Information Systems Frontiers (2023) 25:2139–2157
https://doi.org/10.1007/s10796-021-10146-4

Responsible AI for Digital Health: a Synthesis and a Research Agenda

Cristina Trocin[1] · Patrick Mikalef[1] · Zacharoula Papamitsiou[1] · Kieran Conboy[2]

Accepted: 18 May 2021 / Published online: 26 June 2021

Abstract
Responsible AI is concerned with the design, implementation and use of ethical, transparent, and accountable AI technology in order to reduce biases, promote fairness, equality, and to help facilitate interpretability and explainability of outcomes, which are particularly pertinent in a healthcare context. However, the extant literature on health AI reveals significant issues regarding each of the areas of responsible AI, posing moral and ethical consequences. This is particularly concerning in a health context where lives are at stake and where there are significant sensitivities that are not as pertinent in other domains outside of health. This calls for a comprehensive analysis of health AI using responsible AI concepts as a structural lens. A systematic literature review supported our data collection and sampling procedure, the corresponding analysis, and extraction of research themes helped us provide an evidence-based foundation. We contribute with a systematic description and explanation of the intellectual structure of Responsible AI in digital health and develop an agenda for future research.

Keywords Responsible AI · Artificial intelligence · Ethical concerns · Healthcare · Systematic literature review · Meta-data analysis

1 Introduction

Responsible Artificial Intelligence (AI) is an emerging area that investigates the ethics of AI to understand the moral responsibility in emerging technology (Tigard, 2020). The need for responsible AI has stemmed from a limited understanding of important issues that emerge with the use of such technologies. Recent studies and cases in practice have shown that AI can potentially create unintended consequences such as biases, discrimination, errors or unexpected results, and an overall lack of transparency with regard to how outcomes are achieved (Stahl & Coeckelbergh, 2016). When adopting AI in healthcare, the importance of implementing responsible AI practices is heightened due to the criticality of associated activities and the sensitivity of the data that is used (Morley et al., 2019). Responsible AI is concerned specifically with establishing ethical principles and human values in order to reduce biases and promote fairness, facilitate interpretability and explainability of outcomes, and to ensure robustness and security (Barredo Arrieta et al., 2020; Sambasivan & Holbrook, 2018). The ultimate goal of building AI technologies based on responsible principles is to avoid dramatic negative consequences on human and societal well-being (Dignum, 2019).

These concerns particularly influence the use of AI in healthcare, which integrates and learns from large datasets of clinical data, to support diagnosis, clinical decision making, and personalized medicine. We refer to AI as "*the ability of a system to identify, interpret, make inferences, and learn from data to achieve predetermined organizational and societal goals*" (Mikalef & Gupta, 2021, p. 3). If AI is implemented and used in healthcare responsibly, it can positively contribute to care actors' well-being. However, the use of AI often results in decisions and actions that have moral consequences, undermine ethical principles, and diminish people's rights and dignity (Martin, 2019b). Recent empirical articles highlight how deploying AI is coupled with significant ethical challenges (Floridi & Taddeo, 2016), as the "walking data generators" (individuals/patients) are often unaware of how their medical

✉ Cristina Trocin
cristina.trocin@ntnu.no

Patrick Mikalef
patrick.mikalef@ntnu.no

Zacharoula Papamitsiou
papamitsiou.zacharoula@ntnu.no

Kieran Conboy
kieran.conboy@nuigalway.ie

[1] Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[2] Lero Research Centre & Whitaker Institute, School of Business & Economics, National University of Ireland Galway, Galway, Ireland

data is used, for which purposes and by whom (Newell & Marabelli, 2015). The application of AI in healthcare therefore raises significant concerns of fairness, responsibility, human rights (Floridi & Taddeo, 2016) and can lead to exclusion from essential public services at entirely new levels (Stahl & Markus, 2021).

Yet, the increasing use of AI in healthcare raises questions regarding how to implement and adopt responsible AI practices, which currently is a largely disparate and disconnected field of inquiry. A comprehensive analysis of the intellectual structure of responsible AI for healthcare helps to frame knowledge development work and to set scholars' future research directions (Chen et al., 2019). In the effort to better understand how responsible AI fits into the healthcare context, and what this implies for future research, we conducted a literature review to uncover the most common concerns in utilizing AI in healthcare. We build our understanding of responsible AI through the ethics framework of Mittelstadt et al. (2016). As responsible AI is built on principles of ethics, a framework that adopts a holistic perspective on pertinent issues is deemed as the most suitable way in order to uncover the different relevant facets from a multitude of angles.

Our aim is to provide a synthesis of the most critical issues concerning AI in healthcare and to elaborate a research agenda for future studies. We apply a qualitative systematic literature review (Paré et al., 2015) and rely on its key characteristics such as transparency, replicability, and rigor (Leidner, 2018) to extract most relevant research papers that investigated this area of inquiry. We employed a meta-data analysis to analyse the intellectual structure of the papers selected (Cuccurullo et al., 2016) and to uncover the following research questions:

RQ1. To what extent is current use of AI in healthcare responsible?

RQ2. What important aspects need to be taken into account, and what research questions need to be answered to advance responsible use of AI in healthcare?

From the network and periphery analysis, we identified four types of themes that represent the evolution of the Responsible AI and provided a thematic analysis along four quadrants of the strategic diagram. Then, we reviewed ethical issues emerged from AI in healthcare, based on which we provided a research agenda to guide future studies.

The remainder of the paper is structured as follows. We review the literature on ethical concerns that create the basis for responsible AI in healthcare. We then present our research method followed by meta-data analysis and the synthesis based on the framework developed by Mittelstadt et al. (2016). We conclude with a research agenda to advance responsible approaches for AI in healthcare.

2 Theoretical Background

In this article, we focus on the ethical concerns emerging from AI in digital health based on the six types developed by Mittelstadt et al. (2016), which contribute to developing a responsible AI for healthcare (Dignum, 2019). We use this framework for the synthesis of extant literature.

2.1 Ethical Concerns Stemming from Artificial Intelligence

Ethics has been discussed by philosophers for millennia with the attempt of developing moral statements related to what is good, right, or acceptable (Stahl, 2012). Classical ethical theories developed four main research streams. **Consequentialism** looks at the consequences of an action to determine its ethical status based on the principle maximising the good for most people and minimizing the pain (Davison, 2000). The **deontological ethics** focuses on the rules or processes followed to make a decision regardless of its outcome (Berente et al., 2011) as the rightness of an action is determined by the duty-bond intention of an actor (Chatterjee et al., 2009). This perspective determined the data management plan for conducting research projects. For example, scholars must inform in advance their institutions, the participants, and other stakeholders the purpose of the study, the way the data will be collected, analysed and for how long it will remain stored in specific databases. **Virtue ethics** makes a theoretical distinction between good and bad based on individual's virtues of mind, character, and sense of honesty and not on external aspects of an action (Chatterjee et al., 2009; Gal et al., 2020). Lastly, **pragmatic ethics** rejects any form of absolutism and universality of thought (Davison, 2000) as it assumes there are no universal ethical principles or universal values. Ethical pragmatists acknowledge the existence of the other three normative approaches, but they urge to go beyond them because they are all appropriate but in different contexts.

In addition to these well-established ethical theories, there are more recent approaches specific to technological applications such as information ethics (Floridi, 1999), data ethics (Floridi & Taddeo, 2016), big data ethics (Mittelstadt & Floridi, 2016). Despite a long history of various ethical positions, the current discourse about ethical issues emerging from AI makes little reference to classical ethical theories (Stahl et al., 2021) and relies more on mid-level ethical principles such as biomedical ethics (Mittelstadt & Floridi, 2016), which is concerned to solve practical ethical issues in healthcare. These concerns relate to ensuring that AI do not harm humans and other morally beings and to ensure the moral status of the machines (Bostrom & Yudkowsky, 2014). Most of the high-level interventions into the ethics of AI discussion are principle-based (Floridi & Cowls, 2019) but principles alone cannot guarantee ethical AI (Mittelstadt, 2019). Scholars call

for understanding the ways AI challenges accepted social and ethical norms in several fields such as in healthcare (Mittelstadt et al., 2016). This call is motivated also by AI capacity of tweaking operational parameters and rules, which provided discriminatory results, increased uncertainty about AI decision making process. In response to this, Mittelstadt et al. (2016) developed a map with six types of ethical concerns useful for doing a rigorous diagnosis of ethical concerns emerging from AI in digital health (Fig. 1). We used this map to structure the synthesis of the papers included in this study. Below, we briefly present these dimensions.

Inconclusive evidence refers to the data analysis with inferential statistics and/or machine learning techniques followed to suggest conclusions. The results produce probabilities but also uncertain knowledge, which is not infallible because statistical methods can help identify correlations, but this is not sufficient to posit the existence of a causal connection, which for example might lead to unjustified actions. **Inscrutable evidence** refers to a lack of transparency regarding the data used and a lack of interpretability of how each of the many data-points were used by a machine-learning algorithm contribute to the conclusion it generates. This is the commonly cited 'black-box' issue and can lead to opacity as there are not obvious connections between the data used, how it was used, and its conclusion. **Misguided evidence** refers to the fact that algorithms are subject to a limitation shared by all types of data-processing, which refers to the fact that the output can never exceed the input. Conclusions can only be as reliable (but also as neutral) as the data they are based on. The evidence produced is observer dependent, which can lead to biases. **Unfair outcomes** refer to actions that are based on conclusive, scrutable and well-founded evidence but they have a disproportionate impact on one group of people, which often leads to discrimination. **Transformative effects** refer to algorithmic activities, like profiling that re-ontologise the world by understanding and conceptualising it in new, unexpected ways, triggering and motivating actions based on the insights it generates (Morley et al., 2019). This can lead to challenges for autonomy and informational privacy. **Traceability** refers to problems emerged from the five ethical concerns and it tries to detect the harm caused by algorithmic activity and its cause (Morley et al., 2020). Ethical assessment requires the cause and the responsibility for the harm traced. This can lead to issues with moral responsibility (Tigard, 2020) and thus epistemic and normative ethical issues related to the use of algorithms.

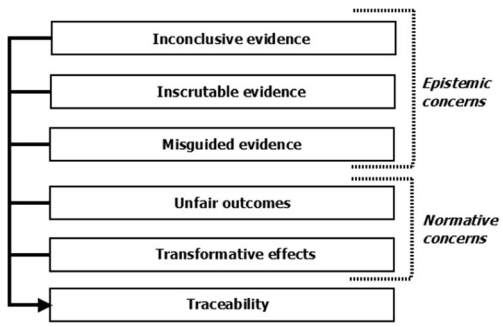

Fig. 1 Six types of ethical concerns raised by algorithms. Source: Mittelstadt et al. (2016)

3 Research Method

Our intellectual structural analysis of Responsible AI for digital health was guided by a systemic literature review for the data collection and sampling and the correspondence analysis, co-word, network and core-periphery analysis for the extraction of research themes. This multimethod data analysis procedure allowed us to spot research gaps and to provide an evidence-based foundation on which to build future research.

3.1 Data Collection and Sampling Procedure

We conducted a systematic literature review (Leidner, 2018; Paré et al., 2015; Schryen et al., 2020) to identify relevant research papers. We followed the guidelines provided by Boell and Cecez-Kecmanovic (2015) and developed a protocol with five subsections, namely *research questions*, *sources searched*, *search terms*, *search strategy*, *inclusion*, and *exclusion criteria*.

In the *first* step, we specified the objective of our review and the road map towards achieving this objective (Templier & Paré, 2015). We investigate ethical concerns emerging from AI in healthcare because it is a key element for Responsible AI, which is engaged with making a proper use of the exchanged information across healthcare organizations. A responsible design of AI increases our trust in the decisions it suggests. An analysis of the most critical ethical issues emerging from AI will allow us to synthesize the intellectual structure of Responsible AI for digital health and to set scholars' future research.

In the *second* step, we identified the pool of journals and databases to extract representative papers. We started to search for research papers in Association for Information Systems "basket of eight" IS journals retrieved from the AIS website www.aisnet.org and leading management journals such as *Academy of Management Journal, Academy of Management Review, Administrative Science Quarterly, Business Ethics Quarterly, Journal of Applied Psychology, Journal of Management, Strategic Management Journal, Organization Science, Information and Organization, Journal of Management Studies, Information and Management*. Then, we continued to search in main online academic databases such as EBSCOhost Business, Searching Interface, Web of Science, Scopus, ACM Digital Library, INFORMS. The maximum coverage of the topic was

achieved with "all databases" option in EBSCO and WOS. Specifically, on Web of Science we searched based on "Topic" for the journals and the AIS electronic library and on "Title", "Abstract", and "Subject" for the conferences. We searched articles published until October 2020.

Third, we focused on papers at the intersection of responsible AI, ethics, and healthcare. To ensure the coverage of potentially relevant search results, we used several variations for Artificial Intelligence (machine learning, algorithms, robots, big data,) for ethics (ethic*, ethical, bioethics, responsible, explainable) and for healthcare (health*, healthcare, care, medical, clinical). The search terms were used with the Boolean "or" operator to ensure that papers that contain these keywords were extracted.

In the *fourth* step we defined our search strategy. We conducted a scoping search to find existing reviews. Then, we searched in selected databases while adding the modifications during the bibliography search to identify key citations for searching further papers through backward and forward reference searching. We searched academic databases and journals to increase the comprehensiveness of the literature review. The search process accumulated a total number of 83 research papers (Fig. 2).

It the *fifth* step, we defined inclusion and exclusion criteria. First, we opted to include papers published in English language that used any methodological approach. Then, peer reviewed academic journals and complete conference papers were preferred for this analysis. Instead, we excluded research in progress, abstracts, workshop proposals, book chapters, demos, and blogs because they are in the exploration phase of the phenomenon. The reason of these restrictions was to exercise quality control on the selected papers. The selection process involved three rounds. In the first phase, we filtered papers from sources searched based on title, keywords, and abstracts. In the second round, we checked whether the keywords were explicitly discussed in the paper. Finally, we conducted forward and backward search of the papers we identified in the second stage. A total number of 34 papers have been selected to respond to our research questions.

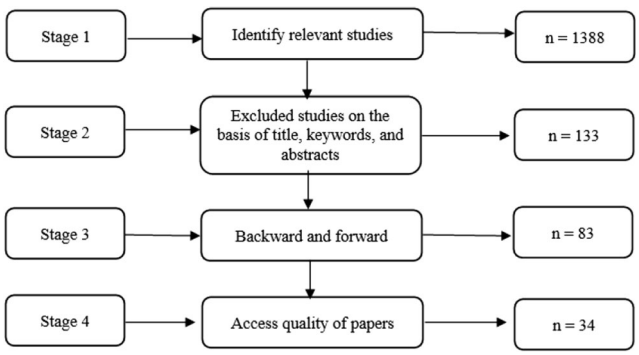

Fig. 2 Stages of the sampling procedure

3.2 Meta-Data Analysis

After having selected the papers to include in our study, we employed a quantified methodology to provide evidence-based insights of the community's research themes (e.g., if they are mature, underdeveloped, emerging, declining, or peripheral), and identify the most studied topics as popular, core, or backbone research topics within the discipline. To do so, we adopted a quantitative analysis, namely co-word analysis, for classifying publications based on the analysis of key-terms from the meta-data of the papers (i.e., author-assigned keywords and machine-extracted key-phrases from abstracts). Co-word analysis has been proposed as a content analysis technique to map the strength of relation between terms in texts and to trace patterns of the associated terms (Callon et al., 1983). The idea behind co-word analysis rests on the assumption that key-terms identified within an article can adequately describe and communicate the content of that article, whilst the co-occurrence of two (or more) key-terms in the same article indicates a linkage between those topics (Callon et al., 1991).

Our dataset consisted of 34 papers published in the time frame 2009–2020; 30 papers had expressively included the keywords (133 author-assigned unique keywords, M = 4.43 keywords per paper). (Fig. 3).

The author-assigned keywords can be potentially biased to human subjectivity (e.g., the authors might use more generic terms to describe their work to ensure its visibility). Thus, the abstracts of the papers were also text-mined to automatically extract from them key-phrases that can describe their contents, based on the "agreement" that the abstract can be seen as a "stand-alone" version of the paper, that synopsizes it in a coherent manner. To extract key phrases from papers' abstracts, we used the TextRank algorithm for text summarization, implemented in Python (Mihalcea & Tarau, 2004; Papamitsiou et al., 2020). TextRank is an extractive and unsupervised text summarization technique that tokenizes and annotates with Part of Speech (PoS). Here, we set the TextRank sliding window to 3, we included nouns (NOUN),

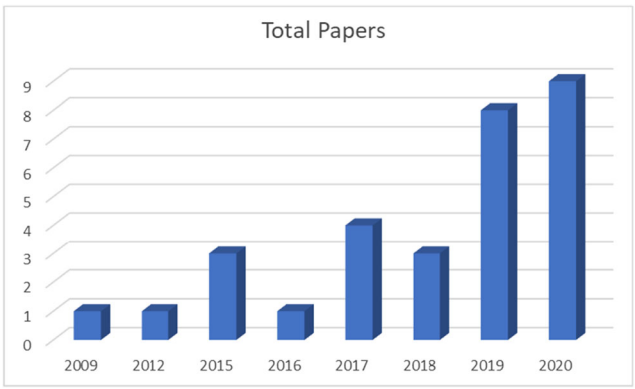

Fig. 3 Number of publications per year (2009–2020)

adjectives (ADJ) and proper nouns (PROPN) as PoS, and we requested for the top-10 phrases.

From the algorithmically performed term extraction, we obtained 158 unique key-phrases (M = 4.65 key-phrases per paper), after manually removing some not highly semantic phrases, such as "findings", "participants", "paper".

Our aim is to identify the most representative research themes (i.e., "hubs") and directions in the information systems field related to ethical aspects of implementing and using AI in healthcare. To find those hubs, a smaller number of highly frequent, i.e., popular terms can be used, as suggested in (Cobo et al., 2011; Liu et al., 2014). The significance of a term in a certain research community is represented in its frequency of use (i.e., the frequency of a keyword is high when more and more researchers are interested in that topic and doing research on it); major research themes can be identified with less than 100 keywords (Liu et al., 2014). Given the limited number of papers considered for analysis, we decided to include the key terms that co-appear more than 2 times (n > = 2) in the considered papers. From the 133 unique keywords, 86 terms appear only once and do not co-appear with other more frequent terms. Thus, from the 133 initial keywords, 47 keywords were used in 27 papers (which describes 90% of our dataset) and considered in our analysis. Similarly, from the 158 unique machine-extracted key-phrases, 57 co-appear more than 2 times and appear in 32 papers, representing 94% of the dataset.

4 Synthesis

In this section, we present the synthesis of the papers selected for this study. First, we present the results of the correspondence, co-word, network, and core-periphery analysis of the intellectual structure of the papers selected. Then, we discuss the ethical concerns stemming from AI and explain their emergence.

4.1 Correspondence Analysis

To develop an initial understanding based on the key-terms (i.e., author-assigned keywords and machine-extracted key-phrases), we applied correspondence analysis (CA), that is suited to graphically and numerically handle categorical data (Greenacre, 2017). We performed CA to plot the overall distribution of topics of interest and how frequently they occur throughout the years.

CA uses a contingency table, i.e., the frequency distribution of years and key-terms, and provides factor scores (coordinates) for both the rows and the columns of the table. In other words, CA decomposes the chi-squared statistic associated with this table into orthogonal factors. These coordinates are used to graphically visualize the association between row and column variables in a two-dimensional space (i.e., a factor map). The results are interpreted based on the relative positions of the points and their distribution along the dimensions; the more words are similar in distribution, the closer they are represented in the map (Cuccurullo et al., 2016). The CA factor map positions the most common key-terms and years on a common set of orthogonal axes. The percentages depicted on the axes correspond to the proportions of the variance in the data that can be explained by the visualization.

4.1.1 Insights from the Correspondence Analysis Factor Map

The correspondence analysis revealed some commonalities between the key-terms assigned by the authors and automatically extracted from abstract, which contributed to cross-checking the findings. Specifically, ethics in AI healthcare emerged as a research interest in 2009, stemming from ethical assessments related to medical information sharing (Fig. 4). On one side, patients were known by many care actors and thus had more opportunities to get treatment, but on the other side were sharing sensitive data without any regulations to protect patients from improper use of data.

In 2012 a new topic emerged related to the morality of artificial agents called also artifactual morality. Scholars raised concerns not only related to machine learning, algorithms, or analytics but also to physical tools such as robots, the use of which increased exponentially. A few years later (2014–2015) scholars created the basis for ethical discourse in digital health and AI with a focus on legal issues, privacy, and their ethical impact on medical research. In 2017, scholars noticed that the issues emerging from medical data sharing and analysis were strictly linked to the structure of the digital devices used for data sharing. Thus, scholars called for ethical designs of Internet of Things (IOT) devices since the tools deeply influenced the ways information was collected, analysed, and visualized by care actors.

Studies between 2018, 2019, and 2020 created an exponential buzz around topics such as black-boxed medicine, data privacy, and data breaches. Scholars extensively investigated the challenges and pitfalls of AI applications in healthcare to understand potential harm and to develop a responsible approach for digital health. The insights from the exploratory correspondence analysis were further investigated with additional, more focused data-analysis methods, namely co-word analysis.

Co-Word Analysis Co-word analysis applies clustering, strategic diagrams, and network analysis to a dataset of terms represented as nodes, and the interactions between terms represented as links (Callon et al., 1991). The terms are clustered into themes according to the correlation matrix of their co-occurrence (e.g., using hierarchical clustering with a distance measurement to maintain content validity and cluster fitness for the highest number of clusters). The relative position of the identified clusters maps the research field using two-

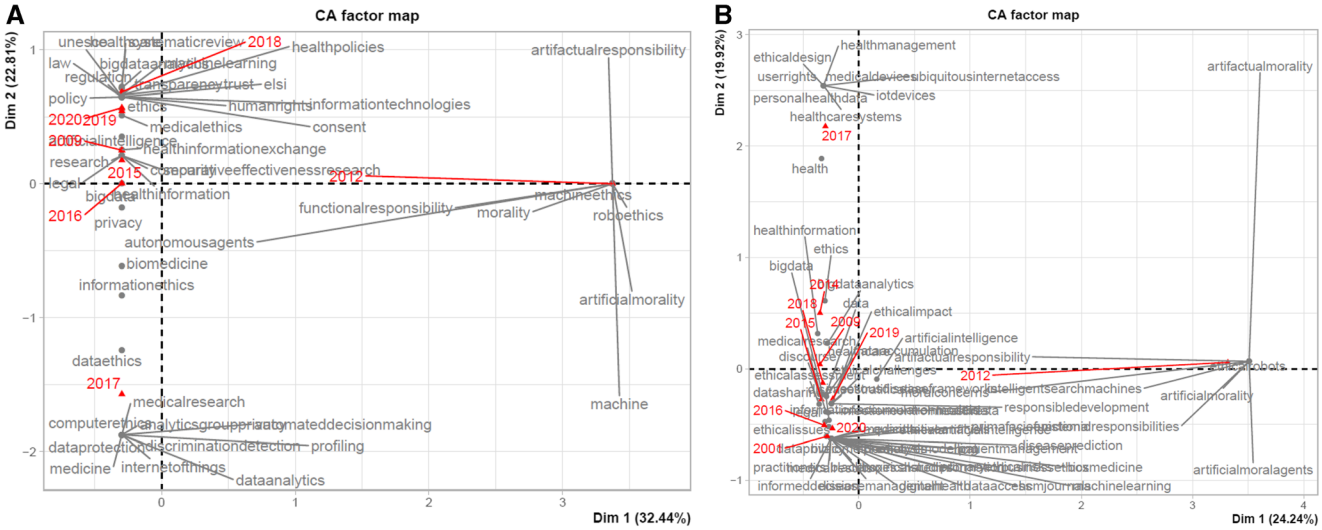

Fig. 4 **a** & **b** - Correspondence analysis (CA) map for ethical concerns stemming from AI in healthcare (2009–2020) (**a**) authorassigned keywords; (**b**) machine-extracted key-phrases

dimensional strategic diagrams (Callon et al., 1991). The positioning is specified using each cluster's centrality (x-axis), i.e., the strength of the links from one term to others, indicating its importance in the development of the field (Liu et al., 2014), and density (y-axis), i.e., the coherence of a cluster and a measure of its internal consistency (He, 1999) – how well the research theme is developed (Fig. 5).

In the strategic diagram, Quadrant I (Q1) contains the mainstream (motor) themes, Quadrant II (Q2) contains themes that are specialized and peripheral to the mainstream work in the field, Quadrant III (Q3) includes themes that are either emerging or disappearing, and Quadrant IV (Q4) covers basic and transversal themes, that hold the potential to become significant.

The co-word network of terms is analysed using the following measures:

- **Key-terms**: subset of terms that constitute a cluster;
- **Size**: number of key-terms in the cluster;
- **Frequency**: how many times all key-terms (in a cluster) appear in the dataset;

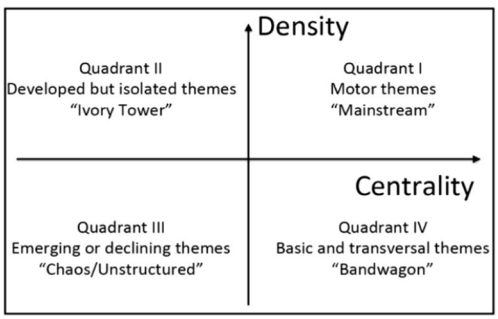

Fig. 5 Strategic diagram of density and centrality (Liu et al., 2014)

- **Co-word frequency**: how many times at-least two key-terms (from a cluster) appear in the same paper. Computing the frequency of two terms appearing together in the same paper results in a symmetrical co-occurrence matrix (Leydesdorff & Vaughan, 2006). In this matrix, values in the diagonal cells are term frequencies, and values in non-diagonal cells are co-word frequencies. High co-occurrence frequency indicates connection between the terms.
- **Centrality**: the degree of interaction of a theme with other parts of the network, i.e., how many other clusters a cluster connects to (Callon et al., 1991); Centrality refers to a group of metrics that aim to quantify the "importance" of a particular node (or cluster) within a network (e.g., betweenness centrality, closeness centrality, eigenvector centrality, degree centrality). Here we used betweenness centrality (C), with $0 \leq C \leq 1$.
- **Density**: how cohesive is the cluster of terms, i.e., the number of direct ties observed for the cluster divided by the maximum number of possible ones (M. Callon et al., 1991). The value range can be any positive number, and can be greater than 1, as density is not "interpreted" as a proportion, but rather as the average number of observed lines (Knoke & Yang, 2019, p. 107).

Based on the clustering results, we plotted the strategic diagram for the both the author-assigned keywords and the machine-extracted key-terms (Fig. 5a and b).

4.1.2 Aligning Authors' Perspectives with Machine's Insights

Clustering analysis of the 47 author-assigned keywords and the 57 machine-extracted key-phrases allowed us to identify

seven clusters in both cases (labelled as C1-C7 and C 1-C 7 respectively), representing the major research themes discussed in the papers we included for this study. The strategic diagrams use the centrality and density of each cluster to help us understand the relative "positions" of these clusters within the overall landscape of ethics stemming from AI in healthcare (Liu et al., 2014; Papamitsiou et al., 2020). In Fig. 5a and b, the axes are centred to the average centrality and density (i.e., 0.258; 1.198 and 0.196; 0.974) of the respective co-word networks. To understand the results, the reader needs to consider the strategic diagram and clusters table together.

As seen from Table 1 and Fig. 6, the analysis of both key-terms datasets yielded quite similar results in terms of what themes have been well developed and are central for the community. One cluster (C4/C 5) appears to be the mainstream theme (i.e., in Q1), and in both cases, that theme covers issues related to bigdata, bigdata analytics and predictive modelling. Furthermore, in both cases, healthcare and ethics create a cluster (C7/C 7) that appears in Q4, i.e., is a basic and transversal theme that has the potential to become mainstream. In addition, issues related to ethical artificial intelligence, primary ethical risks, ethical designs and IoT devices (C6/C 2, C 6) appear to be peripheral topics (i.e., in Q2) that have been well-developed as independent communities and act supportively to the healthcare community.

The only difference shown in the analysis of the two datasets concerns the emerging and declining themes (i.e., in Q3): the analysis of machine extracted key-phrases did not assign any cluster of themes to that quadrant, whilst author-assigned keywords shape themes that either have become trivial and obsolete or they are now starting to attract interest. Those topics are related to information ethics, trust, security, privacy, and health information exchange (C1, C2, C3, C5). The analysis of machine extracted key-phrases assigned those topics to Q2, i.e., identified them as peripheral ones, originating from different research communities. This analysis shows that recently scholars focused on concerns related to security, fairness, and ethics in digital health, which are key components of responsible AI.

Network and Core-Periphery Analysis To better understand the strength of the research themes identified in the diagrams, we visualized their relationship in two granular keyword network maps. Each node in the graphs represents a key-term that is

Table1 **a** & **b** - Clusters of topics related to ethical concerns stemming from AI in healthcare, 2009–2020, (**a**) author-assigned keywords; (**b**) machine-extracted key-phrases, including their quadrant on the strategic diagram (Fig. 6a and b respectively)

ID	Q	Size	Freq	Coword freq	Centrality	Density	Key-terms (the most frequent in bold)
C1	Q3	9	10	101	0,132	1000	consent, elsi, human rights, information technologies, law, policy, regulation, **trust**, unesco
C2	Q3	8	8	56	0,000	1000	artifactual responsibility, artificial morality, autonomous agents, functional responsibility, machine, machine ethics, morality, roboethics
C3	Q3	8	9	83	0,128	1000	**Information ethics**, analytics group privacy, automated decision making, computer ethics, data protection, discrimination detection, medical research, profiling
C4	Q1	2	8	55	0,622	2000	**bigdata**, biomedicine
C5	Q3	6	7	50	0,073	1000	**health information exchange**, comparative effectiveness research, health information, legal, research, security
C6	Q2	4	10	25	0,070	1500	**data ethics**, data analytics, internet of things, medicine
C7	Q4	10	49	132	0,784	0,889	**ethics**, privacy, medical ethics, artificial intelligence, bigdata analytics, healthcare, systematic review, transparency, health policies, machine learning

ID	Q	Size	Freq	Coword freq	Centrality	Density	Key-terms (the most frequent in bold)
C 1	Q2	8	8	64	0,020	1000	artifactual morality, artifactual responsibility, artificial moral agents, artificial morality, ethical robots, functional responsibilities, intelligent search machines, moral concerns
C 2	Q2	9	12	79	0,083	1028	**health**, ethical design, healthcare systems, health management, iot devices, medical devices, personal health data, ubiquitous internet access, user rights
C 3	Q2	8	9	66	0,061	1000	**medical research**, black boxes, box medicine, informed decision, medicine, practitioners, prima facie epistemic, shared information
C 4	Q2	7	7	56	0,040	1000	data accumulation, disease stratification, ethical impact, infection control measures, infectious disease frameworks, information accumulation models, responsible development
C 5	Q1	8	19	96	0,347	1071	**bigdata**, bigdata analytics, disease management, disease prediction, machine learning, medical results, patient management, predictive modelling
C 6	Q2	7	14	73	0,120	1095	**health data**, data privacy, **data sharing**, data access, digital health, ethical artificial intelligence, primary ethical risks
C 7	Q4	10	45	118	0,702	0,622	ethical issues, **healthcare**, health information, artificial intelligence, legal, discourse, ethical assessment, ethics, data, ethical challenges

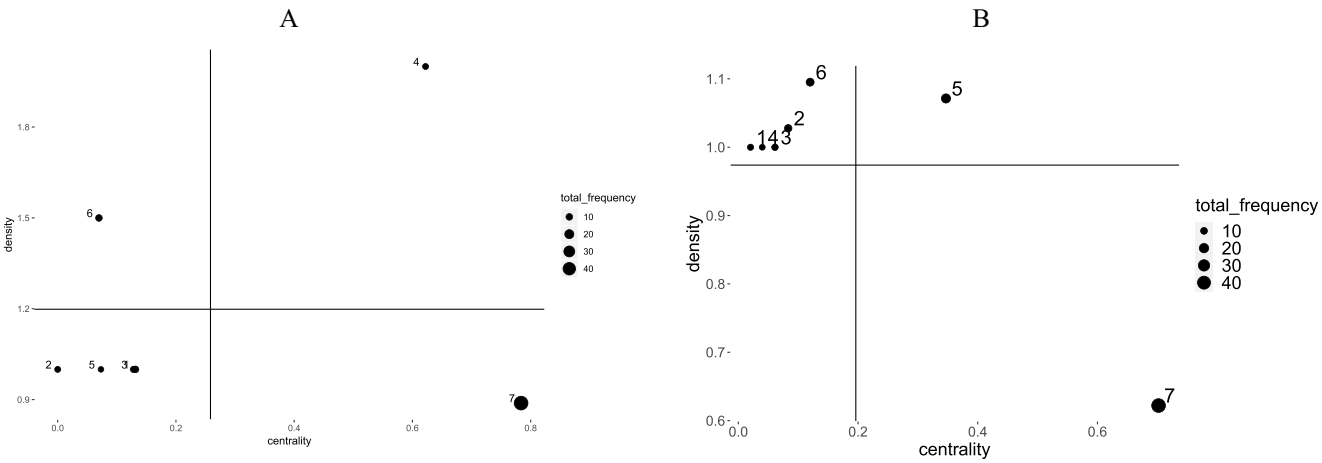

Fig. 6 **a** & **b** - Strategic diagram for ethical concerns stemming from AI in healthcare, 2009–2020, based on (**a**) author-assigned keywords; and (**b**) machine-extracted key-phrases; numbers correspond to cluster IDs in Table 1a and b, respectively

linked to other key-terms that appear in the same paper. The size of the nodes is proportional to the frequency of the key-terms, the colour of the node corresponds to the cluster the key-term has been classified in, and the thickness of the links between nodes is proportional to the co-occurrence correlation for that pair of key-terms. From this analysis, key-terms that appeared less than two times in the initial data set were excluded (as previously explained), and keywords with fewer than three strong ties were excluded to avoid a highly disconnected network.

Figure 7 further confirms previous findings: terms like 'ethics', 'machine learning', 'transparency', 'medical ethics', 'artificial intelligence', 'healthcare' are dominant in both graphs and they have stronger links with the rest of the terms, keeping the network strongly connected, and setting the foundations of a research community. As seen in those two graphs, although the terms identified in the meta-data of the papers are slightly different (i.e., the authors assign keywords to their papers, that are not exactly the same with those that they use in the abstracts of their works), still they point to the same stronger concepts, and capture the bigger picture of this newly raised research area, exhibiting strong interconnectivity. AI appears to have a central role in both aspects of exploring the field; ethics, medical ethics, and ethical issues are also core concepts in both networks. At the same time, slightly different terms are detected in the clusters: e.g., data protection vs. data privacy, IoT devices vs. IoT, machine ethics vs. ethical robots, etc.

Our final analysis identified the core research topics in the field from a whole-network perspective, as individual terms, regardless of the cluster they belong to (this is known as core-periphery analysis). Again, we performed this analysis for both the author-assigned keywords and the machine-extracted key-phrases. The core-periphery analysis yielded

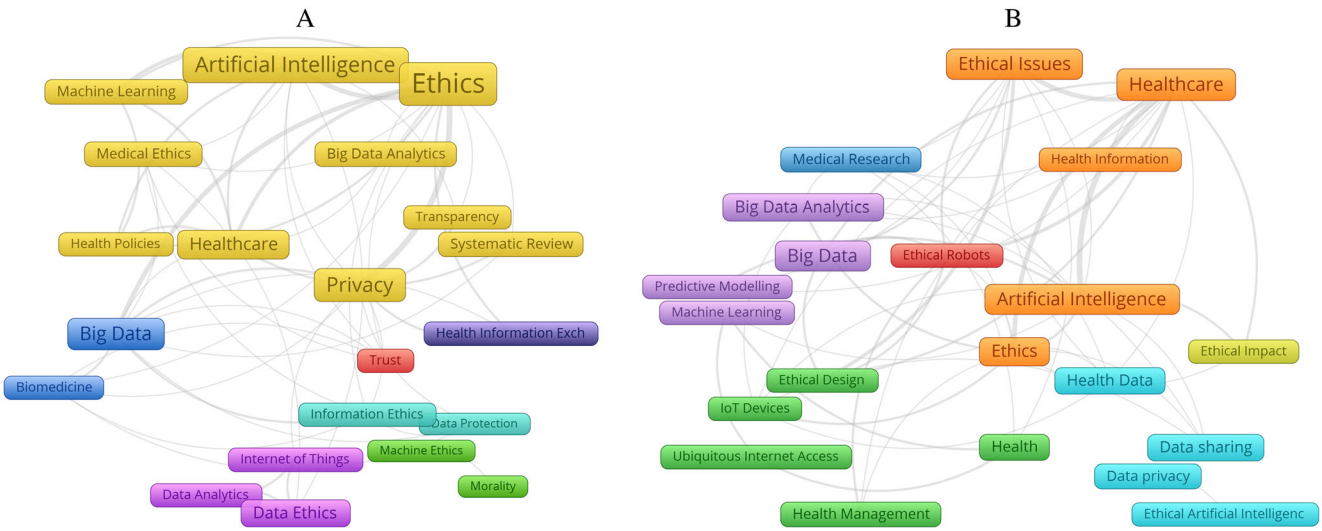

Fig. 7 a & b - Keyword network map for ethical concerns stemming from AI in healthcare2009—2020 based on (**a**) authorassigned keywords; (**b**) machine-extracted key-phrases; each line links two keywords with correlation coefficient ≥ 0.22 and ≥ 0.20 respectively

ten core research topics (terms) in each of the following categories:

- **Popularity**: how frequently a term is used;
- **Coreness**: how connected a term is with other topics; coreness is measured on a [0–1] scale; high coreness value indicates a term that is well connected to other terms.
- **Constraint**: how connected a term is with otherwise distinct terms (i.e., if the term creates a backbone of the field); constraint is measured on a [0–1] scale. High constraint value indicates less structural opportunities a term may have for bridging together otherwise isolated terms, i.e., terms that act as bridges between topics have lower constraint values. Burt's (2004) constraint is commonly used for this purpose.

Table 2a and b synopsize the most popular (high frequency), core (high connection with other topics), and backbone (connection with otherwise isolated topics) thematic areas that emerged during the period 2009–2020. In both tables, six of the most popular themes (identified in bold) are also in the top ten core and backbone themes in the field, suggesting a high consistency between research interests and scientific efforts to maintain the sustainability of the field. One can notice that four out of the six major terms appear in both tables. An interesting note is that although frequent, big data analytics is not a core or backbone term. Further confirming and extending the results from the former analysis, AI appears again to be a driving force on the field.

4.1.3 Thematic Analysis along the Four Quadrants of Strategic Diagram

We identified four types of theme that represent the evolution of the topics investigated by scholars from 2009 till 2020 (Table 1a & b).

Motor Themes (Mainstream – Quadrant 1): Predictive Modelling and Responsible Development According to author assigned keywords, big data, responsible development, ethical impact resulted to be motor theme about ethical issues stemming from AI (custer C4). Indeed, most of the studies discussed the concerns care actors were experiencing while consulting vast amounts of medical information stored in datasets. Authors selected for this study assigned generic terms to categorize their studies in broader research themes, therefore, we also performed an analysis of machine-extracted key-phrases, which resulted to be more specific and fine-grained. From this analysis, we identified terms such as disease prediction, machine learning, predictive modelling, which is represented in cluster C 5. This confirms the keywords assigned by the authors and provides additional information about the motor themes.

Ivory Tower (Developed but Isolated Themes – Quadrant 2): Artificial Morality and Ethical Robots Cluster C6 with keywords as data ethics, Internet of Things, medicine presented well developed and discussed topics according to keywords assigned by the authors. However, they remained somehow isolated from the rest of the discussion as confirmed by the Fig. 4a, where these terms have been mentioned mainly in 2017 by (Mittelstadt, 2017a, 2017b, 2017c).

We identified a more detailed representation of the isolated themes with the analysis of the machine-extracted key-phrases by five main clusters. Cluster C 6 from Table 1b extracted terms as health data, data privacy, data access, ethical AI, primary ethical risks. Whereas cluster C 2 included terms as health, ethical design, healthcare systems, health management, IoT devices, medical devices, personal health data, ubiquitous internet access, user rights. This helped us to understand that authors referred to ethical AI with two perspectives. First, they investigated the primary ethical risks that emerged while sharing and accessing vast databases. Second, they focused on the ethical design of the medical devices used to share and access medical data.

In line with this, cluster C 3 highlighted the emergence of a specific concern related to black-boxed medicine. AI has the potential to support care actors during decision making and knowledge aggregation, however, most of the results are black boxed. When negative consequences emerged from AI-driven decisions, authors called for artificial moral agents, artificial morality, ethical robots (cluster C 1). Lastly, cluster C 4 focused on the ethical impact of infectious disease frameworks, which is concerned with responsible development of AI that can provide suggestions in specific parts of the decision-making process.

Emerging or Declining Theme (Chaos/Unstructured – Quadrant 3): Automated Decision Making and Discrimination Detection From the analysis with the machine-extracted key-phrases, no emerging and declining themes have been identified. Whereas the authors discussed emerging themes such as information ethics, automated decision making, computer ethics, data protection, discrimination detection, medical research, profiling (cluster C3). Scholars were concerned with law, policy, and regulations to protect human rights when using AI for digital health (clusters C1 and C5). This has also been confirmed by the correspondence analysis maps as these terms have been widely used from 2015 to 2020. Although topics related to artificial morality and autonomous agents named also as roboethics, have been developed by some scholars, they remained isolated and turned to be declining themes as they have been mainly investigated in 2012 (Fig. 4a & b).

Basic and Transversal Theme (Bandwagon – Quadrant 4): Transparency, Health Policies, and Ethical Assessment In the last quadrant, basic and transversal themes emerged such as

Table 2 **a** & **b** - Summary of popular, core, and backbone topics of ethical concerns stemming from AI in healthcare, 2009–2020 (**a**) author-assigned keywords; (**b**) machine-extracted key-phrases

A. Author-assigned keywords

#	Frequency		Coreness [0–1]*		Constraint [0–1]**	
1	**Ethics**	12	**Ethics**	0,233	**Ethics**	0,349
2	**Artificial Intelligence**	8	**Privacy**	0,219	**Big Data**	0,360
3	**Privacy**	7	**Big Data**	0,219	**Privacy**	0,486
4	**Big Data**	6	**Healthcare**	0,215	**Healthcare**	0,503
5	**Healthcare**	5	**Artificial Intelligence**	0,212	**Artificial Intelligence**	0,634
6	Data Ethics	4	Health Policies	0,212	**Machine Learning**	0,693
7	Big Data Analytics	3	**Machine Learning**	0,212	Health Policies	0,702
8	**Machine Learning**	3	Health Information Exchange	0,206	Health Information Exchange	0,850
9	Medical Ethics	3	Information Ethics	0,189	Information Ethics	1000
10	Systematic Review	3	transparency	0,189	biomedicine	1000

B. Machine-extracted key phrases

#	Frequency		Coreness [0–1]*		Constraint [0–1]**	
1	**Healthcare**	10	**Healthcare**	0,283	**Healthcare**	0,296
2	**Artificial Intelligence**	9	**Ethical Issues**	0,277	**Ethical Issues**	0,339
3	**Big Data**	8	**Artificial Intelligence**	0,260	**Health data**	0,500
4	**Ethical Issues**	8	**Health data**	0,245	Data sharing	0,500
5	**Ethics**	6	**Big Data**	0,236	**Artificial Intelligence**	0,639
6	Big data Analytics	5	**Ethics**	0,236	**Big Data**	0,798
7	Data sharing	4	Big data Analytics	0,236	**Ethics**	0,856
8	**Health data**	4	Health Information	0,232	Health Information	1000
9	Health	3	Ethical Challenges	0,232	Health	1000
10	Health Management	2	Legal	0,228	Legal	1000

privacy, transparency, health policies, machine learning (cluster C7 from author-assigned keywords). In fact, these key terms have been used by most of the studies to position them as foundation of this phenomenon. The same trend is confirmed by the analysis with machine-extracted key-phrases with words as ethical issues, healthcare, health information, artificial intelligence, legal, discourse, ethical assessment, ethics, data, ethical challenges (cluster C 7).

With co-word analysis, we extracted the main themes that represent the intellectual structure of the topic Responsible AI for digital health and classified them from major to transversal along four quadrants of the strategic diagram.

Ethical Concerns Stemming from Artificial Intelligence in Healthcare Based on the framework by Mittelstadt et al. (2016), we develop a synthesis of the literature that will help understand the current status of literature regarding responsible AI in healthcare. We present a summary of some key points in Table 3, and critically synthesize the extant literature in the sub-sections that follow. Three epistemic concerns address the quality of evidence (inconclusive, inscrutable, and misguided evidence) produced by AI. Two normative concerns (unfair outcomes and transformative effects) focus mainly on the actions itself and the effects they cause on users (patients, healthcare professionals and others). Traceability is a combination of epistemic and normative concerns to debug the harm caused by AI.

4.2 Inconclusive Evidence

To provide data-driven solutions, AI uses inferential statistics and machine learning techniques that can produce uncertain knowledge due to the lack of a causal connection between significant correlations, this calls for an assessment of people's epistemic responsibilities. We identified three main challenges that lead to inconclusive evidence.

First, the data collected with and without patients' direct participation was frequently subject to two types of errors (Burr et al., 2020; Maher et al., 2019). When profiling patients, AI may assign incorrect labels to groups of patients, thus creating a false positive error. For example, patients were occasionally incorrectly categorized as having a disease. On

Table 3 Six types of ethical concerns stemming from AI in healthcare (adapted from Mittelstadt et al., 2016)

Ethical concern		Explanation and consequences	References
Epistemic concerns (quality of evidence)	Inconclusive evidence	Inconclusive evidence refers to conclusions provided by AI based on data analysis with inferential statistics and/or machine learning techniques. The results produce probabilities but also uncertain knowledge, therefore not infallible. Statistical methods can help identify correlations, but this is not sufficient to posit the existence of a causal connection, which can lead to unjustified actions	(Astromskė et al., 2020; Bjerring & Busch, 2020; Crnkovic Dodig & Çürüklü, 2012; Floridi et al., 2020; Garattini et al., 2019; Henriksen & Bechmann, 2020; Morley et al., 2019)
	Inscrutable evidence	Inscrutable evidence refers to a lack of transparency regarding the data used and a lack of interpretability of how each of the many data-points used by a machine-learning algorithm contribute to the conclusion it generates. Not obvious connection between the data used, how it was used, and its conclusion. This is the commonly cited 'black-box' issue and can lead to opacity	(Astromskė et al., 2020; Bjerring & Busch, 2020; Burr et al., 2020; Cohen et al., 2014; Crnkovic Dodig & Çürüklü, 2012; Floridi et al., 2020; Henriksen & Bechmann, 2020; Mittelstadt et al., 2016; Morley et al., 2020; Smith, 2020)
	Misguided evidence	Misguided evidence refers to the fact that algorithms are subject to a limitation shared by all types of data-processing. Conclusions can only be as reliable (but also as neutral) as the data they are based on. The evidence produced is observer dependent, which can lead to bias	(Gray & Thorpe, 2015; Henriksen & Bechmann, 2020; Morley et al., 2019, 2020)
Normative concerns ('fairness' of the action and its effects)	Unfair outcomes	Unfair outcomes refer to actions that are based on conclusive, scrutable and well-founded evidence but it has a disproportionate impact on one group of people, therefore it can lead to discrimination	(Burr et al., 2020; Cohen et al., 2014; Floridi et al., 2019; Garattini et al., 2019; Krutzinna et al., 2019; Maher et al., 2019; Mittelstadt & Floridi, 2016; Morley et al., 2019, 2020)
	Transformative effects	Transformative effects refer to algorithmic activities, like profiling that reontologise the world by understanding and conceptualising it in new, unexpected ways, and triggering and motivating actions based on the insights it generates. This can lead to challenges for autonomy	(Astromskė et al., 2020; Cohen et al., 2014; Henriksen & Bechmann, 2020; Maher et al., 2019; Martin, 2019b; Morley et al., 2019)
Both	Traceability	Traceability refers to problems emerged from the five ethical concerns and it tries to detect the harm caused by algorithmic activity and its cause. Ethical assessment requires the cause and the responsibility for the harm traced. This can lead to issues with informational privacy and moral responsibility	(Cohen et al., 2014; Crnkovic Dodig & Çürüklü, 2012; Krutzinna et al., 2019; Martin, 2019b; Morley et al., 2019; Ocak et al., 2020)

other occasions a false negative occurred when AI incorrectly fails to indicate the presence of a disease when in reality it is present. Additionally, in order to provide these results, AI only scanned the data available, and thus created clusters of patients that excluded the reality outside the database (Astromskė et al., 2020).

Second, AI is used to calculate the most frequent occurrences in the data, whose results are considered evidence-based and are used as sources for decision making (Henriksen & Bechmann, 2020). However, it prone to error over time, and the prediction itself can have negative effects on the automation of knowledge creation and decision making. For example, the size and breadth of the databases may include unrepresentative study groups and may have insufficient statistical power or precision (Gray & Thorpe, 2015; Guan, 2019). In these cases, AI is subject to significant uncertainty, which called for precautionary and safety principles in the data collection and analysis for decision making (Floridi et al., 2020). If practitioners will have an epistemic obligation to rely on AI systems in medical decisions, then policy makers should critically engage in a discussion about the extent to which practitioners should trust AI-driven decisions (Bjerring & Busch, 2020).

Third, AI might be mistakenly considered more objective than people's cognitive abilities due to its computational power. However, this does not necessarily mean that the patterns identified are meaningful because they might suffer from overfitting due to small numbers of samples (Morley et al., 2020). Therefore, the statistics used for calculations might not be sufficient to claim for more objectivity and might lead to erroneous decisions. This issue is strictly linked to the lack of reproducibility, and external validity, of results because AI-

decision making in health are untranslatable between different care settings, which challenges the scientific rigor of these methods (Burr et al., 2020).

4.3 Inscrutable Evidence

AI has the potential to be a good source of evidence when it analyses large datasets and generates results based on this data (Wang et al., 2020). However, when there is a lack of understanding of the exact data used and a lack of transparency and interpretability of the processes followed to generate those results (Floridi et al., 2020), AI provides inscrutable evidence. Several aspects contribute to this concern.

The collection of sensitive information with patients' legal consent is challenging especially when it comes to defining the structure of the 'consent' and the timing of patients' authorization. Informed consents aim to respect the autonomy and human rights of the users (patients, healthcare professionals) involved in projects or medical treatments. This creates the basis for codes of conducts that ideally should be ethical and responsible (Dignum, 2019; Tigard, 2020; Woolley, 2019). However, it is difficult to say in advance how the data will be used by AI and the consequences it will generate. This is also at the heart of the Collindrige dilemma (Mittelstadt et al., 2015; Mittelstadt & Floridi, 2016). AI for digital health challenges current norms related to privacy, confidentiality, and data protection. EU Data Protection Directive in engaged with protecting users with principles such as transparency, legitimacy, and proportionality (Floridi et al., 2019; Kaplan, 2016; Turilli & Floridi, 2009). Although it is clear that principles alone cannot guarantee an ethical and responsible AI for digital health (Mittelstadt, 2019).

The lack of transparency of AI in terms of its content, calculations, and procedures is one of the most discussed concerns, also called "black-boxed medicine" (Astromskė et al., 2020; Crnkovic Dodig & Çürüklü, 2012; Gray & Thorpe, 2015). AI follows complex procedures that are unclear to healthcare professionals, who are not informed about how the data was processed and which protocols have been followed to provide those results (Guan, 2019). An overreliance on AI-driven decisions also challenged the professional role of the physicians, whether to rely on human expertise or on AI suggestions (Morley et al., 2020). It is necessary to be clear about the responsibility and accountability in case of negative effects on patients' health due to opaque results provided by AI (Floridi et al., 2019; Smith, 2020). Additionally, the application of objective metrics limited a deeper analysis, which is usually performed by healthcare professionals who take contextuality, individuality, and equivocality into consideration (Henriksen & Bechmann, 2020; Martin, 2019b).

In response to this, many studies suggested that transparent and controlled approaches for data collection, analysis, and interpretation would solve the black-boxed medicine (Astromskė et al., 2020; Cath et al., 2017; Floridi et al., 2020). If patients are well informed about the way their information will be collected and used during their medical treatments, they will be able to decide their preferences regarding the privacy protection (Noorbakhsh-Sabet et al., 2019). Also, General Data Protection Regulation (GDPR) inserted transparency as one of the central principles to include in the process of sharing users' sensitive information. Although transparency plays an important role in managing health data, it is not enough to protect users from privacy issues. Making the information transparent is costly, time consuming and does not ensure that patients understood it. Therefore, authors call for digital health fiduciaries to protect users during information sharing (Mittelstadt & Floridi, 2016; Woolley, 2019).

4.4 Misguided Evidence

When AI processes medical data, it is subject to several processing limitations. For example, AI results are as reliable as the data they used to provide those results and the evidence AI produces can be misguided, and observer-dependent (Mittelstadt et al., 2016; Morley et al., 2020). Such limitations come from several sources such as the design and implementation phase of AI in organizations, which are highly influenced by the designers' and implementors' values (Gray & Thorpe, 2015; Henriksen & Bechmann, 2020). Biases emerge also from technical constrains and challenges that arise during AI use. Lastly, AI is trained by human experts and consequently it can learn from approaches that are biased from the beginning (Morley et al., 2020). This limitation combined with the biases present in the data used by AI to learn and make suggestions, further reinforces these biases and can be even more harmful for care actors (Schoenberger, 2019).

4.5 Unfair Outcomes

AI-driven actions may result in have a discriminatory effect on minority ethnic communities (Garattini et al., 2019; Morley et al., 2020), which breach the principle of justice, diminish human rights (Martin, 2019b) and lead to unfair outcomes (Burr et al., 2020). Accountability and responsibility principles play a salient role in interpreting unfair outcomes in order to create a responsible approach to healthcare AI (Dignum, 2019).

Accountability refers to the determination of who is responsible for actions taken with AI information (Martin, 2019b). The decision making process is delegated not only to healthcare professional but also to the AI technology used to support these professionals during the decision making process (Guan, 2019; Kaplan, 2016; Smith, 2020). At the intersection of health, technology and law, accountability is associated with the design of AI, the companies that developed AI for specific tasks, the healthcare professions who used AI

during decision making process and the type of liability when AI-driven outcomes cause harm to patients or to the society (Davison, 2000; Schoenberger, 2019). Accountability means understanding the rationale behind the processes followed during decision making (Dignum, 2018; Smith, 2020). Responsibility refers to the role of people when they develop, manufacture, sell and use AI technology. It can be applied both forward, where an entity is in charge of guarantying an intended outcome, and backward to identify the entity that is the appropriate responsible that caused that specific harm(s) (Cath et al., 2017; Morley et al., 2020). It is difficult to identify the causal chain related to the unfair outcome in healthcare for several reasons.

Due to the 'black-boxed' nature of many algorithms, it is challenging to understand the processes followed, the data used, the rules applied, and the people involved (Guan, 2019; Powell, 2019). As a result, patients might be ascribed as morally culpable because the patient did not follow appropriately the medical treatment suggested by the doctor. The ethical burden is shifted to patients. Lastly, data may be biased, where some patients can be considered morally irresponsible because some sensitive information might be less accurate for specific groups of people (minorities), and therefore considered as outliers and excluded (Noorbakhsh-Sabet et al., 2019; Racine et al., 2019).

Martin (2019a) suggests to share the responsibility among all actors involved in data sharing and analysis including the engineers, who developed those AI tools because autonomous robots and other AI technologies behave according to ethical standards and principles inscribed in them by the engineers (Crnkovic Dodig & Çürüklü, 2012). Therefore, the designers of AI also bear a responsibility for the AI functionality (Woolley, 2019).

4.6 Transformative Effects

AI is valued for its capability to identify patterns not visible to human eyes (Henriksen & Bechmann, 2020). However, AI-driven results tend to re-ontologise the world by understanding and conceptualising it in new and unexpected ways, which creates transformative effects and challenges privacy and autonomy (Cohen et al., 2014; Maher et al., 2019; Mittelstadt, 2017b).

When AI makes false negative and false positive errors, these mistakes become part of the dataset used by AI to make suggestion (Martin, 2019b). If these errors are not identified by human experts or AI technology is not taught to detect these errors, the outcomes and recommendations extracted by AI from vast datasets create (negative) transformative effects. Such errors can be further amplified by self-learning or training mechanisms, thus creating a biased cycle of discrimination with little human intervention, which can lead to misdiagnosis or missed diagnosis (Morley et al., 2019).

Thus, there is an urgent need to protect patients and the care actors involved in this process from harm by developing responsible AI for digital health with a clear governance framework (Morley et al., 2020). This can be achieved by considering ethical concerns stemming from AI, which otherwise it can lead to social rejection and/or distorted legislation and policies. Morley et al. (2020) acknowledge also the difficulty in developing a responsible AI and governance frameworks because issues related to privacy, lack of transparency, accessibility and other are not so obvious and difficult to foresee prior to their emergence. The identification of these problems requires input from different disciplines such as computer science, social science, medical science, economics, and others (Guan, 2019).

4.7 Traceability

If detecting erroneous results was difficult due to the complexity and "black box" nature of AI, the traceability of AI results in order to identify the (moral) responsibility of the harm caused was shown to be just as problematic (Guan, 2019; Kaplan, 2016). The most used approach to design AI technology was to support human experts with a limited part of the decision making process, where care actors can control if the profiling and the categorization of patients was correct for a medical diagnosis. Therefore, the final decision is made by doctors and their staff.

Healthcare systems rely on an intertwined series of interactions between humans and AI, which make it very difficult to identify interaction-emerging risks and to allocate liability (Morley et al., 2019). Many people are involved in the use of AI tools for diagnosis for several procedures as organising, collecting, and brokering data, and performing analyses on it. It is extremely difficult to identify each actor's responsibility (Powell, 2019). Therefore, not only are the results of algorithmic-decision making "black boxed" but also the chain of the actors involved in these procedures is extremely complex, making accountability even more difficult (Wearn et al., 2019).

5 Research Agenda

Having described ethical concerns stemming from AI in healthcare, we outline a research agenda following the framework developed by Mittelstadt et al. (2016) for future research (Table 4).

5.1 Inconclusive Evidence

AI has the potential to provide more evidence-based results by taking into consideration a broader range of evidence such as demographic and socioeconomic data, existing diagnosis data,

treatment data, outcome data and others (Morley et al., 2020). Although AI can augment or surpass human abilities by identifying, interpreting, making inferences, and learning from data to achieve predetermined organizational and societal goals (Mikalef & Gupta, 2021), it provides suggestions that are by nature uncertain as it identifies correlation relationships (Floridi et al., 2020). There is a lack of causation between the data used and the results AI provides, which requires medical professionals' involvement (Martin, 2019b). This situation is likely to happen when AI provides suggestions to doctors during decision-making and it uses data collected by other AI tools and human experts, which might diminish the quality of the data. Consequently, healthcare professionals make decisions taking into consideration also AI recommendations that have morally loaded actions and consequences.

This creates new opportunities for future studies by investigating how does AI inform medical professionals during decision-making. Specifically, scholars should focus on how to combine correlation relationships identified by AI with causation relationships elaborated by healthcare professionals to minimize the harm caused by inconclusive evidence. Acknowledging that AI provides superior results when analysing vast amount of data to identify correlation relationships, scholars should uncover what types of tasks are necessary for identifying also causal relationships out of AI recommendations. We believe that a combination of humans' and machines' capabilities might generate new insights about AI involvement in decision-making based on elements of specific situations, which need to be uncovered.

5.2 Inscrutable Evidence

Inscrutable evidence raise concerns about explainability, and transparency of the procedures AI followed to generate results (Barredo Arrieta et al., 2020; Rai, 2020). Additionally, the lack of reproducibility of the results raise questions about scientific rigour (Morley et al., 2020), which increases the difficulty to explain why AI suggests specific actions, which are likely to be black-boxed to healthcare professionals. There is the need to address the explainability issue by investigating how healthcare professionals can achieve better outcomes (in terms of better decisions) and at the same time respect human rights and agency. This perspective will illuminate how to utilize inscrutable evidence provided by AI in healthcare while maintaining safety and explainability. Therefore, scholars should consider not only sociotechnical processes but also social factors that determine AI implementation and in healthcare settings with a specific focus on the role of healthcare professionals, AI developers, implementors and the ways in which responsibilities are shared and managed among multiple stakeholders.

The transparency of the data AI used to generate results and the explainability of results AI suggested deeply influence the ways healthcare professional use and trust AI, however there is a lack of evidence and research on this topic (Burr et al., 2020; Morley et al., 2020). Researchers should study the ways AI influence healthcare professionals' decision making and should inform how they can maintain their own intuition and medical expertise while also leveraging suggestions elaborated by AI. Since AI redefines information processing capabilities and technological design approaches (Martin, 2019a), scholars could investigate to which extent integrate AI in medical decision-making. This creates the moment to rethink how to integrate principles of accountability, responsibility and transparency in the design and development of AI in healthcare in an unobtrusive way.

5.3 Misguided Evidence

Since AI is deeply influenced by the data it uses and analysis to provide suggestions, it faces a well-known limitation where the output of data processing can never exceed the input (quality of data used) (Mittelstadt, 2017a). In line with the "*garbage in, garbage out*" principle, if the data analysed by AI is biased, incomplete or unfair, then also the results provided by AI will have the same limitations and will provided biased suggestions. This calls for a reflection about the neutrality of the data processing, which is observer-dependent (Morley et al., 2020).

Such biases emerge when developers inscribe biased beliefs into AI technology and when AI is trained with datasets that contain noisy data, statistical errors, and others (Henriksen & Bechmann, 2020). This triggers important epistemological and ethical concerns that need to be addressed from the design phase of AI by understanding how to delegate medical decision-making to AI-health solutions, which aspects to consider and what kind of level of interaction is necessary. It is crucial to investigate which ethical considerations are necessary when AI is used to support medical decision-making as healthcare is a complex and well-defined sector with specific rules and procedures to follow. This does not include only quality checks of training datasets but also requires incorporating guidelines for developing responsible approaches for implementing and using AI in healthcare. Furthermore, it is also important to inform the current understanding about how AI biases influence medical decision making to inform design, organizational and IS research about the consequences of AI use in medical practice.

5.4 Unfair Outcomes

When actions driven by AI rely on biased evidence, they can provide unfair outcomes such as discrimination (Mittelstadt et al., 2016). The uncertainty surrounding machine bias has consequences for research investigating the "fairness" of AI results. For example, algorithmic profiling is a method often used to identify correlations or patterns within datasets

Table 4 Research opportunities for developing Responsible AI for digital health

Ethical concerns	Research questions
Inconclusive evidence (inferential statistics, uncertain knowledge)	How does AI inform medical professionals during decision-making? How to combine correlation relationships identified by AI with causation relationships elaborated by healthcare professionals?What types of tasks are appropriate for identifying causal relationships?
Inscrutable evidence (lack of transparency and interpretability, black boxed)	How to utilize inscrutable evidence provided by AI in healthcare while maintaining safety and explainability?How healthcare professionals can maintain their intuition and medical expertise while leveraging AI results?How to integrate principles of accountability, responsibility and transparency in the design and development of AI in healthcare in an unobtrusive way?
Misguided evidence (limitations of data processing)	How to delegate medical decision-making to AI-health solutions? Which aspects to consider? What kind of level of interaction?Which ethical considerations are necessary when AI are used for medical decision-making?How AI biases influence medical decision making?
Unfair outcomes (disproportionate impact on one group of people)	How do we measure the unfairness of AI? Who decides the unfairness? Legislator, clinician? How can AI be guided to commit to fairness and adhere to it during medical decision making?What are the necessary principles to limit potential bias in data training data and in the results provided?
Transformative effects (reontologise things, challenge privacy and autonomy)	How does AI transform the ways through which medical professionals conceptualize patients' information? How does AI transform the content of medical decision-making? How does this transform the collaboration and organization among healthcare professionals?How to prevent potential security breaches that can cause privacy invasion of patients?
Traceability (moral responsibility, accountability)	How to distribute the responsibility of AI results when AI is crucial for medical decision-making?How AI will take responsibility for tasks performed and results suggested?How can AI be controlled once its learning capabilities bring it into states that are only remotely linked to its initial setup?How to reverse-engineer the results elaborated by AI to understand how and why unintended results emerged?

invisible to human eyes, which are used as indicators to classify patients as a member of group. Researchers could investigate how to measure the unfairness of AI results in healthcare, which principles are useful to support such evaluations, and who can decide the unfairness of the suggested provided by AI.

There is a lack of understanding about the ways algorithmic profiling can result in social sorting and harm marginalised groups. The predictions made by AI for each patient are based on proxies extracted at group level, implying that they are not customized on patients' individual characteristics, which can create biased evidence and lead to discrimination (Schoenberger, 2019). Scholars could investigate the processes that lead to discriminatory results and suggest procedures and guidelines to minimize them. This need is motivated also by the fact that such discriminatory practices become self-enforcing with feedback loops, as datasets contain disproportionately data about certain groups of patients, leading to over-monitoring and over-policing of those groups of patients (Lee et al., 2020). Additionally, with an increased complexity of AI, biases will become more sophisticated and difficult to identify, control for, or contest (Racine et al., 2019). Therefore, it is crucial to understand how AI can be guided to commit to fairness and adhere to it during medical decision making and what are the necessary principles to limit potential bias in data training data and in the results provided.

5.5 Transformative Effects

AI influence the ways we conceptualize the world and have transformative effects as AI is increasingly mediating our relationship to reality, our actions and behaviours (Noorbakhsh-Sabet et al., 2019; Tubella et al., 2019). For example, AI recommendations show a sub-part of patients' medical information according to the disease treatment and patients' health conditions. On one side, it aims to extract valuable information for that specific needs, which increases efficiency and efficacy of data processing (Bjerring & Busch, 2020; Stahl et al., 2021). However, on the other side, AI generates risks of data manipulation by cleaning datasets or by excluding information that actually plays an important role; it challenges information privacy and violations of intellectual property rights by limiting patients' access to their data and their ability to understand how their data is being transformed into a recommendations (Mittelstadt & Floridi, 2016).

As these transformations raise a number of ethical concerns, opportunities to address how AI transforms the ways through which medical professionals conceptualize patients' information are continuously emerging. Scholars could focus on the creation process of different content within each group or cluster elaborate by AI according to patients' characteristics. Scholars could study in detail how AI develops classification of behaviour data to inform current understanding

about patients' autonomy protection. Such knowledge could provide new insights for training AI how to 'act ethically' and how to support patients' decisional autonomy. Therefore, more research is needed to investigate how AI transforms the content of medical decision-making and how this transforms the collaboration and organization among healthcare professionals. Lastly, some scholarship suggests examining practices to prevent potential security breaches that can cause privacy invasion of patients (Mittelstadt, 2017b).

5.6 Traceability

The complex procedures followed by AI and its relative opaque results increase the difficulty to identify who should be responsible for the (harmful) consequences of those actions taken based on AI suggestions (Martin, 2019c; Smith, 2020). The close collaboration between human and artificial intelligence during medical decision making calls for important considerations about shared responsibility (Dignum, 2019; Wang et al., 2020). Thus, researchers need to investigate how to distribute the responsibility of AI results when AI is crucial for medical decision-making and how AI will take responsibility for tasks performed and results suggested. Much opportunity for research exists also regarding the ways AI can be controlled once its learning capabilities bring it into states that are only remotely linked to its initial setup. Designer (or developers) and users (healthcare professionals) of AI can be blamed for the harmful results when they have a certain degree of control and intentionality for performing those actions that achieved negative results for patients (Mittelstadt et al., 2016). However, moral responsibility for ethical AI decision-making remains a major question that needs to be addressed by investigating how to reverse-engineer the results elaborated by AI, this will help us to better understand how and why unintended results emerged and decide how to assign the shared responsibility.

In addition to the issues mentioned above, an over-arching concern that affects all capabilities and ethical issues is the absence of any deep theory-driven temporal analysis. Many studies either did not refer to time at all or referred to the increased speed afforded by AI without explicitly articulating exact temporal measures or the source of evidence used to support these claims. In the context of this study, we argue that a debate regarding whether use of AI is ethical or not in a given context often depends on the timing of the AI use and the decision. For example, the amount of time given to digest and fact check an AI-decision may be a significant factor in whether the decision is ethical. Similarly, the overall temporal range of data (in days, weeks, months, or years) used by the AI in that decision may also be a factor.

A particular issue in the context of health and particularly the current pandemic is that the temporal personalities and perceptions of the population or health officials did not feature in any study as far as we are aware. In such an exogenous health shock, the vulnerable, the elderly, and the ill often have a perception of time that is very different from the average. Waiting for a result, or for subsequent treatment, or for the illness to pass can cause time to pass incredibly slowly. Others have a temporal personality that exhibits this feeling of slow passage of time even when there is no logical reason (Ancona et al., 2001; Mosakowski & Earley, 2000; Orlikowski & Yates, 2002). There was no evidence in any of the research that AI applications and the research that contained them either considered or helped this aspect of temporal complexity.

6 Limitations and Future Work

Our study has the following limitations. First, our focus is on responsible approaches for AI development and implementation in healthcare, which is a specific setting with its peculiarities. Although healthcare is a vast industry where the implementation of AI presents common challenges and difficulties with other industries, future studies might consider investigating also other industries such as transportation, law, manufacturing, communication, and others. An inter-industrial analysis of the challenges and key characteristics that emerge from the AI implementation in specific contexts will provide new insights into strategies for increasing AI implementation and use in those settings. Second, our data analysis followed a well-established framework developed by Mittelstadt et al. (2016), which helped us to systematize the current knowledge about responsible AI in health. At the same time, this choice limited the dimensions to consider when analysing this phenomenon. Future scholars might consider other frameworks or core values such as compliance, acceptability, proactivity, reflexivity as suggested by (Stahl & Markus, 2021). Lastly, while we relied on most advanced technologies grouped under the term Artificial Intelligence, we did not include other technologies from which medical data is collected, used, and shared such as Electronic Healthcare Records. It will be beneficial to better understand the interaction and dependences of AI with other technologies in multiple industries.

7 Conclusions

For this study, we reviewed the most discussed ethical concerns that emerged from AI in healthcare. First, we presented the ethical concerns emerging from AI in digital health based on the six types developed by Mittelstadt et al. (2016), which contribute to developing a responsible AI for healthcare (Dignum, 2019). Next, we explained how

the epistemic and normative concerns emerged in healthcare research. Based on this review, we provided a research agenda for future studies. We contribute by providing insights into research themes in this growing research field, especially from the point of view of IS and health informatics scholars.

In the attempt to understand key components of a responsible AI for digital health, we should not ignore potential benefits from analysing vast and small datasets for medical diagnosis and patient monitoring but at the same time we need to be aware of the harm AI might cause to patients and other care actors and how to behave in those situations. The ethical concerns discussed here helps care actors to make a diagnosis ethical issues in future discourses for developing responsible approaches for AI in healthcare.

Funding This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital).

References

Ancona, D. G., Goodman, P. S., Lawrence, B. S., & Tushman, M. L. (2001). Time: A new research lens. *Academy of Management Review, 26*(4), 645–663.

Astromskė, K., Peičius, E., & Astromskis, P. (2020). Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. *AI & SOCIETY.* https://doi.org/10.1007/s00146-020-01008-9.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Berente, N., Gal, U., & Hansen, S. (2011). Ethical implications of social stratification in information systems research. *Information Systems Journal, 21*(4), 357–382. https://doi.org/10.1111/j.1365-2575.2010.00353.x.

Bjerring, J. C., & Busch, J. (2020). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00391-6.

Boell, S. K., & Cecez-Kecmanovic, D. (2015). On being 'systematic'in literature reviews. In *Formulating research methods for information systems* (pp. 48–78). Springer.

Bostrom, N., & Yudkowsky, E. (2014). *The ethics of artificial intelligence*.

Burr, C., Taddeo, M., & Floridi, L. (2020). The ethics of digital well-being: A thematic review. *Science and Engineering Ethics, 26*(4), 2313–2343. https://doi.org/10.1007/s11948-020-00175-8.

Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology, 110*(2), 349–399. https://doi.org/10.1086/421787.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics, 22*(1), 155–205. https://doi.org/10.1007/BF02019280.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235. https://doi.org/10.1177/053901883022002003.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2017). Artificial intelligence and the 'good society': The US, EU, and UK approach. *Science and Engineering Ethics.* https://doi.org/10.1007/s11948-017-9901-7.

Chatterjee, S., Sarker, S., Washington State University, U.S.A., Fuller, M., & Washington State University, U.S.A. (2009). A deontological approach to designing ethical collaboration. *Journal of the Association for Information Systems*, *10*(3), 138–169. https://doi.org/10.17705/1jais.00190

Chen, L., Baird, A., Georgia State University, USA, Straub, D., & Temple University, USA. (2019). An Analysis of the Evolving Intellectual Structure of Health Information Systems Research in the Information Systems Discipline. *Journal of the Association for Information Systems*, 1023–1074. https://doi.org/10.17705/1jais.00561

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382–1402.

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs, 33*(7), 1139–1147. https://doi.org/10.1377/hlthaff.2014.0048.

Crnkovic Dodig, G., & Çürüklü, B. (2012). Robots: Ethical by design. *Ethics and Information Technology, 14*(1), 61–71. https://doi.org/10.1007/s10676-011-9278-2.

Cuccurullo, C., Aria, M., & Sarto, F. (2016). Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains. *Scientometrics, 108*(2), 595–611.

Davison, R. (2000). Professional ethics in information systems: A personal perspective. *Communications of the Association for Information Systems, 3*(8).

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1–3. https://doi.org/10.1007/s10676-018-9450-z.

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

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology, 1*(1), 33–52.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering*

Ethics, 26(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5.

Floridi, L., Luetge, C., Pagallo, U., Schafer, B., Valcke, P., Vayena, E., Addison, J., Hughes, N., Lea, N., Sage, C., Vannieuwenhuyse, B., & Kalra, D. (2019). Key ethical challenges in the European medical information framework. *Minds and Machines, 29*(3), 355–371. https://doi.org/10.1007/s11023-018-9467-4.

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences, 374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360.

Gal, U., Jensen, T. B., & Stein, M.-K. (2020). Breaking the vicious cycle of algorithmic management: A virtue ethics approach to people analytics. *Information and Organization, 30*(2), 100301. https://doi.org/10.1016/j.infoandorg.2020.100301.

Garattini, C., Raffle, J., Aisyah, D. N., Sartain, F., & Kozlakidis, Z. (2019). Big data analytics, infectious diseases and associated ethical impacts. *Philosophy & Technology, 32*(1), 69–85. https://doi.org/10.1007/s13347-017-0278-y.

Gray, E. A., & Thorpe, J. H. (2015). Comparative effectiveness research and big data: Balancing potential with legal and ethical considerations. *Journal of Comparative Effectiveness Research, 4*(1), 61–74. https://doi.org/10.2217/cer.14.51.

Greenacre, M. (2017). *Correspondence analysis in practice: Vol. CRC press* (CRC press; CRC press). CRC press; CRC press. CRC press.

Guan, J. (2019). Artificial intelligence in healthcare and medicine: Promises, ethical challenges and governance. *Chinese Medical Sciences Journal, 34*(2), 76–83.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*(1), 133–159.

Henriksen, A., & Bechmann, A. (2020). Building truths in AI: Making predictive algorithms doable in healthcare. *Information, Communication & Society, 23*(6), 802–816. https://doi.org/10.1080/1369118X.2020.1751866.

Kaplan, B. (2016). How should health data be used?: Privacy, secondary use, and big data sales. *Cambridge Quarterly of Healthcare Ethics, 25*(2), 312–329. https://doi.org/10.1017/S0963180115000614.

Knoke, D., & Yang, S. (2019). *Social network analysis* (Vol. 154). Sage Publications.

Krutzinna, J., Taddeo, M., & Floridi, L. (2019). Enabling posthumous medical data donation: An appeal for the ethical utilisation of personal health data. *Science and Engineering Ethics, 25*(5), 1357–1387. https://doi.org/10.1007/s11948-018-0067-8.

Lee, M. S. A., Floridi, L., & Singh, J. (2020). From fairness metrics to key ethics : A Context-Aware Approach to Algorithmic Ethics in an Unequal Society. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3679975.

Leidner, D. E. (2018). Review and theory symbiosis: An introspective retrospective. *Journal of the Association for Information Systems, 19*(6), 1.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the web environment. *Journal of the American Society for Information Science and Technology, 57*(12), 1616–1628.

Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S., & Kostakos, V. (2014). CHI 1994-2013: Mapping two decades of intellectual progress through co-word analysis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3553–3562.

Maher, N. A., Senders, J. T., Hulsbergen, A. F. C., Lamba, N., Parker, M., Onnela, J.-P., Bredenoord, A. L., Smith, T. R., & Broekman, M. L. D. (2019). Passive data collection and use in healthcare: A systematic review of ethical issues. *International Journal of Medical Informatics, 129*, 242–247. https://doi.org/10.1016/j.ijmedinf.2019.06.015.

Martin, K. (2019a). Designing ethical algorithms. *MIS Quarterly Executive*, 129–142. https://doi.org/10.17705/2msqe.00012.

Martin, K. (2019b). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 160*(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411. https://www.aclweb.org/anthology/W04-3252

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management, 58*(3), 103434. https://doi.org/10.1016/j.im.2021.103434.

Mittelstadt, B. D. (2017a). Designing the health-related internet of things: Ethical principles and guidelines. *Information, 8*(3), 77. https://doi.org/10.3390/info8030077.

Mittelstadt, B. D. (2017b). Ethics of the health-related internet of things: A narrative review. *Ethics and Information Technology, 19*(3), 157–175. https://doi.org/10.1007/s10676-017-9426-4.

Mittelstadt, B. D. (2017c). From individual to group privacy in big data analytics. *Philosophy & Technology, 30*(4), 475–494. https://doi.org/10.1007/s13347-017-0253-7.

Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679.

Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics, 22*(2), 303–341. https://doi.org/10.1007/s11948-015-9652-2.

Mittelstadt, B. D., Stahl, B. C., & Fairweather, N. B. (2015). How to shape a better future? Epistemic difficulties for ethical assessment and anticipatory governance of emerging technologies. *Ethical Theory and Moral Practice, 18*(5), 1027–1047. https://doi.org/10.1007/s10677-015-9582-8.

Morley, J., Machado, C., Burr, C., Cowls, J., Taddeo, M., & Floridi, L. (2019, November 13). The debate on the ethics of AI in health care: A reconstruction and critical review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3486518.

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine, 260*, 113172. https://doi.org/10.1016/j.socscimed.2020.113172.

Mosakowski, E., & Earley, P. C. (2000). A selective review of time assumptions in strategy research. *Academy of Management Review, 25*(4), 796–812.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems, 24*(1), 3–14. https://doi.org/10.1016/j.jsis.2015.02.001.

Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., & Abedi, V. (2019). Artificial intelligence transforms the future of health care. *The American Journal of Medicine, 132*(7), 795–801. https://doi.org/10.1016/j.amjmed.2019.01.017.

Ocak, S., Köseoglu, M. A., & Yildiz, M. (2020). Business ethics research in healthcare management: A systematic review. *International Journal of Healthcare Management, 13*(2), 170–176. https://doi.org/10.1080/20479700.2017.1336882.

Orlikowski, W. J., & Yates, J. (2002). It's about time: Temporal structuring in organizations. *Organization Science, 13*(6), 684–700.

Papamitsiou, Z., Giannakos, M., Simon, −, & Luxton-Reilly, A. (2020). Computing education research landscape through an analysis of keywords. *Proceedings of the 2020 ACM Conference on*

International Computing Education Research, 102–112. https://doi.org/10.1145/3372782.3406276

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management, 52*(2), 183–199.

Powell, J. (2019). Trust me, I'm a Chatbot: How artificial intelligence in health care fails the Turing test. *Journal of Medical Internet Research, 21*(10), e16222. https://doi.org/10.2196/16222.

Racine, E., Boehlen, W., & Sample, M. (2019). Healthcare uses of artificial intelligence: Challenges and opportunities for growth. *Healthcare Management Forum, 32*(5), 272–275. https://doi.org/10.1177/0840470419843831.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5.

Sambasivan, N., & Holbrook, J. (2018). Toward responsible AI for the next billion users. *Interactions, 26*(1), 68–71.

Schoenberger, D. (2019). Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology, 27*(2), 142–170. https://doi.org/10.1093/ijlit/eaz002.

Schryen, G., Wagner, G., Benlian, A., & Paré, G. (2020). A knowledge development perspective on literature reviews: Validation of a new typology in the IS field. *Communications of the AIS*, *46*(7), 134–186. https://doi.org/10.17705/1CAIS.04607.

Smith, H. (2020). Clinical AI: Opacity, accountability, responsibility and liability. *AI & SOCIETY*. https://doi.org/10.1007/s00146-020-01019-6.

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., Patel, A., Ryan, M., & Wright, D. (2021). Artificial intelligence for human flourishing – Beyond principles for machine learning. *Journal of Business Research, 124*, 374–388. https://doi.org/10.1016/j.jbusres.2020.11.030.

Stahl, B. C. (2012). Morality, ethics, and reflection: A categorization of normative IS research. *Journal of the Association for Information Systems, 13*(8), 1.

Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems, 86*, 152–161. https://doi.org/10.1016/j.robot.2016.08.018.

Stahl, Bernd Carsten, & Markus, M. L. (2021). Let's claim the authority to speak out on the ethics of smart information systems. MIS Quarterly. Special Issue*:* . https://dora.dmu.ac.uk/handle/2086/20578

Templier, M., & Paré, G. (2015). A framework for guiding and evaluating literature reviews. *Communications of the Association for Information Systems, 37*(1), 6.

Tigard, D. W. (2020). Responsible AI and moral responsibility: A common appreciation. *AI and Ethics*. https://doi.org/10.1007/s43681-020-00009-0.

Tubella, A. A., Theodorou, A., Dignum, V., & Dignum, F. (2019). Governance by glass-box: Implementing transparent moral bounds for AI behaviour. *ArXiv:1905.04994 [Cs]*. http://arxiv.org/abs/1905.04994

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology, 11*, 105–112. https://doi.org/10.1007/s10676-009-9187-9.

Wang, Y., Xiong, M., & Olya, H. G. T. (2020). Toward an understanding of responsible artificial intelligence practices. In *Hawaii international conference on system sciences* (p. 10).

Wearn, O. R., Freeman, R., & Jacoby, D. M. P. (2019). Responsible AI for conservation. *Nature Machine Intelligence, 1*(2), 72–73. https://doi.org/10.1038/s42256-019-0022-7.

Woolley, J. P. (2019). Trust and justice in big data analytics: Bringing the philosophical literature on trust to bear on the ethics of consent. *Philosophy & Technology, 32*(1), 111–134. https://doi.org/10.1007/s13347-017-0288-9.

Cristina Trocin is a postdoctoral researcher with the ERCIM "*Alain Bensoussan*" Fellowship at the Norwegian University of Science and Technology (NTNU). She is doing research at the intersection of technology, work, and organizational change. Currently, her research projects include the development and introduction of Artificial Intelligence (AI) in healthcare organizations, in Human Resource Management, future of work with a sociomaterial approach. She holds a PhD in Management from Ca′ Foscari University of Venice, Italy. Her research has been presented at international conferences such as *Academy of Management (AOM), International Conference on Information Systems (ICIS), European Group for Organizational Studies (EGOS)*.

Patrick Mikalef is an Associate Professor in Data Science and Information Systems at the Department of Computer Science. He has been a Marie Skłodowska-Curie post-doctoral research fellow working on the research project "*Competitive Advantage for the Data-driven Enterprise*" (CADENT). He received his B.Sc. in Informatics from the Ionian University, his M.Sc. in Business Informatics for Utrecht University, and his Ph.D. in IT Strategy from the Ionian University. His research interests focus on the strategic use of data science and information systems in turbulent environments. He has published work in international conferences and peer reviewed journals including the *European Journal of Information Systems, British Journal of Management, Information and Management, European Journal of Operational Research, and Information Systems Frontiers*.

Zacharoula Papamitsiou is a senior researcher at the Computer Science Department of NTNU. She holds a Ph.D. in adapting the learning services for supporting learners' decision-making using learning analytics. Her research interests include used modeling, learner-computer interaction, and autonomous learning. She has published articles in ranked international journals including *Computers in Human Behavior, British Journal of Educational Technology, Journal of Computer Assisted Learning, IEEE Transactions on Learning Technologies, and Educational Technology Research and Development,* as well as in international conferences such as the *ACM UMAP, LAK, ACM ICER*. She is a recipient of the ERCIM fellowship.

Kieran Conboy is a Professor in Information Systems and leader of the Lero research group at NUI Galway. He previously worked for Accenture Consulting and the University of New South Wales in Australia. He has collaborated with many organizations such as Atlassian, Cisco Systems, Suncorp, and Fidelity Investments and many SMEs. He has published over 200 papers in leading international journals and conferences including *Information Systems Research, ISE, TOSEM, the European Journal of Information Systems, IEEE Software, and IEEE Computer*.