Kempton, Alexander Moltubakk, and Vassilakopoulou, Polyxeni (2021): Accountability, Transparency & Explainability in AI for Healthcare. 8th International Conference on Infrastructures in Healthcare. DOI: 10.18420/ihc2021_018

Accountability, Transparency and Explainability in AI for Healthcare

Alexander Moltubakk Kempton, Polyxeni Vassilakopoulou
University of Oslo, University of Agder
alexansk@ifi.uio.no, *polyxenv@uia.no*

Abstract. The multiplicity of actors and the opacity of technologies involved in data management, algorithm crafting and systems´ development for the deployment of Artificial Intelligence (AI) in healthcare create governance challenges. This study analyzes extant AI governance research in the context of healthcare focusing on accountability, transparency and explainability. We find that a significant part of this body of research lacks conceptual clarity and that the relationship between accountability, transparency and explainability is not fully explored. We also find that papers written back in the 1980s, identify and discuss many of the issues that are currently discussed. Up to today, most published research is only conceptual and brings contributions in the form of frameworks and guidelines that need to be further investigated empirically.

Introduction

The increased availability of health data creates new knowledge-creation opportunities that can transform clinical practice. The rapid explosion in Artificial Intelligence (AI) allows leveraging health data for the development of powerful models that can automate diagnoses, enable a precision approach to medicine by tailoring treatments and increase the efficiency and effectiveness in the use of resources (Panch, Mattie, & Celi, 2019). A growing number of information systems supporting healthcare embed AI technologies. AI refers to machines performing the cognitive functions typically associated with humans, including perceiving, reasoning and learning (McCarthy, Minsky, Rochester, & Shannon, 2006; Rai, Constantinides, & Sarker, 2019). These technologies are used in diverse healthcare application areas including processing medical imaging, supporting triage assessments, expediting disease diagnoses, automating patient monitoring and supporting biopharmaceutical development. As AI is infusing nearly every aspect of healthcare delivery, the governance of AI-enabled systems emerges as a growing concern.

A significant challenge in AI governance is the "black box" problem. This is the problem of putting in place an opaque process of transforming data inputs to insight outputs (e.g. related to diagnoses, or outcomes predictions). AI solutions that use machine learning are particularly susceptible to the black box problem. This opaqueness can pose limits on involving humans in operating and monitoring AI-enabled healthcare information systems. Transparency and explainability are directly related to addressing the black-box problem.

Furthermore, the multiplicity of actors and technologies involved in a) the registration, storage and management of the data required for the AI algorithms, b) the development, validation and maintenance of the algorithms and c) the development, deployment and handling of the applications that embed AI algorithms creates accountability challenges. The academic community can contribute sociotechnical approaches for accountability in AI systems (Vassilakopoulou, 2020). Ensuring accountability for decisions and actions within these complex infrastructural arrangements is key for introducing AI technologies in clinical practice (Habli, Lawton, & Porter, 2020).

The seminal work by Bovens and colleagues on accountability and governance (Bovens, 2007, 2010; Bovens, Schillemans, & Goodin, 2014) provides a comprehensive accountability definition that includes three complementary aspects: the obligation of actors to answer for and justify their actions, the interrogation ability of those affected and sanctions when systems work in unacceptable ways. This comprehensive definition indicates some interdependency between accountability and addressing the "black box" problem. One of the key aspects of accountability is the interrogation ability of those affected by systems and black-boxing may impair the ability to interrogate.

There is a strong interest on AI for healthcare in both academia and practice resulting to a growing volume of related research. We developed a synthesis of the extant body of research, by performing a structured literature review following the process proposed by Kitchenham (2004). The review was guided by the following research questions:

- RQ1: What are the key insights provided by extant research on AI accountability in healthcare?
- RQ2: How is accountability defined in extant research on AI accountability in healthcare?
- RQ3: How are the relations between accountability, explainability, transparency understood and conceptualized?

The remainder of the paper is organized as follows. First, we present the method used, then, we present our findings, finally, we conclude by discussing the implications for further research and end with overall concluding remarks.

Method

The approach followed for the literature review is based on the three-step structured literature review process proposed by Kitchenham (2004). The steps include: a) planning the review, where a detailed protocol containing specific search terms and inclusion/exclusion criteria is developed, b) conducting the review, where the identification, selection, appraisal, examination and synthesis of published research is performed and c) reporting the review, where the write-up is prepared. We used these steps as our methodological framework.

To identify journal and conference articles to be reviewed, we combined two different sections of search terms with AND operators. The first section represents technologies associated with the objective of our research (AI OR "artificial intelligence") AND (Healthcare OR Health). The second section reflects the concern for accountability in AI governance (accountability OR accountable). Combining these three sections we searched in the abstract, title or keywords within published research. We utilized Scopus as our search engine. This search yielded 73 articles. Furthermore, to ensure as wide coverage of related literature as possible, we searched for articles that are related to AI accountability in health but do not mention the words health or healthcare in their abstract, title or keywords. For this purpose, we searched for papers on (AI OR "artificial intelligence") AND (accountability OR accountable) without (Healthcare OR Health) within outlets flagged by Scopus as medical, nursing or healthcare related. This search yielded 15 additional articles. The search was performed on April 5th, 2021.

For screening the papers, specific exclusion criteria were used. We excluded documents that are: a) not research articles (e.g. interviews, research proposals), b) not focusing on AI (but only causally mention AI), c) not engaging with AI governance (but only casually mention accountability), d) not focusing on healthcare (but only casually mentioning the word as one of many potential contexts for AI use), e) not written in English. We did not set any specific time period for our search. Interestingly, most papers identified are recent (published after 2018) but a couple of them are from the earlier AI spring of the 1980s. In total, 21 articles were included from the 88 that were initially identified. Within these 21 articles, 9 also include the terms explainability or Transparency in their abstract, title or keywords. We flagged these papers in order to explore the relations between accountability, explainability and transparency (RQ3) and we coded the papers´ content in a spreadsheet shared by the authors. Throughout the process, we held meetings to ensure consensus in the coding. The distribution of papers over time is presented in Table I. The full list of articles reviewed is included in the Annex.

Table I. Articles distribution over publication years

Publication Year	Selected Articles (in parenthesis the papers that include accountability and also explainability/transparency)
1986	2 (0)
2018	2 (0)
2019	8 (2)
2020	5 (5)
2021	4 (2)

Findings

The articles reviewed cover different AI healthcare application domains (see table II). AI for medical image analysis is the most frequent domain (2 articles discussing image analysis for breast tissues, 1 on the analysis of cancer tissues in general and 1 on image analysis for regulating radiation doses in CT scanners). AI applications for mental health (digital consultation services) are discussed in 3 articles. AI enabled surgery robotics are discussed in 2 articles and 1 article is on the use of AI for Electrocardiogram (ECG) interpretation. About half of the articles reviewed (11 of 21) do not engage with specific healthcare application domains but they discuss healthcare AI applications in general.

Table II. AI in healthcare application domains in articles reviewed

Application Domain	Articles
AI for medical image analysis	4
AI for mental health	3
AI for surgery robotics	2
AI for ECG interpretation	1
AI for healthcare in general	11

Around half of the articles reviewed are "solution-oriented" offering suggestions for what needs to be done in practice to ensure Accountability for AI in healthcare (12 out of 21 articles). The suggested solutions include design principles and requirements, technical artifacts (algorithms, software applications) and ethical guidelines. Furthermore, a significant number of the articles reviewed contribute to the literature at the conceptual level (9 articles) by discussing issues and identifying challenges and opportunities or suggesting frameworks and conceptual maps. The remaining articles include legal essays and insights on users´ perceptions and behavior. It is interesting to note that three articles include both conceptual frameworks and solution-oriented contributions.

Table III. Articles´ Contribution Types (three papers had more than one type of contribution)

Type of Contribution		Articles
Solution-oriented	Design Principles and Requirements	4
	Ethical Guidelines	4
	Technical Artifacts (algorithms, software applications)	4
Conceptual	Discussion on Challenges and Opportunities	6
	Framework – Conceptual Mapping	3
Other	Legal Essays	2
	Perceptions and Behavioural Insights	1

By analyzing the articles, we found that a significant part of them lacks conceptual clarity. Around half of the articles reviewed (13 of 21) do not include a definition for the accountability concept. This is a significant issue as accountability is a malleable term and its liberal use easily leads to conceptual confusion. We classified the definitions provided in the articles that do define the concept using the work of Bovens on accountability and governance (Bovens, 2007, 2010; Bovens et al., 2014). Specifically, Bovens suggested a comprehensive definition which covers a) the obligation of actors involved in the development and deployment of AI systems to answer for and justify their actions, b) the interrogation ability of those affected by AI systems and c) the sanctioning potential that entails specifying what is acceptable in the use of AI and what happens when AI systems work in unacceptable ways. Within the articles reviewed we found one that defines the concept in a comprehensive way covering all three aspects (obligation, interrogation, sanctioning) and four that cover two of these aspects. The remaining three only cover one aspect (two only cover "obligation", one covers "interrogation"). Overall, accountability is most frequently conceptualized as an obligation or responsibility of the actors involved (table IV). When accountability is defined as the ability to interrogate about AI-enabled systems, the focus is shifted from those involved in the development and deployment of AI-enabled systems to those affected by these systems and their own ability to pose questions and make sense of AI applications. Finally, the "sanctioning" type of definitions points to the need to have in place rules for what happens when AI is not within what is acceptable.

Table IV. Accountability Definitions in the papers reviewed

Definitions in the papers reviewed		Articles
Accountability Aspects	Obligation (also expressed as "responsibility for")	6
	Interrogation ability	3

	Posthoc sanction potential (for blamable agents)	3
No Definition		13

The analysis helped us identify differences in the way the relationship between accountability and transparency/explainability is conceptualized. We were motivated to analyze this relationship by observing how these concepts are treated in non-scientific literature. We noticed that in popular press they are frequently used interchangeably. Interestingly, this issue appears in only one of the articles reviewed. Among the articles that include accountability and also explainability and/or accountability, most discuss them as discrete characteristics that can be assessed separately (four articles) or as discrete but enabling (four articles). In the latter category, explainability and transparency are conceptualized as being enabling for achieving accountability. Table V provides an overview.

Table V. Overview of the relationships between accountability and explainability/transparency

Relationship between accountability and explainability/transparency	**Articles**
Discrete	4
Discrete but enabling	4
Tightly related - conflated	1

Finally, an interesting observation, is that the majority of the articles reviewed are not empirical. Specifically, only 2 of the 21 articles reviewed include empirical data: one analyses trials with 14 pathologists on AI-enabled diagnoses for histopathological cancer tissues and one investigates women´s views on AI for the diagnostic interpretation of screening mammograms (922 participants). This suggests that more research efforts should be geared towards empirical studies.

Discussion and Conclusion

AI opens up great opportunities for leveraging health data to generate insights that can improve healthcare delivery. In order to harness these opportunities, it is important to address AI governance challenges ensuring accountability in AI use. We performed a literature review to map and synthesize extant related research literature and we identified 21 articles published from 1986 to 2021. There is a gap in publishing activity from 1986 to 2018 which reflects the long "AI winter" that lasted from the late 1980´s till the early 2010´s (when the widespread use of machine learning rejuvenated the interest (and funding) on AI applications). Interestingly, the two papers that were written back in 1986, discuss the same issues that are discussed today. Specifically, Hartman suggests the adoption of guidelines to increase programmer and provider accountability for clinical software, training clinicians to understand the limits of artificial intelligence, and determining the legal and ethical status of software (Hartman, 1986a). Furthermore, in a second paper published the same year, he points to the challenges created by black-box algorithms and to the need for delineating types of use (Hartman, 1986b). Enthusiasm for AI during the 1980s was followed by a severe cutback in interest which lasted for two decades resulting to limited advancement for the governance of AI. The same issues that were identified back in the 1980s were brought again in public debate and in academic discourse during the past few years (the first relevant papers were identified in our review in 2018). This time, national and international regulatory authorities are aiming to move swiftly. Recent activity at the policy level (European Commission, 2021) addresses AI use in high risk domains, setting rules and mechanisms to minimize unintended negative consequences of the rapid explosion in AI and introducing a risk based approach to categorize different applications of AI.

In our literature review we find that a significant part of the literature lacks conceptual clarity. This is problematic. When researchers do not define accountability or define it in different ways, they end up addressing different accountability issues, practices, and challenges. In other words, the discourse becomes fragmentated. Future research should therefore be explicit in how key terms are used. We also found that the relationship between accountability and transparency/explainability is yet not fully explored. Some prior research, identifies transparency and explainability as significant accountability enablers (O'Sullivan et al., 2019; Rjoob et al., 2020; Sabol et al., 2020; Tosun et al., 2020). Nevertheless, this relationship may not be straightforward. Durán and Jongsma (2021) suggest that demanding explainability, including full technical transparency for AI may be overdemanding and not really needed for accountability. They use the example of physicians operating other technologies which they do not fully understand or cannot fully explain their inner working of (e.g. MRI scans), yet in these cases, physicians are sufficiently in control to be considered responsible. This review contributes to conceptual clarification by mapping the different accountability definitions and points to the need to further explore the relationship between transparency, explainability and accountability.

Another important finding of this review is the scarcity of empirical research identified. This is understandable when considering the novelty of the phenomenon. It is only recently that the availability of data and the use of AI techniques started to make an impact on organizations. Nevertheless, the high proportion of conceptual and literature-based papers suggests that research efforts should be geared towards empirical studies. The reviewed literature could provide a starting point for empirical investigations. For example, the guidelines and frameworks suggested could be empirically evaluated. There is also a possibility to utilize the literature in an abductive manner: mismatches between empirical material and existing conceptually derived theories can be used as opportunities for theorizing (Markus & Rowe, 2021). This review can provide a basis for development helping researchers orient themselves and position their own work.

Acknowledgments

This work was performed within the AI4Users project, funded by the Norwegian Research Council, IKTPLUSS program (project nr. 311680).

References

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European law journal, 13*(4), 447-468.

Bovens, M. (2010). Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. *West European Politics, 33*(5), 946-967.

Bovens, M., Schillemans, T., & Goodin, R. E. (2014). Public accountability. *The Oxford handbook of public accountability, 1*(1), 1-22.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*(5), 329-335.

European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), published on 21 April 2021. Retrieved from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in health care: accountability and safety. *Bulletin of the World Health Organization, 98*(4), 251.

Hartman, D. E. (1986a). Artificial intelligence or artificial psychologist? Conceptual issues in clinical microcomputer use. *Professional Psychology: Research and Practice, 17*(6), 528.

Hartman, D. E. (1986b). On the use of clinical psychology software: Practical, legal, and ethical concerns. *Professional Psychology: Research and Practice, 17*(5), 462.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele University Technical Report, UK, TR/SE-0401*(2004), 1-26. doi:https://doi.org/10.1.1.122.3308

Markus, M. L., & Rowe, F. (2021). Guest Editorial: Theories of Digital Transformation: A Progress Report. *Journal of the Association for Information Systems, 22*(2), 11.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine, 27*(4), 12-12.

O'Sullivan, S., Nevejans, N., Allen, C., Blyth, A., Leonard, S., Pagallo, U., . . . Ashrafian, H. (2019). Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery, 15*(1), e1968.

Panch, T., Mattie, H., & Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. *NPJ digital medicine, 2*(1), 1-3.

Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's comments: next-generation digital platforms: toward human–AI hybrids. *Mis Quarterly, 43*(1), iii-x.

Rjoob, K., Bond, R., Finlay, D., McGilligan, V., Leslie, S. J., Rababah, A., . . . McShane, A. (2020). Towards Explainable Artificial Intelligence and Explanation User Interfaces to Open the 'Black Box'of Automated ECG Interpretation *Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications* (pp. 96-108): Springer.

Sabol, P., Sinčák, P., Hartono, P., Kočan, P., Benetinová, Z., Blichárová, A., . . . Jašková, A. (2020). Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. *Journal of biomedical informatics, 109*, 103523.

Tosun, A. B., Pullara, F., Becich, M. J., Taylor, D. L., Chennubhotla, S. C., & Fine, J. L. (2020). HistoMapr™: An Explainable AI (xAI) Platform for Computational Pathology Solutions *Artificial Intelligence and Machine Learning for Digital Pathology* (pp. 204-227): Springer.

Vassilakopoulou, P. (2020). Sociotechnical approach for accountability by design in AI systems. *European Conference on Information Systems (ECIS 2020).*

Annex: List of Articles included in the Review (in chronological order)

Authors	**Title**	**Year & Outlet**	**Healthcare domain**	**Relation between Accountability & Transparency/ Explainability**	**Key contribution**
Durán J.M., Jongsma K.R.	Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI	2021 Journal of Medical Ethics	generally about AI in healthcare	Discrete: physicians can be responsible, in terms of accountability without fully knowing or understanding inner workings.	Ethical Guidelines
Gómez Rivas J., … Grossmann R.	Autonomous robots: a new reality in healthcare? A project by European Association of Urology-Young Academic Urologist group	2021 Current opinion in urology	robotics (autonomous robots)	N/A	Discussions on Pros and Cons
Ongena Y.P., Yakar D., Haan M., Kwee T.C.	Artificial Intelligence in Screening Mammography: A Population Survey of Women's Preferences	2021 Journal of the American College of Radiology	image analysis (mammograms)	N/A	Perceptions and Behavioural Insights
Rjoob K., Bond R., …Peace A.	Towards Explainable Artificial Intelligence and Explanation User Interfaces to Open the 'Black Box' of Automated ECG Interpretation	2021 Lecture Notes in Computer Science	ECG interpretation	Enabling: lack of explainability and interpretability (which results to less transparency) increases difficulty of accountability.	Technical Artifacts (algorithms, software applications)
Abràmoff M.D., Tobey D., Char D.S.	Lessons Learned About Autonomous AI: Finding a Safe, Efficacious, and Ethical Path Through the Development Process	2020 American Journal of Ophthalmology	generally about AI in healthcare	Discrete	Frameworks and Mappings Ethical Guidelines
Basu T., Engel-Wolf S., Menzer O.	The ethics of machine learning in medical sciences: Where do we stand today?	2020 Indian Journal of Dermatology	generally about AI in healthcare	Discrete	Ethical Guidelines
Mattei P.	Digital governance in tax-funded European healthcare systems: From the Back office to patient empowerment	2020 Israel Journal of Health Policy Research	generally about AI in healthcare	Discrete	Discussions on Pros and Cons
Sabol P., Sinčák P., … A., Jašková A.	Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images	2020 Journal of Biomedical Informatics	image analysis (histopathological cancer tissues)	Enabling: "explainability improves the accountability of the proposed classifier"	Technical Artifacts (algorithms, software applications)
Tosun A.B., Pullara F., … Fine J.L.	HistoMapr™ : An Explainable AI (xAI) Platform for Computational Pathology Solutions	2020 Lecture Notes in Computer Science	image analysis (breasts)	Enabling: explainability enables accountability	Technical Artifacts (algorithms, software applications)
Forcier M.B., Gallois H., Mullan S., Joly Y.	Integrating artificial intelligence into health care through data access: Can the GDPR act as a beacon for policymakers?	2019 Journal of Law and the Biosciences	generally about AI in healthcare	N/A	Legal Essays

Authors	**Title**	**Year & Outlet**	**Healthcare domain**	**Relation between Accountability & Transparency/ Explainability**	**Key contribution**
Larson D.B., Boland G.W.	Imaging Quality Control in the Era of Artificial Intelligence	2019 Journal of the American College of Radiology	image analysis (control image quality, regulate CT scan radiation)	N/A	Design Principles and Requirements
Lysaght T., Lim H.Y., Ngiam K.Y.	AI-Assisted Decision-making in Healthcare: The Application of an Ethics Framework for Big Data in Health and Research	2019 Asian Bioethics Review	generally about AI in healthcare	Tightly related -conflated and coupled to the "black box" problem	Ethical Guidelines
Milosevic Z.	Ethics in digital health: A deontic accountability framework	2019 IEEE 23rd Enterprise Distributed Object Computing Conference	generally about AI in healthcare	N/A	Technical Artifacts (algorithms, software applications) & Frameworks and Mappings
O'Sullivan S., Nevejans N.,… Ashrafian H.	Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery	2019 International Journal of Medical Robotics and Computer Assisted Surgery	robotics (autonomous robots in surgery)	Enabling, although transparency and explainability are not much discussed in the paper	Discussions on Pros and Cons
Price W.N., II, Cohen I.G.	Privacy in the age of medical big data	2019 Nature Medicine	generally about AI in healthcare	N/A	Discussions on Pros and Cons
Schönberger D.	Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications	2019 International Journal of Law and Information Technology	generally about AI in healthcare	N/A	Legal Essays
Vakkuri V., Kemell K.-K., Abrahamsson P.	Implementing Ethics in AI: Initial Results of an Industrial Multiple Case Study	2019 Lecture Notes in Computer Science	generally about AI in healthcare	N/A	Design Principles and Requirements & Frameworks and Mappings
Martinez-Martin N., Kreitmair K.	Ethical issues for direct-to-consumer digital psychotherapy apps: Addressing accountability, data protection, and consent	2018 Journal of Medical Internet Research	mental health (using AI and conversational agents)	N/A	Discussions on Pros and Cons
Pesapane F., Volonté C., Codari M., Sardanelli F.	Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States	2018 Insights into Imaging	generally about AI in healthcare	N/A	Discussions on Pros and Cons
Hartman D.E.	Artificial Intelligence or Artificial Psychologist?. Conceptual Issues in Clinical Microcomputer Use	1986 Professional Psychology: Research & Practice	mental health	N/A	Design Principles and Requirements
Hartman D.E.	On the Use of Clinical Psychology Software. Practical, Legal, and Ethical Concerns	1986 Professional Psychology: Research & Practice	mental health	N/A	Design Principles and Requirements