RESEARCH ETHICS FRAMEWORKS FOR ARTIFICIAL INTELLIGENCE: THE TWOFOLD NEED FOR COMPLIANCE REQUIREMENTS AND FOR AN OPEN PROCESS OF REFLECTION AND ATTENTION

Anais Resseguier

Trilateral Research (Ireland)

Anais.resseguier@trilateralresearch.com

EXTENDED ABSTRACT

This paper proposes to enhance research ethics frameworks for research projects developing and/or using Artificial Intelligence (AI). It highlights that these frameworks need both (a) requirements for compliance with emerging ethical and legal norms to govern this technology and (b) an open process of reflection and attention to research and innovation in this area. The field of AI ethics has seen intense developments since 2015 with numerous governmental and international bodies, institutions and companies creating guidelines, frameworks, and sets of principles for AI governance. Jobin et al. (2019) have analysed 84 of these documents and point to significant convergence on key principles, in particular transparency, justice and fairness, non-maleficence, and responsibility. However, these initiatives have also received sharp critiques from experts in the field, including that of being a form of "ethics washing" (Wagner, 2018; Resseguier and Rodrigues, 2020) or of reproducing existing power structures and inequalities (D'Ignazio and Klein, 2020). The proposed approach seeks to address these critiques by focusing on review processes as handled by research ethics committees (RECs), also called institutional review boards (IRBs). As the AI ethics field is currently working toward its operationalisation, research ethics constitute a powerful, but so far underdeveloped framework to make AI ethics more effective at the level of research (Santy et al. 2021).

A two-pronged approach to the operationalisation of AI ethics

The present paper proposes a two-pronged approach to the operationalisation of AI ethics in research ethics frameworks: (a) compliance with requirements imposed on researchers and (b) an open process of attention and reflection. In the words of the philosopher George Canguilhem, while the former aspect of ethics is about engaging with the norms, the second one attends to the *capacity* to determine the norms, i.e., the "normative capacity" (Canguilhem, 1991). Before presenting what this means concretely for AI research ethics (section 2), this paper makes a detour by the theory of ethics (section 1). It does so by drawing from works by Gertrude E.M Anscombe (1958) and Charles Mills (2005) that help provide conceptual clarity on the notion of ethics used primarily in AI ethics since around 2015 and ways to avoid critical pitfalls of this approach. It shows indeed how a clarification between the level of the norms (a) and that of the open process of reflection and attention (b), i.e., the "normative capacity" in Canguilhem's terms, helps to lift a confusion in AI ethics, a confusion that has weakened its potential effectiveness and has led to a number of legitimate critiques.

Compliance requirements and the potential role of the European AI Act

In the second section, this paper formulates a series of concrete recommendations for AI research ethics, based on the conceptual framework identified in the first section. To begin with, this paper encourages the imposition of particular requirements within research ethics frameworks embedded in institutions. This corresponds to the side of the norms requiring compliance (a) as identified in the model described in the first section. These norms, principles, or requirements, should be accompanied by mechanisms to ensure compliance, such as through the possibility of withdrawing funding if these are not fulfilled (this is for instance the case with the ethics appraisal scheme for research projects funded under the Horizon Europe Funding Program of the European Commission). Requiring compliance with certain criteria allows to put red lines and better orient AI research in a way that avoids potential harms caused by this technology, such as mass surveillance or discrimination. In this sense, research ethics takes the shape of "soft law" requiring compliance with certain obligations. This would help address the critique AI ethics has received of being "toothless", a form of "ethics washing", due to the absence of enforcement mechanisms.

Requirements from the European Union's AI Act currently under development will assuredly constitute a key reference for research ethics norms. Although, in the current form of the draft (as of June 2023), the obligations of the AI Act do not apply to scientific research, it is most likely that these obligations will nonetheless have a strong impact on AI research considering the need to anticipate placement on the market or to test in real world conditions (European Parliament, 2023). This paper explores the implications of the AI Act for research ethics frameworks and especially what the legal obligations in this regulation will mean for research ethics requirements and mechanisms to ensure compliance with these. In particular, it will investigate what the risk-based approach in the AI Act implies for research ethics and how to ensure compliance with the obligations at the different risk levels.

An open process of reflection and attention

In addition, ethics review frameworks offer a space for an open process of reflection and attention (b). The focus here is on questioning established norms and ways of doing through an open reflection and a continuously renewed form of attention to both technical advances in the field and social developments and concerns. This corresponds to the level of the "normative capacity", to use Canguilhem's terms as defined in the first section, i.e., the capacity to pay attention to the new situation, reflect on it, and challenge existing norms if needed to best adapt to the novelty one faces. Considering the uncertainty AI brings to societies, this constantly renewed attention and reflection is essential. For instance, in-depth critical social science and humanity (SSH) studies are crucial to engage such open reflection and renewed attention (e.g., Crawford, 2021). The submission of a societal impacts statement as part of an ethics submission for AI research projects can serve to embed such reflection within the ethics review process (Bernstein et al., 2021; Ada Lovelace Institute, 2022). Another option would be to carry out discussions with an expert on the ethical and social impacts of the AI system under development at several stages of the research project development. Strengthening the open process of reflection and attention at the research ethics level would help address the critique made toward AI ethics according to which it would fail to address structures of power and inequalities.

By distinguishing the level of the norms and that of the open process of attention and reflection, highlighting their respective values, and the way they relate to each other, this paper contributes

to advancing further ai ethics through its operationalisation in research ethics frameworks. The aim is eventually to make ai ethics more effective but also more thoughtful.

KEYWORDS: AI ethics; Research ethics review; AI Act; Ethics washing; Research ethics committees.

REFERENCES

Ada Lovelace Institute. (2022, Dec). Looking before We Leap. Expanding Ethical Review Processes for AI and Data Science Research. Retrieved from https://www.adalovelaceinstitute.org/report/looking-before-we-leap/

Bernstein, M. S., Levi, M., Magnus, D., Rajala, B. A., Satz, D., Waeiss, Q. (2021 Dec). Ethics and Society Review: Ethics Reflection as a Precondition to Research Funding. *Proceedings of the National Academy of Sciences* 118(52).

Canguilhem, G. (1991). *The Normal and the Pathological*. Translated by Carolyn R. Fawcett. Princeton: Princeton University Press.

Crawford, K. (2021) *Atlas of AI*. New Haven & London: Yale University Press.

D'Ignazio, C., Klein L. F. (2020). *Data Feminism,* Cambridge, MA; London, England: MIT Press.

European Parliament (2023, June), Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)) https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

Jobin, A., Ienca M., Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1(9), 389–99.

Mills, C. (2005). "Ideal Theory" as an Ideology. *Hypatia*, 20(3), 165–84.

Santy, S., Rani, A., & Choudhury, M. (2021). Use of Formal Ethical Reviews in NLP Literature: Historical Trends and Current Practices. *CoRR*, *abs/2106.01105*.

Rességuier, A., Rodrigues, R. (2020). AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics. *Big Data & Society* 7(2).

Wagner, B. (2018). Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping. In *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, ed. Emre Bayamlioglu et al., Amsterdam: Amsterdam University Press.