Situated Accountability: Ethical Principles, Certification Standards and Explanation Methods in Applied AI

Anne Henriksen, Simon Enni and Anja Bechmann

Aarhus University, Denmark

ABSTRACT

Artificial intelligence (AI) has the potential to benefit humans and society by its employment in important sectors. However, the risks of negative consequences have underscored the importance of accountability for AI systems, their outcomes, and the users of such systems. In recent years, various accountability mechanisms have been put forward in pursuit of a responsible design, development, and use of AI. In this article, we provide an in-depth study of three such mechanisms, as we analyze Scandinavian AI developers' encounter with (1) ethical principles; (2) certification standards, and; (3) explanation methods.

By doing so, we contribute to closing a gap in the literature between discussions of accountability on the research and policy level, and accountability as a responsibility put on the shoulders of developers in practice.

Our study illustrates important flaws in the current enactment of accountability as an ethical and social value which, if left unchecked, risks undermining the pursuit of responsible AI. By bringing attention to these flaws, the article signals where further work is needed in order to build effective accountability systems for AI.

OBJECTIVES

The research questions guiding the study are:

How are ethical principles, certification standards, and explanation methods enacted? How are they responded to and reflected on by developers in applied AI? To what extent do these mechanisms promote accountability and the use of responsible approaches in design and development processes?

The article aims to provide situated bottom-up perspectives on accountability and responsible AI, and, consequently, on the governance of AI and responsible innovation.

Through our case study-based analysis and discussion, we aim to bridge the gap between accountability as discussed at the policy and research level, and accountability as a responsibility put on the shoulders of engineers working with AI in practice.

THEORY

Accountability as a mechanism (Bovens 2010) involves a relationship between an actor and a forum with expectations for (1) what kind of formal or informal account the actor should give in order to justify its conduct; (2) how and by whom (which forum) the actor giving an account should be questioned and passed judgment on with regards to the adequacy of the account, or the legitimacy of the actor's conduct, and; (3) which consequences are mandated in case of a negative judgment. Some would consider the judgement by the forum, or even just the justification by the actor, to be enough to qualify a relation as an accountability mechanism.

By drawing on Bovens (2010), we understand ethical principles, certification standards, and explanation methods as mechanisms intended to facilitate accountability in socio-technical systems (cf. Ananny and Crawford 2018).

METHOD

The empirical data underlying the article was collected on the basis of an *ethnographic case study* (Yin 1994; Davies 2008).

- Duration: late 2018 until early 2020

A *follow-the-actors* approach (Latour 1985) was used for studying how developers at an AI company in Scandinavia *practiced* AI design and development.

Methods & data:

- *Participant observation* during the whole period including 6 months of day-to-day observations involving on-the-spot interviews
- *Semi-structured interviews* (N: +20) with managers, data scientists, data modelers etc.
- *Analysis of documents*, e.g. project descriptions describing the AI techniques used

Both ethical principles, certification standards, and explanation methods emerged as analytical themes from the empirical material. This material was analyzed by means of an iterative process, open to the themes emerging from the empirical data and yet informed by our research interest.

RESULTS

Ethical Principles

The developers felt that AI ethical guidelines were largely irrelevant to the type of work they were doing and, in fact, somewhat harmful to their business. Still, they obviously felt targeted by them.

> –The general misunderstanding of what AI is has really surprised me. Really, AI is just 'statistics on speed' and nothing more than that. I don't understand why people question what AI is but don't question, for example, what MRI [Magnetic Resonance Imaging] is, because, in my opinion, MRI is just as unstable as an ML algorithm may be. It's not that I am against legislation but I just think the general discussion is too generalizing and stereotypical, and is missing the point. In fact, I think it is damaging to the work that we're *actually* doing. (Engineer, Feb. 2020)

Certification Standards

The developers were motivated to demonstrate their accountability and integrity through conformity to ISO standards. However, they were met by little to no guidance on how to conform to standards as a supplier of AI for healthcare, and a certification process unfit to deal with AI systems in depth.

> –We are about to apply for the certification in ISO 13485 on medical devices but there is absolutely *nothing* for us to follow in order to implement the standard. Our best bet is some FDA guidelines from the US; we're not even ready in the EU yet! Seriously, wake up, please!...I've talked to the national medicines agency that has to handle these things [provide guidance] but they knew *nothing*...The agency has announced that it will develop some new guidelines as if all of this was *completely* new, whereas I'm just thinking: "Stop, please, and just look at the papers from the FDA". (Director, Feb. 2020)

Explanation Methods

The developers learned that explanations generated with xAI methods potentially could counteract usability and might discard information that would be otherwise important in the evaluation of a patient's condition.

> –From the beginning, explanation has been foregrounded as something that ought to be given at the level of a user interface: "Ohh, it's a black box! This means we cannot use AI for anything *at all*!" That's from the perspective of a doctor, you know: "I have to know what the reasons are" and so on. However, I don't believe this will be necessary because AI is not going to be applied like: "Does this guy have cancer or not?" Rather, I believe algorithms will be used for eliminating parts of working processes and triggering actions. (Director, Feb. 2020)

DISCUSSION & CONCLUSION

Given that AI developers play a major role in ensuring accountability for AI systems, their outcomes, and the users of such systems, we need them to pursue accountable, ethical, and responsible approaches. Therefore, we suggest that these actors are involved in the policymaking processes aimed at generating a responsible design and use of AI. Based on our empirical findings, we have present several recommendations to remedy the flaws identified in the current ways that ethical principles, certification standards, and explanation methods are enacted as mechanisms of accountability in pursuit of responsible AI. Our hope is that these recommendations may contribute to the discussion of how accountability is ensured in practice in a way that accounts for the perspectives of both developers, researchers, and the general public.

REFERENCES

Ananny, M.; and Crawford, K. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *Big Data & Society* 20(3): 973-989. doi.org/10.1177/1461444816676645

Bovens, M. 2010. Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. *West European Politics* 33(5): 946–67. doi.org/10.1080/01402382.2010.486119.

Davies, C. A. 2008. *Reflexive Ethnography – A Guide to Researching Selves and Others.* London & New York: Routledge.

Dignum, V. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Cham, Switzerland: Springer. doi.org/10.1007/978-3-030-30371-6.

Yin, R. K. 2014. *Case Study Research: Design and Methods*, 5th Edition. Los Angeles, LA, USA: Sage Publications.

ACKNOWLEDGEMENTS & CONTACT

Thanks to the director and employees at the AI company for participating in interviews and allowing for participant observation. The work has been funded by Aarhus University and Aarhus University Research Foundation for funding this work.

Corresponding authors:

annehenriksen@cc.au.dk - enni@cc.au.dk