commission in the summer of 2018, published a definition of trustworthy AI on the one hand and ethics guidelines for the development and use of trustworthy AI systems on the other. (The book KI-Kompass für Entscheider, which will be published by Hanser in August 2020, deals with these guidelines in chapter 5.4).

The guidelines, which fill about 40 pages, claim to support developers and users of AI systems in meeting basic ethical principles. The guidelines set three basic requirements for trustworthy AI:

- It should be legally compliant and follow all applicable laws and regulations.
- It should be based on ethical principles and values.
- It should be reliable, both technically and socially, because even AI developed with the best of intentions can cause unintended damage.

Even if the guidelines are only the ideas of a group of experts and do not have the status of regulations or EU laws, they are to be welcomed as guidelines. It was high time that ethical frameworks were defined in which the further development of AI was put on a socially desirable leash.

The four basic principles of the European Union are cited and declared to be the general basis of AI ethics: respect for human self-determination, damage prevention, fairness and explainability. Thus, the guidelines place human autonomy absolutely in the foreground. Self-controlling systems, which replace the decision-making power and responsibility of humans, are therefore not to be classified under trustworthy AI.

According to these guidelines, artificial intelligence systems should take into account the special vulnerability of certain groups - for example, children, the disabled and other people who are repeatedly excluded. The asymmetry of power between rich and poor, entrepreneurs and dependent employees, economy and consumers should be counteracted rather than reinforced by AI systems. This is based on the experience that a number of AI systems currently in use seem to pursue the exact opposite objective. For

example, job applications or housing requests can be pre-sorted and filtered in a racist or misogynist manner, depending on how the AI used has been trained. Artificial intelligence systems should therefore, according to the guidelines from Brussels, always be verifiable as to whether they comply with these ethical principles.

The second part of the guidelines explains seven concrete requirements that trustworthy AI should fulfil:

1. Human action, fundamental rights and ultimately human supervision must be guaranteed.
2. AI must offer technical robustness and security.
3. Data protection and data management also and even more so apply to AI.
4. The functioning of AI systems, and in particular their respective decision making, should be transparent and traceable.
5. AI systems should serve diversity, non-discrimination and fairness.
6. The environmental and social well-being must be given the highest priority in the design of AI systems.
7. The results of AI use must be verifiable, negative impacts should be minimized, legal remedies and accountability should be ensured.

In order to ensure compliance with these principles, the development and provision of AI systems should include technical and non-technical aspects.

In the third part of the guidelines, practical suggestions show how an evaluation list can help in the development and use of trusted AI systems. Finally, some very concrete examples are given, which either demonstrate the particular benefits of AI application, for example in the health sector, or the dramatic damage when it comes to autonomous, lethal weapon systems.

The guidelines are not a European law or a basic order like the GDPR. But they should be made known to a wider public, because they can help to turn European AI systems into role models. "Compliance with the guidelines, as the experts expressed in the introduction, should not hinder the introduction of AI, but should itself become a trademark that significantly increases the economic success of trustworthy systems worldwide". (Ulrich Sendler, KI-Kompass für Entscheider, Hanser Verlag, Munich, ISBN: 978-3-446-46295-3, published in August 2020, chapter 5.4)