Effective Human Oversight: Conditions and Implications of the Proposed EU AI Act from an Interdisciplinary Perspective

2023-07-31 15:34:43

Authors

Holger Hermanns, Anne Lauber-Rönsberg, Markus Langer, Kevin Baum, Sven Hetmank, Sarah Sterz, Philip Meinel, Sebastian Biewer

Abstract

With its proposal for an AI Act, the EU is planning to adopt a horizontal approach, according to which artificial intelligence systems are to be regulated to varying degrees depending on the risk associated with their use. A human-centric model of artificial intelligence is to be pursued as a central element. Therefore, an essential component is the human oversight of AI systems, which according to Art. 14 must be carried out whenever such systems are applied in high-risk situations. The purpose of this contribution to the AISoLA workshop is to discuss the different requirements for the design of human supervision envisaged by the proposals of the European Commission, the Council and the Parliament, highlighting subtle differences. We discuss how these are to be evaluated according to legal, ethical, psychological, and system design principles. For this purpose, we will first outline the allocation of obligations for the implementation of effective human oversight to the agents involved.

From a legal perspective, we discuss the stringency of the distribution of duties and highlight possible ambiguities. According to the New Legislative Framework of the EU, these uncertainties are to be eliminated by the creation of harmonized standards. We will examine whether this mechanism is useful, also in view of the proposals of the Parliament and Council to regulate General Purpose AI Systems (GPAI) and foundation models.

From an ethical perspective, the proposed human oversight of AI systems raises several important considerations that we will discuss. They relate to fairness and transparency of, as well as the accountabilities and responsibilities within the overall system and its use context, paired with the need for meaningful human control and intervention. In addition, the ethical implications of human oversight of AI systems also affect the actual public trust in AI systems (as well as in decision-making processes involving AI systems) and the extent to which that trust is justified.

Turning to a psychological perspective, we will discuss the implications of having to exert ?effective human oversight? from a perspective of the operator who will be responsible for this oversight in practice. We will also discuss several challenges to effective oversight that are only partly mentioned in the AI Act (e.g., automation bias).

We will finally take on the system design perspective and discuss concrete system requirements that are to contribute to effective human oversight. This will cover the question how far the technological basis for effectuating human oversight is actually ready for use, respectively being developed in research labs. This will pinpoint a number of challenges for AI and software research. For instance, this involves tools to explicate decision processes or trace system malfunctions, as well as explainability approaches that aim to enable humans to effectively oversee systems.

Overall, this contribution thus provides an interdisciplinary evaluation of the current proposals for human oversight from four perspectives crucial for a successful future implementation of the AI Act, highlighting current limitations, open questions, and possible ways forward.

Keywords

AI Act, AI ethics, AI regulation, automation bias