Implementing AI Ethics in the Design of AI-assisted Rescue Robots

Désirée Martin
Institute for Technology Assessment and Systems Analysis (ITAS)
Karlsruhe Institute of Technology
Karlsruhe, Germany
https://orcid.org/0009-0007-1511-0594

Michael W. Schmidt
Institute for Technology Assessment and Systems Analysis (ITAS)
Karlsruhe Institute of Technology
Karlsruhe, Germany
https://orcid.org/0000-0002-4602-1478

Rafaela Hillerbrand
Institute for Technology Assessment and Systems Analysis (ITAS)
Karlsruhe Institute of Technology
Karlsruhe, Germany
https://orcid.org/0000-0002-7890-3715

Keywords—ethics guidelines, values, principles, indicator system, VCIO model

I. Introduction

For implementing ethics in AI technology, there are at least two major ethical challenges. First, there are various competing AI ethics guidelines and consequently there is a need for a systematic overview of the relevant values that should be considered. Second, if the relevant values have been identified, there is a need for an indicator system that helps assessing if certain design features are positively or negatively affecting their implementation. This indicator system will vary with regard to specific forms of AI technology. An adequate indicator system for the ethical development of recommendation algorithms, for example, will diverge considerably from another for autonomous road vehicles, although both are based on shared values. In this contribution, we propose solutions to both challenges with regard to the special case of the development of an AI-assisted rescue robot.

II. Towards an Normative Consensus on AI

In order to find a solution to the first challenge, we compare prominent AI ethics guidelines [1]–[4] and recent proposals for AI regulation [5], [6]. On this basis, we identify shared values and principles and systematize them. To base the ethical development of AI technology on shared values is important, since liberal democracies exhibit a reasonable pluralism concerning comprehensive moral doctrines [7]. In our systematization, we take action-guiding, normative principles to aim at (or be in accordance with) the values of "justice", "well-being" and "understanding". "Value", in this context, refers broadly to states, which are morally desirable or are instrumentally valuable for such desirable states. Principles provide orientation on how to act to reach such a state. Explainability, for example, is a principle that aims at the value of understanding [8]. Understanding and explainability have a moral dimension when we think of people affected by AI, that do not have the chance to understand what happens. Understanding might be regarded as a final value, but is also of instrumental value to other final values, such as justice and its corresponding principles, like accountability. The result is a relational list of shared values and principles that should be implemented in AI technology.

III. Towards an Indicator System for the Ethical Development of AI-assisted Rescue Robots

Based on the resulting relational list of shared values and principles and in an interdisciplinary dialogue with technical developers and end-users we propose an indicator system for the ethical design of AI-assisted rescue robots. A theoretical background for the construction of the indicator system is the so-called VCIO model, which is hierarchically composed of values, criteria, indicators, and observables [9]. Our proposal complements the VCIO model with the category of principles. Their inclusion is necessary, since principles figure prominently in the considered ethics guidelines and allow for a more nuanced structure of the indicator system. Principles are further specified by criteria. Indicators, in turn, point to observable features of the technological system that allow to measure the degree to which the system meets the relevant criteria reliably. Suppose, for example, that an AI-assisted robot is intended to support the reconnaissance of hazardous substances. Consider the value of understanding, the higher-level principle of explainability, and the lower-level principle of transparency. A corresponding criterion might be the transparency of uncertainties in the AI-assisted detection process. One indicator in that case might be "Are uncertainties made transparent throughout the operation?". A possible observable is: "Yes, information on uncertainties is visualized for the operator of the robot in a map."

Research presented in this contribution has been supported by the German Federal Ministry of Education and Research (BMBF) within the subproject "Ethical Issues Concerning the Opportunities and Risks of AI-Assisted Robotics for Radiological Hazards" of the collaborative KIARA project (grant no. 13N16277).

978-1-6654-5713-2/23/$31.00 ©2023 IEEE

References

[1] 2017 Asilomar conference (Beneficial AI), "Asilomar AI Principles," *Future of Life Institute*, 2017. https://futureoflife.org/ai-principles/ (accessed Oct. 08, 2021).

[2] L. Floridi *et al.*, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds & Machines*, vol. 28, no. 4, pp. 689–707, Dec. 2018, doi: 10.1007/s11023-018-9482-5.

[3] High-Level Expert Group on Artificial Intelligence set up by the European Commission, "Ethics guidelines for trustworthy AI," European Commission, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[4] The IEEE Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, version 2," Version 2, 2017. [Online]. Available: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

[5] EU Commission, *Proposal for a Regulation of the European Parliament and of the Council laying down harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, COM/2021/206 final,*. 2021.

[6] White House Office of Science and Technology Policy, *Blueprint for an AI Bill of Rights*. 2022.

[7] J. Rawls, *Justice as Fairness: A Restatement*. Cambridge, Massachusetts: Harvard University Press, 2001.

[8] W. Fleisher, "Understanding, Idealization, and Explainable AI," *Episteme*, pp. 1–27, Nov. 2022, doi: 10.1017/epi.2022.39.

[9] S. Hallensleben *et al.*, "From Principles to Practice - An interdisciplinary framework to operationalise AI ethics," VDE, Bertelsmann Stiftung, Frankfurt a. M. / Gütersloh, 2020. [Online]. Available: https://www.ai-ethics-impact.org

2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS) | 978-1-6654-5713-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ETHICS57328.2023.10155062

Authorized licensed use limited to: Kennesaw State University. Downloaded on October 13,2025 at 18:36:29 UTC from IEEE Xplore. Restrictions apply.