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

Anne Henriksen
DATALAB
Aarhus University
Aarhus, Denmark
annehenriksen@cc.au.dk

Simon Enni
Department of Computer Science
Aarhus University
Aarhus, Denmark
enni@cs.au.dk

Anja Bechmann
DATALAB
Aarhus University
Aarhus, Denmark
anjabechmann@cc.au.dk

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 the 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.

CCS CONCEPTS

• Social and professional topics~Professional topics~Computing and business~Socio-technical systems • Social and professional topics~Professional topics~Computing and business~Computer supported cooperative work • Social and professional topics~Professional topics~Computing and business~Automation • Social and professional topics~Computing / technology policy • Social and professional topics~Professional topics~Management of computing and information systems~System management~Technology audits **• Social and professional topics~Professional topics~Management of computing and information systems~Project and people management~Systems development**

AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05...$15.00.
https://doi.org/10.1145/3461702.3462564

KEYWORDS: AI, Machine learning, Algorithmic systems, Accountability, Responsible AI, AI ethics, Certification, Explainable AI, Case Study, Ethnography

ACM Reference format:
Anne Henriksen, Simon Enni, and Anja Bechmann. 2021. Situated Accountability: Ethical Principles, Certification Standards, and Explanation Methods in Applied AI. In *Proceedings of the 2021 AAAI/ACM conference on AI, Ethics, and Society (AIES'21), May 19-21, 2021.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3461702.3462564

1 Introduction

Perhaps the greatest example of artificial intelligence (AI) put to use for people and society is AI techniques applied to healthcare, with promises of better, cheaper, and more efficient care for patients. In fact, such techniques, which draw on machine learning (ML)[1] models applied to big data, have already shown advances in a number of areas, e.g. image classification [34, 45, 57], medical prognosis [28, 55], and automated processing of Electronic Health Records [84]. Yet, the incorporation of automated algorithmic systems into critical infrastructures of society does not come without concerns. In recent years, studies have documented cases in which AI systems have had adverse effects on people subjected to them. Such cases include discriminatory services [15, 92], capricious and unfair social decision systems [35, 76], and runaway feedback loops in criminal justice [33]. Additional perils have been identified specifically in healthcare such as misleading prognoses and unsafe treatment [21], and an inability to handle uncertainty and ambiguity of medical data [18]. Also risks of deskilling and a decreased diagnostic accuracy have been documented [19].

The discovery of these potential consequences has underscored the importance of accountability for AI systems, their outcomes, and the users of such systems *before* and *after* they have been developed and deployed. Arguably, such accountability is intrinsically linked to the responsible design and use of AI [31, 58, 59]. As Smith [87] argues, "one of the ways that responsible actors demonstrate their responsibility is by being accountable". In consequence, different *mechanisms for accountability* [8] have been put forward in pursuit of so-called 'responsible AI'. For instance, political and professional organizations have issued guidelines prescribing ethical principles for the design and use of AI systems [54]. Additionally, researchers have come up with various theoret-

[1] For an overview of ML's position as a subfield of AI, see [82].

ical and technical approaches to accountable AI algorithms and computer systems (see e.g. [29, 60, 65]).

In this article, we study how accountability and consequently responsible AI is being pursued and practiced, focusing on the three accountability mechanisms of (1) ethical principles,[2] (2) certification standards, and (3) explanation methods. Hence, we adopt a view of these three as mechanisms intended to facilitate accountability in socio-technical systems incorporating AI systems [3, 8]. These mechanisms are not legally binding per se and do not have any formal legal status in themselves. Hence, from a legal perspective, they may be considered as 'soft law' rather than 'hard law' [1]. Generally, soft law is defined as "written and unwritten instruments and influences that shape administrative decision-making" [88]. In the literature, this type of law is highlighted as an appropriate tool for governance on an international level, which "often has much more influence than legislative standards" have in practice [88]. Yet, much criticism has been raised against the use of soft law approaches to the governance of AI. Recent studies have criticized especially ethical guidelines for being "toothless" [79] and called for research studying the use and effect of such guidelines closer to the realm of applied AI (see e.g. [70]). Such criticism underlines the general need to study mechanisms, which we expect to facilitate more substantiate accountability and promote responsible AI, *in context.*

This article contributes to bridging this gap through an empirical study of Scandinavian AI developers' encounter with the three accountability mechanisms outlined. These mechanisms were explicitly discussed in the research case which we draw upon. They are furthermore widely discussed by researchers and policymakers in the debate on how to ensure accountability in and after AI development processes, and the responsible design, development, and use of AI. 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 in and after design and development processes, and the use of responsible approaches during such processes?*

To study these questions we, firstly, expound the three accountability mechanisms studied from a theoretical perspective and elaborate on the distinct problems that AI poses to accountability. In doing so, we draw on theory and literature from related fields such as governance and public administration, philosophy of information, and ethics and information technology. Secondly, we present the central case study and argue for the methods used to collect and analyze the data underlying the study. Next, we present our empirical findings. Finally, we conclude with a critical discussion of the issues identified in our analysis, and provide our concluding remarks and recommendations.

The article aims to provide situated bottom-up perspectives on accountability and responsible AI and, consequently, on the governance of AI [10, 31, 94]. 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 on the development of AI systems in practice.

[2] In this paper, we use 'ethical principles' as a collective name for ethical principles, codes, and guidelines.

2 Accountability Mechanisms

As noted by many scholars, accountability is a multifaceted concept with a long history and many applications [86]. Mark Bovens distinguishes between two different usages of the term; as a *virtue* and as a *mechanism* [8]. Despite this distinction, the two concepts are "closely related and mutually reinforcing" [8].

When used as a virtue, accountability is regarded as "virtuous behavior" which organizations should strive for, i.e. "a willingness to act in a transparent, fair, compliant, and equitable way" [8]. In this sense, accountability refers to "substantive norms for the behavior of actors" [8]. Meanwhile, it is in the sense of a social, political, or administrative mechanism that accountability has its historical and semantic roots. Staying close to these roots, accountability can be understood as a mechanism that involves "an obligation to explain and justify conduct" [8]. Today, however, it is the *relation* through which "an agent can be held to account by another agent or institution" that is the crux of the concept [9]. This relation can be more or less formal or informal by involving consequences that are either formalized or based on unwritten rules [8]. Briefly explained, accountability as a mechanism 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 judgement [8, 9].

As noted by Bovens [8], 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. In this way, we may understand accountability mechanisms to form a *web* of different interrelated mechanisms intended to enable accountability. In line with this idea, researchers and policymakers consider ethical guidelines to have an effect on AI stakeholders' conduct *only in interaction with* the legislations and regulations of a country or union (see e.g. [63]). We acknowledge this observation and understand the three accountability mechanisms in our study to be working in conjunction with other mechanisms at different levels of government, including regulatory and legislative instruments. Yet, we will leave a specific treatment of such instruments for future research and focus on the three mechanisms concerned.

Recent studies have shown how accountability is complicated by the incorporation of AI systems into the decision-making processes of (public) institutions, as AI development firms thus become a part of the accountability relationship between the institution and its customers or citizens [68, 101]. However, most attention within research on accountability in relation to AI is given to difficulties with producing effectful accounts of the functioning of AI systems and the decisions made with such systems (see e.g. [60, 58, 101]). Such difficulties were already pointed to in 1994 by Helen Nissenbaum [74] who called for accountability in the use of complex computer systems in critical sectors. Yet, the situation

now has been further complicated by issues unique to modern AI systems.

In particular, systems created with ML models introduce difficult barriers to accountability outside of what could normally be expected for complex computer systems. Especially the *opacity* of complex ML models creates problems for the production of accounts of AI systems and their outcomes [16, 32, 98]. This opacity is the result of ML models being "learned" from data rather than written by human programmers. Such data-driven learning often leads to the production of highly complex and large mathematical models which cannot easily be 'picked apart' into meaningful units and inspected individually [16, 85]. Since such models end up appearing as 'black boxes' to outside human inspectors, it can become almost impossible to distinguish between and distribute responsibility for the many different kinds of errors and mistakes that the models can make [16]. Furthermore, since many ML models continually learn and adapt to new input data when applied in practice, it may not even be possible to investigate the particular offending instantiation of the model in a post-hoc fashion if problems occur [83, 105]. Yet, AI systems are always "fit for certain uses" [58] and thus rely on design choices, values, and underlying goals that are important for the resulting (social) outcomes, and for which the developers, owners, and users of the systems can be held accountable [10, 31].

Given the different problems and risks arising in the light of modern AI systems, it is important that developers are held accountable for their innovations; provide accountability for the users of their AI systems including the people subjected to them; and proceed with caution and respect for ethical and social values when designing and developing such systems. For these purposes, different mechanisms for accountability have been, and are, discussed by researchers and policymakers. This includes ethical principles, certification standards, and explanation methods. We analyze these three mechanisms individually in order to provide focused insights.

2.1 Ethical Principles

Ethical principles, or applied ethics, prescribe the norms for what is socially and ethically acceptable and preferable [36, 37]. In this way, ethical principles work to "steer society in the right direction" by defining what ought to be done and what ought to be avoided, thereby outlining the conduct which actors should strive for *beyond* mere legal compliance [37]. In recent years, a number of guidelines prescribing ethical principles for the design and use of AI have been issued by professional and political organizations. These are, for instance, the FAT-ML community [30], AI4People forum [39], OECD [75], IEEE [23], and the High-Level Group on AI appointed by the European Commission [4]. Common to the various guidelines is that they seek to guide the application of AI towards uses that generally benefit society and contribute to the common good. They do so by pointing to the principles and intrinsic values to pursue, and the ethical risks and social harms to avoid [20, 94]. In a systematic literature review of ethical guidelines, Jobin, Ienca, and Vayena [54] found a convergence towards principles of transparency, justice and fairness, non-maleficence, responsibility, and privacy. However, there was disagreement on how these principles should be interpreted and implemented.

Some guidelines include suggestions or requirements for how principles and the associated values are implemented, thereby assuming an operational dimension (see e.g. [4]). Yet, as suggested earlier, the main strength of ethical principles as mechanisms for accountability presumably lies in their *normative forces* [37]. By prescribing norms for the behavior of actors towards positive societal impacts, ethical principles place expectations on actors, not only for which acts they ought to perform but also for which obligations they have a responsibility to fulfill [12, 36, 37]. It is in this way that ethical principles themselves may function as accountability mechanisms, namely by establishing expectations for the conduct and obligations of actors, and by establishing a public (the society) to which the actors are accountable [88]. If AI firms do not fulfill these expectations, or explicitly choose *not* to adopt otherwise universally agreed-upon ethical principles, they may suffer reputational losses, resulting in severe consequences for the credibility of their businesses [47, 104].

Despite the presumed 'power' of AI principles to exert pressure on stakeholders to strive for building a so-called "Good AI society" [39], ethical principles have been harshly criticized for being too 'weak' an accountability mechanism, failing to have an actual effect on AI stakeholders' conduct. In particular, ethical principles are criticized for lacking oversight processes and sanctions, thereby making the "critical audience" [56] ill-equipped to assess the conducts of those developing and employing AI systems. In this way, ethical principles expect AI stakeholders to be able to interpret what it means to uphold and practice ethical principles in the context of their own work [11, 47, 70, 73, 99, 100].

2.2 Certification Standards

A certification is used for attesting whether an object of certification meets the requirements of a standard on the basis of an audit. Hence, certifications, standards, and audits are inextricably linked [77]. Historically, certification standards have been used in manufacturing sectors and developed by, among others, non-governmental standards setting organizations [13, 17, 26]. Standards concerning AI design and development are currently under development by several standards organizations, including the International Organization for Standardization (ISO).[3] Such standards are considered to offer a method for formalizing AI ethical principles and substantiating their implementation in practice, in this way incentivizing their adoption [102, 24].

Certification standards are described as "consensus-based agreed-upon ways of doing things, setting out how things should be done" [14]. Additionally, they are defined as "technical specifications and other precise criteria, which ensures that materials, processes, services, systems, or persons are fit for their intended purpose" [51]. Usually, it is an authorized independent body from the private sector, a trusted third party, that assesses whether the object of certification controlled by the applicant meets the specified criteria and builds on the best practices of the standard [7, 11]. In this way, it is certification bodies which "attest to the broader

[3] https://www.iso.org/home.html

public that an AI system is transparent, accountable, and fair" [4]. Certification standards thus function in a very direct way as a mechanism for accountability, since they establish a clear accountability relationship between the actor seeking the certification, and the certification body to whom this actor must be accountable as a proxy for the customer and the wider public. A formal account is given, and the consequences of non-conformity to the criteria and best practices of the standard is the denial or revocation of the certification. Such denial or revocation may result in reputational losses or even more severe consequences, if the certification is a mandatory regulatory requirement. In this latter case certification standards may be interpreted as 'hard law' [102, 13, 24].

The use of certification standards as accountability mechanisms in AI has been criticized for potentially promoting an instrumentalization of ethical values without necessarily eliminating irresponsible and unethical behavior [38, 99]. Additionally, scholars have stressed the limitations of the commonly used post-hoc audit when applied to AI systems [83, 105]. This has led to recommendations for how audits should be performed specifically in relation to AI systems. For example, Andrew Tutt [96] has recommended that an entirely new regulatory body is instituted to oversee the quality and compliance of automated algorithmic systems, inspired by similar bodies in other fields such as the American Food and Drug Administration (FDA). In contrast, researchers at Google have presented an end-to-end framework for conducting internal audits during the AI development process [78]. Hence, they suggest that AI developers themselves should play a larger role in the audit process.

2.3 Explanation Methods

As previously mentioned, the inherent opacity of AI systems using ML models naturally impedes accountability in the sense of "providing answers for your behavior" [9]. This has made researchers and policy bodies point to the need for explanations, or explainability, in order to provide accounts alongside predictions of AI systems (see e.g. [4, 11]). Such explanations were spurred on by the GDPR, which pushes for "a right to explanation" but without legally mandating it [43, 97]. The field of explainable AI (xAI) has recently seen a resurgence of interest as a means to face this challenge by enabling more substantial transparency and thereby accountability [46, 71].

In xAI, ML models deliver explanations alongside predictions in order to compensate for the lack of transparency and understanding of their inner workings. The explanations generated by the use of xAI methods are solutions to certain optimization problems, and many of the differences in xAI methods are the result of different assumptions and formulations of these optimization problems. Common variations include a focus on explaining a single outcome of the model (local explanation) [80] versus explaining the general behavior of the model (global explanation) [71], and optimizing the accuracy and predictive power of 'inherently transparent' models [81] versus generating post-hoc explanations for the behavior of a given black box model [44].

xAI is often envisioned as a mechanism for accountability that mediates the relationship between the AI system and its uses by allowing the system *itself* to account for its behavior through automatically generated explanations that the user can query and question [46, 65]. In this accountability relationship, the accounts given are judged by single individuals, be they doctors, researchers, or laypeople. Rather than referring to external standards and guidelines, each recipient of an explanation evaluates the extent to which it is a valid account of the AI system and its outcomes [32, 46, 80, 98]. Thus, explanation methods center individual actors in the accountability relationship rather than certification bodies or the wider public and lack, like ethical principles, explicit consequences in cases of misconduct. In turn, explanation methods are sometimes highlighted as ways of enabling other types of consequences, such as by revealing breaches of legal or professional standards [32].

While the xAI field has seen a surge of popularity recently, the aims and methods of the field have also been the target of extensive criticism [81]. The first part of this critique is technical. Explanations are found to be statistically fragile [42] and oversensitive to spurious correlations [6]. Furthermore, they can be manipulated to deceive users [62]. The other part of the critique is conceptual. Adrian Weller has, for instance, criticized the idea that transparency will solve the problems with using ML for decision-making [98]. Similarly, Zachary C. Lipton has challenged the assumption that explanations in the form of simplified model approximations will even improve transparency in ML in the first place [66]. In spite of this criticism, explainability remains important for accountability in AI, whether it is realized through xAI or some other method. It is given a central role, not only in the GDPR but also in a number of ethical guidelines (see e.g. [4, 30, 39, 75]).

3 Case & Method

Both ethical principles, certification standards, and explanation methods emerged as analytical themes from the empirical material underlying the article. This material was collected on the basis of an extensive ethnographic case study [103, 25], conducted by the first author from late 2018 until early 2020. Using a *follow-the-actors* approach [64], the author studied how the developers at an AI company in Scandinavia *practiced* AI design and development. In this way, the case study resembles earlier empirical studies focusing on the practices surrounding AI development (see e.g. [2, 40, 91, 53]). The study is unique especially by virtue of its accounts of modern AI development in a commercial company.

The first author had asked for permission to study the developers' work close up with regard to a predictive system for healthcare developed within a research and innovation project. But as this work involved various practices and considerations that linked to other work and AI projects in the firm, the data collection came to encompass, for instance, developers' work on the development of explanation methods and attainment of ISO certifications. Similarly, it came to include developers' occasional discussions of AI ethics in general and the AI HLEG ethics guidelines [4] specifically, which were issued during the data collection period.

Various different ethnographic methods were used for collecting the data in order to generate *thick descriptions* [41]. Participant

observation [89] was conducted at numerous meetings and workshops held in relation to the AI research and innovation project during the entire period. Furthermore, the first author stayed with the company on an everyday basis from March to August 2019, where she performed several spontaneous on-the-spot interviews with developers and conducted day-to-day observations of their everyday work. As the first author had agreed to assist the team developing the predictive system for healthcare, she made the observations and spontaneous interviews in the role as a *participant observer* [25]. Additionally, she conducted more than 20 semi-structured interviews [61] with e.g. managers, business developers, data scientists, and data modelers. These interviews were conducted during two periods: August-September 2019 and January-February 2020. Semi-structured interviews were used for producing more details on aims, and observed actions and statements. They furthermore delved into the developers' roles, experiences, and views. In line with the 'follow the actors'-approach, the semi-structured interviews were conducted in an exploratory manner based on a set of loosely structured questions [61] with a view to cover key themes relevant to the exploration of AI design and development practice. Interviews and field notes were additionally supplemented with *documents* collected in the field, e.g. project descriptions [103].

All the data collected and used in this article have been transcribed and subjected to a prolonged analysis process. More specifically, the data have been analyzed by means of initial categorization based on "the participants' voice" resulting in preliminary themes and topics [67], repeated readings to generate more condensed meaning units [25], and simultaneous writing and thinking to produce more well-found interpretations of the data [27, 90]. Hence, the way from data to findings has been a highly iterative process, open to the themes emerging from the empirical data and yet informed by our research interest, i.e. to understand how accountability and responsible AI is pursued and practiced. Through this process, our findings of the study have constantly been analyzed and questioned. It should be noted that the quotes stated in the findings section have been translated into English from the original language.

At the time when the data were collected, the AI company was 10 years old and had approximately 30 employees, primarily engineers. However, the company had recently also engaged profiles from non-technical disciplines, including an anthropologist. This was in order to adopt a more user-centered development approach for the purpose of developing high quality AI products. Whereas the company had originally provided data consultancy for public healthcare institutions, it had recently started to move further into AI product development. The aim of the company was to utilize the extensive centralized health data records, which the Scandinavian countries are known for keeping [95], to improve the delivery of public healthcare. At the time of the study, the developers' focus was on using deep neural networks [82] to improve, among other things, medical diagnostics.

With our ethnographic accounts, we strive to elucidate the dynamics, conflicts, and complexities that the enactment of accountability and responsible AI involve in practice [48]. We believe that the results in the following section and the patterns discovered within the research case may provide important lessons, and serve as a background for future in-depth qualitative case studies and discussions of accountability and responsible AI.

4 Results

In this section, we report on the empirical data underlying the study as we analyze AI developers' encounter with the three accountability mechanisms outlined in the previous sections.

4.1 Ethical Principles

It is clear from the empirical material that the developers' reactions to ethical principles were highly negative. This is not to say that they did not act according to such ethical principles, however, our data show that they chiefly did so on their own initiative. As an example of such an initiative, the developers had worked on ensuring the strict traceability of their development process by constructing an advanced log system, as is also recommended in policy documents (see e.g. [4]). Rather, the reason behind their negative response to ethical principles had to do with a frustration with the *extent* to which and the *way* in which AI ethics were discussed in guidelines and the public debate. This frustration was indicated by the multiple times that developers talked of the discussion of ethical principles and issues as being irrelevant or out of proportions.

The developers indicated that the focus on AI ethics was excessive, primarily for two reasons. Firstly, they believed that being accountable and responsible was an *essential* premise of their business, if they should hope to have the healthcare sector adopt their AI systems. Secondly, they did not consider ethics and, more specifically, ethical principles to be genuine responses to the *actual* problems that they were facing 'on the ground'. Especially when discussions centered on moral dilemmas, they found them to be irrelevant to their work with developing AI systems. For these reasons, the focus on AI ethics which they considered to be extensive came to seem almost like a provocation, as this quote suggests:

> Believe me, we know that we have to be utterly impeccable…Here in the EU we have been sleeping on the job, and now we suddenly have to take up ethics. We need to *move away* from that kind of rhetoric! Really, nobody will say that they are not ethical. A philosopher raises one hundred questions but has no answers. (Director, Aug. 2020)

In particular, the developers were provoked by the *AI Trustworthy Guidelines* [4] issued by the AI HLEG which they considered to be a waste of time and suspected was made to cover up the fact that the EU was lagging behind in the 'AI race' [47] and, more specifically, in devising legislation:

> I think it's extremely frustrating that they [the national and European authorities] release all kinds of things [i.e. ethical guidelines] and make all kinds of statements instead of just rolling up their sleeves! Seriously, we [i.e. the EU] cannot allow ourselves to just do *nothing*! Really, we do *everything* that we can because we know that

> our business *cannot* survive if it's compromised by any of this. (Director, Feb. 2020)

Furthermore, the developers expressed several times that the discourse surrounding AI ethics was largely misguided. In their opinion, this led the expectations for and demands placed on AI to be excessive compared to other technologies and human beings themselves:

> 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)

We find the reason for this view to be caused by diverging and conflicting notions of AI, prompting a misalignment between the discussions of AI ethics and the problems faced by the developers working with developing AI systems in practice. Generally, the developers understood AI as merely one tool and knowledge production method among others which ought not to be used in isolation. Clearly, they did not consider this understanding to be reflected in the general and more abstract discussions of AI ethics.

The developers generally found the ethical principles to be of little benefit to their work because they did not reflect the realm of their work on applications of AI for healthcare closely enough. This observation has similarly been stated in other studies (see e.g. [70]). Instead, the developers stressed in their critique of ethical principles how important a 'level playing field' [37] with respect to legislation was for the ability of smaller companies like theirs to compete effectively. Thereby, they worried that if discussions and principles hinder or delay (changes in) legislation, it might be severely damaging to smaller AI companies.

4.2 Certification Standards

In the case study, we found that the developers were highly motivated to apply for internationally recognized certifications from the International Organization of Standardization (ISO). During the period of data collection, the company underwent an ISO/IEC 27001 certification and was furthermore preparing for an ISO 13485 certification. Whereas the latter is required in order to obtain a CE mark, which is mandatory for companies wanting to market medical devices in Europe, the former is optional as a means of documenting the information security management of an organization [93, 49, 50]. Yet, the ISO/IEC 27001 certificate was described by the company director as one of the most prestigious certificates that the company could achieve, and is by ISO itself highlighted as one of their "popular standards" [52]. Despite the clear motivation for obtaining the ISO certifications, our analysis reveals some important problems with the use of certification standards as mechanisms for accountability.

First of all, we observed that the top management viewed the extensive preparatory work which employees needed to engage in as something that had to be 'done', so they could 'get back to work'. This suggests that the certifications merely served as seals of approval to the AI company, and that the best practices prescribed by the ISO/IEC 27001 standard were not considered to affect the actual work on developing AI systems in any significant way. Furthermore, and perhaps even more importantly, our data suggest that while the certification system was immensely resource demanding for the developers to navigate, they did not experience that their great preparatory efforts were reciprocated by the public and regulatory authorities. For one thing, they found that there was no guidance on how they as suppliers of AI systems for healthcare could comply with standards, not even in the case of the ISO 13485 required in order to obtain a CE mark:

> 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)

Whereas the developers felt they were doing everything they could in order to meet the expectations for them as suppliers of AI products for healthcare, they clearly did not think this was the case for the authorities involved in the certification process. The fact that they were met by an, in their opinion, underfunded and deprioritized certification body when certified in the ISO/IEC 27001 confirmed their view on this:

> There's a great lack of people who are able to certify others; there are simply not enough people with the right competences [with regards to understanding data modelling]. They [certification bodies] are faced with a gigantic readjustment as they fundamentally do not understand agile development, but they have accepted that this is how things work...Everything has been based on assumptions like: "We have this requirement specification, we are developing this product, and we test it in this way"...But we need to remember that when we're working with data products, it's not only the *way* you work that is agile but also the *basis* that you are working on [i.e., the data]. This is *not* taken into account in the audits. (Director, Feb. 2020)

So, while the developers were very motivated to apply for the ISO certifications, they were frustrated by the poor quality of the guidance that they received, compared to the immense effort required of them, and disappointed by the level of the certification process. It was clear that based on these experiences, the credibility of the certification system had degraded in their eyes. This may have pushed them further towards a shallow and *pro-forma* adher-

ence to the standards rather than substantially incorporating them into their work.

4.3 Explanation Methods

In the beginning of the data collection period, the developers were committed to providing explanations for models and their outcomes to healthcare practitioners through xAI. Specifically, the developers wanted to apply local explanation methods to the temporal convolutional network they had developed for predictive and diagnostic purposes. After prolonged experimentation, they finally settled on Layer-wise Relevance Propagation [5] with deep Taylor decomposition [72]. The manager stated that he partly felt the GDPR pushed them to provide explanations, in this way enabling practitioners to account for decisions and actions to patients. Yet, in practice, the motivation rather was to build 'good' and useful AI products, and make AI systems intelligible to practitioners, thereby promoting their trustworthiness and adoption. However, the developers' commitment to providing explanations for users decreased over time, and so did their reliance on xAI for this purpose.

For one thing, they learned through observation conducted in clinical settings that providing explanations via the user interface of AI systems was at the risk of counteracting the efficiency gained from using AI in the first place, as users were given additional information to process. This change in attitude coincided with a change of focus from developing decision support systems to instead developing what they termed "microservices". Such microservices were meant to ease clinical work by, for instance, automatically starting the medical examinations necessary to test a likely diagnosis, rather than producing the diagnosis directly:

> 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)

The developers worried that implementing user-facing explanations would run the risk of disrupting an otherwise smooth work practice to such a degree that it would not contribute to easing clinicals' work, but in fact introduce even *more* work on the computer. This highlights a problem which is often neglected in scholarship on accountability in AI: Explanations at the level of the user interface may counteract the efficiency gained from applying AI in the first place, creating a tension between the value of accountability and the value of usability. Owing to these experiences, the developers changed their understanding of explanations; they centered explanations around the purpose that they serve, rather than only looking at which part of the AI system that explanations target as in the usual classifications of xAI methods [44]. In doing so, they identified four distinct purposes for explanations: (1) Making models intelligible and usable to users of AI systems; (2) Enabling users to trace outcomes and errors; (3) Understanding, debugging, and improving models in the development process, and; (4) Assessing models in auditing situations.

Simultaneously, the developers' extensive project on xAI showed that the explanations generated did not match their expectations for how such explanations would work in a clinical setting. In particular, they learned that xAI provides *functional* explanations of how predictive models work *at a general level*. In this way, explanations might discard information which could be of importance to a physician, thereby potentially causing misunderstandings of the patient condition:

> This is the model's image of a sepsis-patient, in principle…It's a *functional* explanation in the sense that it explains how the model works *in general*. It's not an explanation of the model's *complexity* but just an explanation that reflects the model's *image* of reality. So, if the model has understood that pulse and blood pressure always correlate, then, in principle, we don't know if the model has put less weight on one feature than the other even though they matter equally…You might be able to get some good explanations from it, but they are still explanations that are conditioned on the model's image of reality. And, if it [the model] has forced [the weight on] feature x to zero, this will be reflected in the explanation. (Chief Engineer, Feb. 2020)

While the developers found the use of xAI methods in a clinical setting to be somewhat problematic, they learned that they were quite valuable in the development process for the purpose of debugging models and identifying, for instance, which features to include in the training of ML models. This change in attitude towards xAI explanations is in line with recent studies and criticism of such explanations also outlined earlier in the article [6]. The experiences of the developers combined with theoretical critiques paints a picture of explanations generated with xAI techniques as unable to serve the role that they are often given in policy documents (see e.g. [4, 32, 75]).

5 Discussion & Conclusion

This article has studied how the accountability mechanisms of (1) ethical principles, (2) certification standards, and (3) explanation methods are enacted as well as responded to and reflected on by developers in applied AI, and to what extent these mechanisms promote accountability in and after AI development processes and the use of responsible approaches during such processes. We have studied this empirically on the basis of data from an ethnographic case study conducted at an AI company in Scandinavia. For this purpose, we have drawn on Mark Bovens' conceptualization of accountability as a mechanism [8], along with theory from the fields of philosophy of information, and ethics and information technology. Our analysis reveals an important gap between the way accountability as a social and ethical value is discussed and conceptualized at the policy and research level, and the way that accountability is enacted in practice, where it becomes a responsibility put on the shoulders of engineers.

With regards to ethical principles, we found that they were regarded by the developers to be of little use in their work, and this

became a source of frustration to them. The developers felt that the discussions of ethical principles and issues in guidelines and the public debate were based on a general misunderstanding of AI and the nature of the applications they were developing. While they believed AI to be just another tool or method that might be used to assist in certain tasks, they encountered discussions centered around the premise of autonomous AI decisions and general AI. Thus, they felt that AI ethical guidelines were largely irrelevant to the type of work they were doing and, in fact, somewhat harmful to their business; however, they obviously still felt targeted by them. Here, perhaps, we see the crux of the matter: The high-profile, important and principled discussions of AI ethics reflected in ethical guidelines may often be developed in response to worst-case scenarios and very disruptive applications of AI, e.g. mass surveillance, drones, and self-driving cars. Meanwhile, less attention is paid to the smaller scale efforts to use AI as a supportive technology to assist professional work in important sectors. Although healthcare may be regarded as critical infrastructure, not all actions and decisions that AI may support are equally critical. The evaluation as to *how* critical a decision or action is thus becomes crucial, not least to the developers faced with important design choices in AI development processes.

As an accountability mechanism, the purpose of ethical principles is not to outline specific instructions but rather to establish normative expectations and obligations [37]. Seen in this light, we might still observe an important effect of the extensive discussions of AI ethics, as the developers recognized the great expectations there were to their integrity and conduct, and tried to act accordingly. In this way, the question of *whether* to act ethically had been long settled, and rather, the developers craved guidance on *how* to act ethically and responsibly as suppliers of AI systems for healthcare. Centering ethical discussions and principles around more concrete and sector-specific uses of AI could, perhaps, help to ensure that guidelines would have a greater benefit to developers. This recommendation is in line with other studies (see e.g. [70]). For example, our study revealed that the developers were faced with important choices as to *when* to give prioritization to explanation as a means of accountability over other important values such as usability, which they mainly were concerned with. Undoubtedly, it is important to ensure accountability for patients as these are the ones ultimately affected, and, therefore, incentives for developers to provide such accountability are vital. Although many ethical guidelines suggest incorporating perspectives of *all* users, including end users, and the GDPR pushes for an informal 'right to explanation', these requests are seemingly not enough.

Our data also revealed an air of distrust and suspicion towards the regulatory authorities, particularly the European Commission, as the developers suspected that the ethical principles developed by the AI HLEG [4] were nothing more than an attempt to cover up the commission's lack of progress with regards to legislation in the EU. Rather than a vague discussion of ethical principles, the AI developers preferred to have clear expectations and (changes in) legislation in place as early as possible. This would allow them to adjust their business accordingly and thereby safeguard their competitiveness as a small-scale firm on the global market. On this basis we furthermore recommend, regardless of the validity of the developers' suspicions, that extra care is taken that ethical principles will not be used as an excuse to delay (changes in) legislation that otherwise could bring clarity to the requirements of AI systems in concrete cases. This recommendation is in accordance with recent studies, cautioning against the misuse of ethical guidelines to delay legislation (see e.g. [38]).

As for certification standards, we found that the developers were much more motivated to demonstrate their accountability and integrity through conformity to such standards, than they were to explicitly adhere to ethical principles. The primary reason was that certifications provided them with clear advantages: The ISO 13485 certification was an essential stepping stone in order to achieve the CE mark required when marketing a medical device in Europe, and the ISO/IEC 27001 was considered a prestigious certificate, documenting their level of information security management. While the motivation to do the extensive preparatory work required to implement the standards was pragmatic, the certifications were still a very serious matter to the developers because of their importance to the competitiveness and credibility of the company. The vast amount of effort that the developers put into preparing for certifications and navigating the complicated certification system made it clear that achieving the certifications was a top priority for the company. Yet, they worried that the great effort required would bar smaller AI firms from being certified and thereby make important standards practically unattainable for such firms. This and the complexity of the certification system is indeed critical. However, it might be even more critical if AI companies are faced with little to no guidance on how to conform to such standards and a certification process unfit to deal with AI systems in depth, as our study suggested.

Given that certification standards as mechanisms for accountability establish an accountability relationship [8, 9] in which an authorized certification body attest to the wider public that an AI system works in a desired way, it is paramount that the integrity and reliability of the certification process are not compromised. Therefore, if certification standards are to be used for implementing ethical principles and play as great a role in the accountability system for AI as they have for other sectors [13, 17, 26], it is vital that the certification system is provided with sufficient resources. Proper infrastructure is needed so that certification bodies actually *can* attest to the wider public that an AI system meets the expectations placed on its developers. Otherwise, we may risk that the credibility of the certification system is jeopardized, and that certification fails to advance its goal of ethically responsible AI, commonly referred to as *means-ends de-coupling* [24]. In this latter case, the consequence may further be that accountability as a value is eroded. The power of regulatory bodies to push companies towards ensuring accountability for their products and users should not be underestimated. However, a poor certification process might result in the opposite effect, as developers are pushed to comply with standards that appear to them as arbitrary or frivolous in order to achieve a certification that is *de facto* mandatory for their business. We recommend that more research is done on the use of certification standards in order to further illuminate the problems indicated by our study.

As for explanation methods, the AI developers initially had high expectations for the use of user-facing explanations generated with xAI techniques, and had therefore started a large project exploring this branch of explanation techniques. However, as their project matured, the developers learned that these techniques were somewhat problematic in the user-facing role which they had expected them to fit into. One of the greatest problems they faced was that the display of automatically generated explanations to users might undermine the efficiency gained from applying AI in the first place. This realization coincided with a change of focus from developing decision support systems to developing what they termed 'microservices', aimed at easing the work of clinical practice by automatically starting small actions in the medical examination flow. Rather than centering explanations around the ML model, as is normally the case in xAI [44, 71], the developers learned that a more salient distinction for their work was to look at the purpose of the explanation for the *user*, keeping in mind the role of the explanation in practice. The fact that explanations must be *usable* in order to be *used* should be taken into account in future work in xAI, as 'usability' is an important aspect of the explanation which is neglected if the explanation is not considered in relation to its recipient [69].

Furthermore, the developers discovered that the reliability and stability of explanations generated with xAI techniques potentially made it problematic to use explanations in a clinical setting; the level of such explanations may not match the level of the information required when dealing with medical conditions of patients. This is a sentiment which echoes recent scholarship on the topic [42, 62, 81]. Instead, the developers found xAI techniques to be of great use in the development process in order to understand deep-learning models better and debug them. In this way, the use of xAI explanations as internal tools for development could potentially have the added benefit of improving the traceability of AI developments, as choices in the design and development process are backed up by explanations, giving the developers a better idea of the logics embedded in their models. Such traceability could itself improve the accountability for AI systems, as developers will be more able to account for the choices made when creating a system which is often requested in ethical guidelines, standards, and research studies (see e.g. [4, 22, 58]). These experiences of the developers combined with theoretical critiques paints a picture of explanations generated with current xAI techniques as unable to serve the role that such explanations often are given in policy documents to automatically provide accounts of AI systems and outcomes to the users of such systems [4, 32, 75]. In order to serve as an accountability mechanism in this regard, our study suggests that xAI techniques will still need further development. In particular, the explanations need to become more reliable and stable, and they need to be tailored to the specific context in which they are used. Until then, xAI seems to be relegated to supporting development processes rather than ensuring accountability and transparency for users of AI systems.

The issues discussed in this section should be taken seriously going forward, as they risk eroding the many efforts made to ensure that AI systems for critical sectors like healthcare are designed and developed to be accountable and in line with ethical and social values. Although this article was based on a single yet extensive empirical case study, the developers' experiences in the study may very well be shared by many, and can serve as fruitful perspectives on high and mid-level policymaking. 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 ensuring the responsible design, development, and use of AI. Based on our empirical findings, we have presented several recommendations to remedy the flaws identified in the current ways that ethical principles, certification standards, and explanation methods are enacted as mechanisms for 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 wider public.

ACKNOWLEDGMENTS

We want to thank the director and employees at the AI company for participating in interviews and allowing for participant observation. Also, we want to thank the project steering group that made it possible to follow the design and development of one particular AI system. Finally, thanks to the reviewers for their fruitful comments which helped improve the article.

This work was funded by Aarhus University and Aarhus University Research Foundation, grant number AUFF-E-2015-FLS-8-55.

REFERENCES

[1] Abbott, K. W.; and Duncan, S. 2000. Hard and Soft Law in International Governance. *International Organization* 54 (3): 421–56. doi.org/10.1162/002081800551280.

[2] Agre, P. E. 1997. Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work, edited by G. C. Bowker; L. Gasser, S.L. Star; and B. Turner. USA: L. Erlbaum Associates Inc.

[3] 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.

[4] AI HLEG. 2019. Ethics Guidelines for Trustworthy AI. Report for the European Commission by the High-Level Expert Group on Artificial Intelligence. Report no. B-1049. Brussels, Belgien. Available at: ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[5] Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS One* 10(7): e0130140. doi.org/10.1371/journal.pone.0130140.

[6] Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J. Puri, R.; Moura, J. M. F.; and Eckersley, P. 2020. Explainable Machine Learning in Deployment. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 648–57. FAT* '20. Barcelona, Spain: Association for Computing Machinery. doi.org/10.1145/3351095.3375624.

[7] Blair, M. M. 2008. The New Role for Assurance Services in Global Commerce. *Journal of Corporation Law* 33: 325-360. Available at: scholarship.law.vanderbilt.edu/faculty-publications/30.

[8] 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.

[9] Bovens, M.; Schillemans, T.; and Goodin, R. E. 2014. Public Accountability. *The Oxford Handbook of Public Accountability*, edited by M. Bovens; and R. E. Goodin 1(1): 1–22. doi: 10.1093/oxfordhb/9780199641253.001.0001.

[10] Brundage, M. 2016. Artificial intelligence and responsible innovation. *Fundamental Issues of Artificial Intelligence*, edited by V. E. Müller, 543-555. Switzerland: Springer International Publishing.

[11] Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H. et al. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv preprint. ArXiv:2004.07213 [Cs.CY]. Available at: arxiv.org/abs/2004.07213.

[12] Bryson, J. 2018a. Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics. *Ethics and Information Technology* 20(1): 15–26. doi.org/10.1007/s10676-018-9448-6.

[13] Bryson, J. 2018b. AI & Global Governance: No One Should Trust AI. *United Nations University Centre for Policy Research* (blog). 2018. Available at: cpr.unu.edu/ai-global-governance-no-one-should-trust-ai.html.

[14] Bryson, J.; and Winfield, A. 2017. Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems. *Computer* 50(5): 116–19. doi.org/10.1109/MC.2017.154.

[15] Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of Machine Learning Research, edited by A. F. Sorelle; and C. Wilson, 81: 77–91. New York, NY, USA: PMLR. proceedings.mlr.press/v81/buolamwini18a.html.

[16] Burrell, J. 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3(1): 1-12. doi.org/10.1177/2053951715622512.

[17] Büthe, T.; and Mattli, W. 2011. *The New Global Rulers: The Privatization of Regulation in the World Economy*. Oxford; Princeton, N.J.: Princeton University Press.

[18] Cabitza, F.; Ciucci, D.; and Rasoini, R. 2019. A Giant with Feet of Clay: On the Validity of the Data That Feed Machine Learning in Medicine. *Organizing for the Digital World*, edited by F. Cabitza; C. Batini; and M. Magni, 121–36. Lecture Notes in Information Systems and Organisation. Cham: Springer International Publishing. doi.org/10.1007/978-3-319-90503-7_10.

[19] Cabitza, F.; Rasoini, R.; and Gensini, G. F. 2017. Unintended Consequences of Machine Learning in Medicine. *JAMA* 318(6): 517–18. doi.org/10.1001/jama.2017.7797.

[20] Canca, C. 2020. Operationalizing AI Ethics Principles. *Communications of the ACM* 63(12): 18–21. doi.org/10.1145/3430368.

[21] Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15: 1721–30. Sydney, NSW, Australia: Association for Computing Machinery. doi.org/10.1145/2783258.2788613.

[22] CEN-CENELEC. 2020. CEN-CENELEC Focus Group Report: Road Map on Artificial Intelligence (AI). CEN-CENELEC Road Map Report on AI, version 2020-09. Available at: ftp.cencenelec.eu/EN/EuropeanStandardization/Sectors/AI/CEN-CLC_FGR_RoadMapAI.pdf.

[23] Chatila, R.; and Havens, J. C. 2019. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Robotics and Well-Being*, edited by M. I. A. Ferreira; J. S. Sequeira; G. S. Virk; M. O. Tokhi; and E. E. Kadar, 95: 11–16. Cham: Springer International Publishing. doi.org/10.1007/978-3-030-12524-0_2.

[24] Cihon, P.; Kleinaltenkamp, M. J.; Schuett, J.; and Baum, S. D. In review. AI Certification: Advancing Practice by Reducing Information Asymmetries. IEEE Transactions on Technology and Society.

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

[26] Delimatsis, P. 2015. *The Law, Economics and Politics of International Standardisation*. Cambridge International Trade and Economic Law. Cambridge, United Kingdom: Cambridge University Press.

[27] Denzin, N. K. 2013. Writing and/as Analysis or Performing the World. *The SAGE Handbook of Qualitative Data Analysis*, edited by U. Flick, 569-584. London, UK: Sage Publications.

[28] Deo, R. C. 2015. Machine Learning in Medicine. *Circulation* 132(20): 1920–30. doi.org/10.1161/CIRCULATIONAHA.115.001593.

[29] Diakopoulos, N. 2015. Algorithmic Accountability. *Digital Journalism* 3(3): 398–415. doi.org/10.1080/21670811.2014.976411.

[30] Diakopoulos, N.; Friedler, S.; Arenas, M.; Barocas, S.; Hay, M.; Howe, B.; Jagadish, H. V. et al. 2017. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. *FAT/ML*. Available at: fatml.org/resources/principles-for-accountable-algorithms.

[31] 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.

[32] Doshi-Velez, F., Kortz, M.; Budish, R.; Bavitz, C.; Gershman, S.; O'Brien, D.; Scott, K. et al. 2019. Accountability of AI Under the Law: The Role of Explanation. arXiv preprint. ArXiv:1711.01134 [Cs.AI]. Available at: arxiv.org/abs/1711.01134.

[33] Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway Feedback Loops in Predictive Policing. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81: 160–71. proceedings.mlr.press/v81/ensign18a.html.

[34] Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* 542(7639): 115–18. doi.org/10.1038/nature21056.

[35] Eubanks, V. 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York, NY, USA: St. Martin's Press.

[36] Floridi, L. 2017. Infraethics–on the Conditions of Possibility of Morality. *Philosophy & Technology* 30: 391–94. doi.org/10.1007/s13347-017-0291-1.

[37] Floridi, L. 2018. Soft Ethics, the Governance of the Digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376(2133). doi.org/10.1098/rsta.2018.0081.

[38] Floridi, L. 2019. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology* 32(2): 185–93. doi.org/10.1007/s13347-019-00354-x.

[39] Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C. et al. 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28(4): 689–707. doi.org/10.1007/s11023-018-9482-5.

[40] Forsythe, D.; and Hess, D. J. 2001. *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Writing Science. Stanford, CA: Stanford University Press.

[41] Geertz, C. 1973. *The Interpretation of Cultures*. New York, NY, USA: Basic Books.

[42] Ghorbani, A: Abid, A.; and Zou, J. 2019. Interpretation of Neural Networks Is Fragile. In Proceedings of the AAAI Conference on Artificial Intelligence 33(01): 3681–88. doi.org/10.1609/aaai.v33i01.33013681

[43] Goodman, B.; and Flaxman, S. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38(3): 50–57. doi.org/10.1609/aimag.v38i3.2741.

[44] Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* 51(5) Article 93: 1-42. doi.org/10.1145/3236009.

[45] Gulshan, V.; Peng, P.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S. et al. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316(22): 2402–10. doi.org/10.1001/jama.2016.17216.

[46] DARPA. 2017. Broad Agency Announcement: Explainable Artificial Intelligence (XAI), DARPA-BAA-16-53. Available at: darpa.mil/attachments/DARPA-BAA-16-53.pdf.

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

[48] Halkier, B. 2011. Methodological Practicalities in Analytical Generalization. *Qualitative Inquiry* 17(9): 787–97. doi.org/10.1177/1077800411423194.

[49] ISO. 2013. ISO/IEC 27001:2013. International Organization for Standardization. Available at: iso.org/isoiec-27001-information-security.html.

[50] ISO. 2016. ISO 13485:2016 Medical Devices — Quality Management Systems — Requirements for Regulatory Purposes. International Organization for Standardization. Available at: iso.org/standard/59752.html.

[51] ISO. 2021a. Consumer Standards: Partnership for a Better World. International Organization for Standardization. Available at: iso.org/sites/ConsumersStandards/index.html#top

[52] ISO. 2021b. Popular Standards. International Organization for Standardization. Available at: iso.org/popular-standards.html.

[53] Jaton, F. 2017. We get the algorithms of our ground truths: Designing referential databases in digital image processing. *Social Studies of Science* 47(6): 811-840. doi.org/10.1177/0306312717730428.

[54] Jobin, A; Ienca, M.; and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1(9): 389–99. doi.org/10.1038/s42256-019-0088-2.

[55] Kannel, W. B.; Doyle, J. T.; McNamara, P. M.; Quickenton, P.; and Gordon, T. 1975. Precursors of Sudden Coronary Death. Factors Related to the Incidence of Sudden Death. *Circulation* 51 (4): 606–13. doi.org/10.1161/01.CIR.51.4.606.

[56] Kemper, J.; and Kolkman, D. 2019. Transparent to Whom? No Algorithmic Accountability without a Critical Audience. *Information, Communication & Society* 22 (14): 2081–96. doi.org/10.1080/1369118X.2018.1477967.

[57] Kooi, T.; Litjens, G.; van Ginneken, B.; Gubern-Mérida, A.; Sánchez, C. I.; Mann, R.; den Heeten, A.; and Karssemeijer, N. 2017. Large Scale Deep Learning for Computer Aided Detection of Mammographic Lesions. *Medical Image Analysis* 35 (January): 303–12. doi.org/10.1016/j.media.2016.07.007.

[58] Kroll, J. A. 2018. The Fallacy of Inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133): 20180084. doi.org/10.1098/rsta.2018.0084.

[59] Kroll, J. A. 2020. Accountability in Computer Systems. *The Oxford Handbook of Ethics of AI*, 181.

[60] Kroll, J. A.; Huey, J.; Barocas, S.; Felten, E. W.; Reidenberg, J. R.; Robinson, D. G.; and Yu, H.. 2017. Accountable Algorithms. *University of Pennsylvania Law Review* 165: 74.

[61] Kvale, S. (2008). *Doing interviews*. London, UK: Sage Publications.

[62] Lakkaraju, H.; and Bastani, O. 2020. 'How Do I Fool You?': Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85. AIES '20. New York, NY, USA: Association for Computing Machinery. doi.org/10.1145/3375627.3375833.

[63] Larsson, S. 2020. On the Governance of Artificial Intelligence through Ethics Guidelines. *Asian Journal of Law and Society*, 1–15. doi.org/10.1017/als.2020.19.

[64] Latour, B. 1987. Science in action: How to follow scientists and engineers through society. Harvard University Press.

[65] Lepri, B.; Oliver, N.; Letouzé, E.; Pentland, A.; and Vinck, P. 2018. Fair, Transparent, and Accountable Algorithmic Decision-Making Processes. *Philosophy & Technology* 31(4): 611–27. doi.org/10.1007/s13347-017-0279-x

[66] Lipton, Z. C. 2018. The Mythos of Model Interpretability. *Queue* 16(3): 31-57. doi.org/10.1145/3236386.3241340

[67] Malterud, K. 2012. Systematic text condensation: a strategy for qualitative analysis. *Scandinavian Journal of Public Health* 40(8): 795–805. doi.org/10.1177/1403494812465030

[68] Martin, K. 2019. Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics* 160(4): 835–50. doi.org/10.1007/s10551-018-3921-3

[69] Miller, T. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267: 1–38. doi.org/10.1016/j.artint.2018.07.007.

[70] Mittelstadt, B. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence* 1 (11): 501–7. doi.org/10.1038/s42256-019-0114-4.

[71] Mittelstadt, B.; Russell, C.; and Wachter, S. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 279–88. Atlanta, GA, USA: ACM Press. doi.org/10.1145/3287560.3287574.

[72] Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; and Müller, K.-R. 2017. Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recognition* 65: 211–22. doi.org/10.1016/j.patcog.2016.11.008.

[73] Morley, J.; and Floridi, L. 2019. Enabling Digital Health Companionship Is Better than Empowerment. *The Lancet Digital Health* 1 (4): e155–56. doi.org/10.1016/S2589-7500(19)30079-2.

[74] Nissenbaum, H. 1994. Computing and Accountability. *Communications of the ACM* 37 (1): 72–81. doi.org/10.1145/175222.175228

[75] OECD. 2020. OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. Organisation for Economic Co-operation and Development.

[76] O'Neil, C. 2017. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. London: Penguin Books.

[77] Power, M. 1997. *The Audit Society: Rituals of Verification*. Oxford, [England]; New York: Oxford University Press.

[78] Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P.. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. ArXiv:2001.00973 [Cs], January. Available at: arxiv.org/abs/2001.00973.

[79] Rességuier, A.; and Rodrigues, R. 2020. *AI ethics should not remain toothless!* A call to bring back the teeth of ethics. *Big Data & Society*, 1-5. doi.org/10.1177/2053951720942541

[80] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. KDD '16. New York, NY, USA: Association for Computing Machinery. doi.org/10.1145/2939672.2939778.

[81] Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1 (5): 206–15. doi.org/10.1038/s42256-019-0048-x.

[82] Russell, S. J.; and Norvig, P. 2016. *Artificial Intelligence: A Modern Approach*, 3rd Edition. Pearson Education, 2009.

[83] Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; and Young, M. 2014. Machine Learning: The High Interest Credit Card of Technical Debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*.

[84] Shickel, B.; Tighe, P. J.; Bihorac, A.; and Rashidi, P. 2018. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics* 22 (5): 1589–1604. doi.org/10.1109/JBHI.2017.2767063.

[85] Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the Black Box of Deep Neural Networks via Information. *ArXiv:1703.00810 [Cs]*. Available at: arxiv.org/abs/1703.00810.

[86] Sinclair, A. 1995. The Chameleon of Accountability: Forms and Discourses. *Accounting, Organizations and Society* 20 (2–3): 219–37. doi.org/10.1016/0361-3682(93)E0003-Y.

[87] Smith, H. 2020. Clinical AI: Opacity, Accountability, Responsibility and Liability. *AI & SOCIETY*, July. doi.org/10.1007/s00146-020-01019-6.

[88] Sossin, L.; and Smith, C. W. 2003. Hard Choices and Soft Law: Ethical Codes, Policy Guidelines and the Role of the Courts in Regulating Government. *Alberta Law Review*: 867–89. doi.org/10.29173/alr1344.

[89] Spradley, J. P. 1980. *Participant Observation*. New York: Holt, Rinehart and Winston.

[90] St. Pierre, E. A. 2011. Post qualitative research: The critique and the coming after. In *The SAGE Handbook of Qualitative Research*, edited by A. K. Denzin and Y. S. Lincoln, 4th Edition, 11-25. London, UK: Sage Publication.

[91] Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.

[92] Sweeney, L. 2013. Discrimination in Online Ad Delivery. *Queue* 11 (3): 10–29. doi.org/10.1145/2460276.2460278

[93] the European Commission. 2020. Commission Implementing Decision (EU) 2020/437. Available at: eur-lex.europa.eu/eli/dec_impl/2020/437/oj.

[94] Theodorou, A.; and Dignum, V. 2020. Towards Ethical and Socio-Legal Governance in AI. *Nature Machine Intelligence* 2(1): 10–12. doi.org/10.1038/s42256-019-0136-y.

[95] Tupasela, A.; Snell, K.; and Tarkkala, H. 2020. The Nordic Data Imaginary. *Big Data & Society* 7 (1): 2053951720907107. https://doi.org/10.1177/2053951720907107.

[96] Tutt, A. 2016. An FDA for Algorithms. *69 Admin. L. Rev. 83*. Rochester, NY: Social Science Research Network. Available at: papers.ssrn.com/abstract=2747994.

[97] Wachter, S.; Mittelstadt, B.; and Floridi, L. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* (7): 76–99. Oxford University Press. doi.org/10.1609/aimag.v38i3.2741.

[98] Weller, A. 2019. Transparency: Motivations and Challenges. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, edited by Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L.; Müller, K. R. 23–40. Lecture Notes in Computer Science. Cham: Springer International Publishing. doi.org/10.1007/978-3-030-28954-6_2.

[99] Whittaker, M.; Crawford, K.; Dobbe, R.; Fried, G.; Kaziunas, E.; Mathur, V.; West, S. M.; Richardson, R.; Schultz, J.; and Schwartz, O. 2018. *AI Now Report 2018*. AI Now Institute at New York University, New York.

[100] Whittlestone, J.; Nyrup, R.; Alexandrova, A.; and Cave, S. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions, In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (*AIES '19*): 195–200. Association for Computing Machinery, New York, NY, USA. doi.org/10.1145/3306618.3314289

[101] Wieringa, M. 2020. What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18. FAT* '20. Barcelona, Spain: Association for Computing Machinery. doi.org/10.1145/3351095.3372833.

[102] Winfield, A. F. T.; and Jirotka, M. 2018. Ethical Governance Is Essential to Building Trust in Robotics and Artificial Intelligence Systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133): 20180085. doi.org/10.1098/rsta.2018.0085.

[103] Yin, R. K. 2009. *Case Study Research: Design and Methods*, 4th Edition. Los Angeles, LA, USA: Sage Publications.

[104] Zerilli, F. M. 2010. The Rule of Soft Law. *Focaal—Journal of Global and Historical Anthropology* (56): 3–18. doi.org/10.3167/fcl.2010.560101.

[105] Zhang, J. M.; Harman, M.; Ma, L.; and Liu, Y. 2019. Machine Learning Testing: Survey, Landscapes and Horizons. *IEEE Transactions on Software Engineering*. doi: 10.1109/TSE.2019.2962027.