The Ethics of AI Ethics, Thilo Hagendorff (2020)

Thilo Hagendorff's paper "The Ethics of AI Ethics: An Evaluation of Guidelines" was published in the Minds and Machines journal in February 2020. At the time, Hagendorff worked for the Cluster of Excellence *Machine Learning: New Perspectives for Science* at University of Tuebingen. Currently he leads a research group at the University of Stuttgart and focuses on the topics AI Ethics and the intersection of Machine Learning and Cognitive Science.
The motivation behind this semi-systematic overview on AI ethics guidelines stems from the plethora of publications on ethical principles targeted to regulate recent AI technologies. Hagendorff aims to depict the current state of AI ethics and discusses the issues that are currently limiting the transfer of such guidelines into practice. In terms of methodology, the author identifies common AI ethics guidelines that were published in the preceding five years. This selection encompasses 22 guidelines with the majority being scientific publications. Among the remaining literature pieces are 3 positional papers that can be connected to the three "AI superpowers" China, United States and the European Union as well as corporate guidelines by Google, OpenAI and other important players in the industry. Regarding Hagendorff's approach, he notes which of the chosen 22 AI principles are discussed within these publications. Some principles that Hagendorff and related papers such as Jobin (2019) and Fjeld et al. (2020) identified as important in AI ethics are privacy, fairness, accountability, transparency, safety, sustainability, control, inclusion, explainability, diversity in the field of AI, and many others. Hagendorff notes that the aspects of accountability, privacy, and fairness appear in 80% of all examined guidelines and therefore represent the minimal requirement for building an ethically sound AI system.

Hagendorff criticizes that certain principles and issues are wrongfully omitted from the general discussions in AI ethics. Examples for such factors would be the political abuse of AI systems, the lack of diversity in the AI community and the contradiction of AI developments with sustainability goals. Moreover, the close link between science and the private sector poses a threat in terms of buy-outs. The ethicist raises concerns regarding the ongoing international race between the US, China and the EU when it comes to AI advances. He diagnoses this in- and outgroup thinking to be a serious danger, as it promotes recklessness and the likelihood of skipping ethical considerations in the process. When it comes to evaluating whether ethical concerns are effective in practice, Hagendorff raises harsh criticism. In his opinion, these normative ethical goals are rather weak and underachieved, as they can not be enforced in practice by binding legal frameworks. Instead, companies rely on internal self-commitments that are failing since they are overshadowed by decisions that are driven by economic motives. On a positive note, he mentions that for some principles such as privacy and fairness, progress is made in terms of technical fixes.

Considering the overall unfortunate state of AI ethics, the author suggests two solution approaches to address the challenges in the field. His first proposal is to provide supplementary technical instructions with ethical guidelines so that they can immediately be translated into action. This is meant to close the gap between abstract ethical values and technical design of AI systems. For this to have an effect, ethicists have to attain a thorough

understanding of data collection processes, algorithm design and many other aspects. Hagendorff calls for a movement to the micro-level of ethics. An example for such practices would be standardized data sheets proposed by Gebru et al. (2018). The second proposed approach is to transition from deontological ethics with a technology focus to social and personality-related (virtue) aspects that are situation-sensitive, look at individuals, and strengthen self-responsible actions. This approach involves considering the social implications of AI development, such as how it affects individuals and society as a whole. It emphasizes the importance of a moral character and self-responsibility in ensuring that ethical principles are followed throughout the development process.

The paper by Hagendorff was generally well received by the research community as an overview of ethical and responsible use of AI. It was referenced by following papers such as Christoforaki (2022) who noted that there are some values frequently missing in AI ethics guidelines and Ibáñez (2021) and Solanki (2022) who agreed with Hagendorff on the notion that a principle based approach to AI ethics is not sufficient. Since the publication of the paper in 2020, there have been developments in AI ethics both on the technical and the conceptual side. With recent developments in the field of generative AI such as the emergence of tools like ChatGPT, additional considerations are required to address pressing ethical issues. For instance, the use of copyrighted materials in training data sets and generated content may result in copyright infringement. Similarly, authorship of AI-generated content raises questions about the creator and their legal and moral responsibilities. It is also crucial for users to assess the truthfulness and accuracy of information generated by AI systems to prevent misleading content. Deepfakes and synthetic media pose a risk to make this assessment harder.

As technology continues to advance at a rapid pace, it is vital to develop additional guidelines and legislation to address these ethical issues surrounding AI. This points to the general need of AI ethics to incorporate these new developments into its ethical considerations. A recent conceptual development is the publication of the UNESCOs recommendation the Ethics of AI, which was accepted by 193 member states in 2021. The recommendation focuses on the broader ethical implications of AI systems and how they relate to human rights and the sustainable development goals. Key principles for ethical AI development among others are safety and security, fairness and non-discrimination, sustainability, transparency and explainability, and multi-stakeholder collaboration. This document is a conceptual milestone because it provides advice on specific policy action and has the potential to harmonize AI ethics across different countries. A recent legislative development is the introduction of the AI act of the European Union. It proposes a regulatory framework that classifies AI systems according to their risk and mandates development requirements. It is the first legislative piece to introduce legally binding rules instead of voluntary principles and self-commitments and in that way promotes the vision of sustainable and trustworthy AI.

For additional reading we propose papers by Jobin (2019) and Fjeld et al. (2020), which both map values of AI ethics through a systematic analysis and go into more detail on the concepts of the individual values compared to the original Hagendorff paper. Additionally, two follow-up papers by Hagendorff (2022a, 2022b) touch on the frequent blind spots in AI

ethics guidelines and propose ways to incorporate AI ethics values in practice through a virtue-based framework.

In conclusion, "The Ethics of AI Ethics" provides a critical analysis of the current state of AI ethics and highlights areas for improvement. At the time of its publication, Hagendorff contributed to the field of AI ethics by highlighting its weaknesses and providing solution approaches to increase its effectiveness. The recent developments targeting generative AI show that AI ethics is an ever-evolving field that continuously struggles to adapt to the reality of fast-paced technological advances.

References

Christoforaki, M., & Beyan, O. (2022). AI Ethics—A Bird's Eye View. *Applied Sciences*, *12*(9), 4130. https://doi.org/10.3390/app12094130

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3518482

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2018). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hagendorff, T. (2022a). A Virtue-Based Framework to Support Putting AI Ethics into Practice. *Philosophy & Technology*, *35*(3), 55. https://doi.org/10.1007/s13347-022-00553-z

Hagendorff, T. (2022b). Blind spots in AI ethics. *AI and Ethics*, *2*(4), 851–867. https://doi.org/10.1007/s43681-021-00122-8

Ibáñez, J. C., & Olmeda, M. V. (2022). Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI & SOCIETY*, *37*(4), 1663–1687. https://doi.org/10.1007/s00146-021-01267-0

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Solanki, P., Grundy, J., & Hussain, W. (2023). Operationalising ethics in artificial intelligence for healthcare: A framework for AI developers. *AI and Ethics*, *3*(1), 223–240. https://doi.org/10.1007/s43681-022-00195-z