Journal Pre-proof

The Ethics of AI in Health Care: a Mapping Review

Jessica Morley, Caio C.V. Machado, Christopher Burr, Josh Cowls, Indra Joshi, Mariarosaria Taddeo, Luciano Floridi

PII: S0277-9536(20)30391-9

DOI: https://doi.org/10.1016/j.socscimed.2020.113172

Reference: SSM 113172

To appear in: *Social Science & Medicine*

Revised Date: 22 June 2020

Accepted Date: 23 June 2020

Please cite this article as: Morley, J., Machado, C.C.V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., Floridi, L., The Ethics of AI in Health Care: a Mapping Review, *Social Science & Medicine*, https://doi.org/10.1016/j.socscimed.2020.113172.

The Ethics of AI in Health Care: a Mapping Review

Abstract

This article presents a mapping review of the literature concerning the ethics of artificial intelligence (AI) in health care. The goal of this review is to summarise current debates and identify open questions for future research. Five literature databases were searched to support the following research question: 'how can the primary ethical risks presented by AI-health be categorised, and what issues must policymakers, regulators and developers consider in order to be 'ethically mindful?'. A series of screening stages were carried out—for example, removing articles that focused on digital health in general (e.g. data sharing, data access, data privacy, surveillance/nudging, consent, ownership of health data, evidence of efficacy)—yielding a total of 156 papers that were included.

Ethical issues can be (a) epistemic, related to misguided, inconclusive or inscrutable evidence; (b) normative, related to unfair outcomes and transformative effectives; or (c) related to traceability. We further find that these ethical issues arise at six levels of abstraction: individual, interpersonal, group, institutional, and societal or sectoral. Finally, we outline a number of considerations for policymakers and regulators, mapping these to existing literature, and categorising each as epistemic, normative or traceability-related and at the relevant level of abstraction. Our goal is to inform policymakers, regulators and developers of what they must consider if they are to enable health and care systems to capitalise on the dual advantage of ethical AI; maximising the opportunities to cut costs, improve care, and improve the efficiency of health and care systems, whilst proactively avoiding the potential harms. We argue that if action is not swiftly taken in this regard, a new 'AI winter' could occur due to chilling effects related to a loss of public trust in the benefits of AI for health care.

The Ethics of AI in Health Care: a Mapping Review

Abstract

This article presents a mapping review of the literature concerning the ethics of artificial intelligence (AI) in health care. The goal of this review is to summarise current debates and identify open questions for future research. Five literature databases were searched to support the following research question: how can the primary ethical risks presented by AI-health be categorised, and what issues must policymakers, regulators and developers consider in order to be 'ethically mindful?. A series of screening stages were carried out—for example, removing articles that focused on digital health in general (e.g. data sharing, data access, data privacy, surveillance/nudging, consent, ownership of health data, evidence of efficacy)—yielding a total of 156 papers that were included in the review.

We find that ethical issues can be (a) epistemic, related to misguided, inconclusive or inscrutable evidence; (b) normative, related to unfair outcomes and transformative effectives; or (c) related to traceability. We further find that these ethical issues arise at six levels of abstraction: individual, interpersonal, group, institutional, and societal or sectoral. Finally, we outline a number of considerations for policymakers and regulators, mapping these to existing literature, and categorising each as epistemic, normative or traceability-related and at the relevant level of abstraction. Our goal is to inform policymakers, regulators and developers of what they must consider if they are to enable health and care systems to capitalise on the dual advantage of ethical AI; maximising the opportunities to cut costs, improve care, and improve the efficiency of health and care systems, whilst proactively avoiding the potential harms. We argue that if action is not swiftly taken in this regard, a new 'AI winter' could occur due to chilling effects related to a loss of public trust in the benefits of AI for health care.

Keywords

Artificial Intelligence; Ethics; Healthcare; Health Policies; Machine Learning.

1. Introduction

Healthcare systems across the globe are struggling with increasing costs and worsening outcomes (Topol, 2019). This presents those responsible for overseeing healthcare systems with a 'wicked problem', meaning that the problem has multiple causes, is hard to understand and define, and hence will have to be tackled from multiple different angles. Against this background, policymakers, politicians, clinical entrepreneurs and computer and data scientists increasingly argue that a key part of the solution will be Artificial Intelligence (AI), particularly Machine Learning (Chin-Yee & Upshur, 2019). The argument stems not from the belief that all healthcare needs will soon be taken care of by "robot doctors" (Chin-Yee & Upshur, 2019). Instead, the argument rests on the classic definition of AI as an umbrella term for a range of techniques that can be used to make machines complete tasks in a way that would be considered intelligent *were* they to be completed by a human. For example, as mapped by (Harerimana et al., 2018), decision tree techniques can be used to diagnose breast cancer tumours (Kuo et al., 2001); Support Vector Machine techniques can be used to classify genes (Brown et al., 2000) and diagnose Diabetes Mellitus (Barakat et al., 2010); ensemble learning methods can predict outcomes for cancer patients (Kourou et al., 2015); and neural networks can be used to diagnose stroke (Jiang F, Jiang Y, Zhi et al 2017). From this perspective, AI represents a growing *resource* of *interactive*, *autonomous*, and often *self-learning* (in the machine learning sense) *agency*, that can be used on demand (Floridi, 2019), presenting the opportunity for potentially transformative cooperation between machines and doctors (Bartoletti, 2019).

If harnessed effectively, such AI-clinician cooperation, where AI is used to provide comprehensive evidence-based clinical decision-support to the clinician (AI-Health), could offer great opportunities for the improvement of healthcare services and ultimately patients' health (Taddeo & Floridi, 2018) by significantly improving human clinical capabilities in diagnosis (Arieno et al., 2019; De Fauw et al., 2018; Kunapuli et al., 2018), drug discovery (Álvarez-Machancoses & Fernández-Martínez, 2019; Fleming, 2018), epidemiology (Hay et al., 2013), personalised medicine (Barton et al., 2019; Cowie et al., 2018; Dudley et al., 2015) and operational efficiency (H. Lu & Wang, 2019; Nelson et al., 2019). However, as Ngiam & Khor (2019) stress, if these AI solutions are to be embedded in clinical practice, then a clear governance framework is needed to protect people from harm, including harm arising from unethical conduct. We use the term 'cooperation' here and suggest that AI will be chiefly used for clinical decision support. This differentiates from arguments often made by the popular press which suggest that AI will be used to 'replace' clinicians.

To support policymakers, the task of the following pages is to classify the ethical risks presented by AI-health, align these with specific questions that must be answered by policymakers, and provide example actions that could be taken by healthcare governing bodies to develop the requisite governance framework. The intention is to ensure that the ethical challenges raised by implementing AI in healthcare settings are tackled *proactively* (Char et al., 2018). We seek to do this because if the ethical risks are not tackled proactively, by encouraging AI-health policymakers, developers and regulators to be ethically mindful, there is a potential risk of incurring significant opportunity costs (Cookson, 2018). For instance, ethical mistakes or misunderstandings may lead to social rejection and/or distorted legislation and policies, which in turn cripple the acceptance and advancement of [the necessary] data science. Encouraging this kind of proactive ethical analysis is essential but also challenging because, although bioethical principles for clinical research and healthcare are well established, and issues related to privacy, effectiveness, accessibility and utility are clear (Nebeker et al., 2019), other issues are less obvious (Char et al., 2018). For example, AI processes may lack transparency, making accountability problematic, or may be biased, leading to unfair, discriminatory behaviour or mistaken decisions (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). Identification of these less obvious concerns requires input from the medical sciences, economics, computer sciences, social sciences, law, and policy-making. Yet, research in these areas is currently happening in siloes, is overly focused on individual level impacts (Morley & Floridi, 2020a), or does not consider the fact that the ethical concerns may vary depending on the stage of the algorithm development pipeline (Morley et al., 2019).

Whilst AI-Health remains in the early stages of development and relatively far away from having a major impact on frontline clinical care (Panch, Mattie, & Celi, 2019), there is still time to develop this framework. However, this window of opportunity is closing fast, as the pace at which AI-Health solutions are gaining approval for use in clinical care in the US is accelerating (Topol, 2019). Both the Chinese (Zhang et al., 2018) and British governments (Department of Health and Social Care, 2019) have made it very clear that they intend on investing heavily in the spread and adoption of AI-Health technologies. It is for these reasons that the goal of this article is to offer a cross-disciplinary mapping review of the potential ethical implications of the development of AI-Health in order to support policy discussion, which will in turn orient the development of better design practices, and transparent and accountable deployment strategies. We will do this in terms of digital ethics. That is, we will focus on the evaluation of moral problems related to data, algorithms and corresponding practices (Floridi & Taddeo, 2016), with the hope of enabling governments and healthcare systems looking to adopt AI-Health to be ethically mindful (Floridi, 2019a). Specifically, the research question is: how can the primary ethical risks presented by AI-health be categorised, and what issues must policymakers, regulators and developers consider in order to be ethically mindful?

2. Methodology

A mapping review methodology (Grant & Booth, 2009) was used to find literature from across disciplinary boundaries that highlighted ethical issues *unique* to the use of AI algorithms in healthcare. This type of review is used to map and categorise existing literature on a particular topic (in this case the ethics of AI) and contextualise the findings within broader literature. The mapping review methodology was developed by the Evidence for Policy and Practice Information and Co-ordinating Centre in London to offer policymakers, practitioners and researchers an explicit and transparent means of identifying narrower policy and practice-relevant review questions (Grant & Booth, 2009). As our goal is to support the policy discussion and with these issues orient the development of better design practices, and transparent and accountable deployment strategies, this was the most appropriate methodology.

Our review question focused on: how can the primary ethical risks presented by AI-health be categorised, and what must policymakers, regulators and developers consider in order to be ethically mindful? We were concerned with categorising issues in order to facilitate future research and discussion. We chose five literature databases that are relevant to these issues and that are at the cross-section of the technical, medical, ethical and social science literature: Scopus, Google Scholar, Philpapers, Web of Science, Pub Med. Our literature review searches were conducted in April 2019, with references being added or removed throughout the drafting iterations. The search engines are not identical, so we used variations of the following generic search term string: ethic* AND algorithm* OR AI* OR "Artificial Intelligence"* OR "Machine Learning"* AND health* (see Appendix for details on results and search queries). Initial results were screened on title. Those that were deemed relevant were downloaded so that the abstracts and keywords could be reviewed for relevance. At this stage, we excluded any results that were focused on issues related to digital health in general (e.g. data sharing, data access, data privacy, surveillance/nudging, consent, ownership of health data, evidene of efficacy) to remain focused on mapping the current debate about the ethics of AI specifically. Recorrds that the authors had prior knowledge of, and which were relevant to the research question but not included in the initial database searches, were also added

To ensure that the focus stayed on the *unique* ethical issues, the map, developed by (Mittelstadt et al., 2016), of the epistemic, normative, and overarching ethical concerns related to algorithms was used as a base. The typology offered by Mittelstadt et al. identifies problems pertaining to algorithmic decision making and their possible causes, such as error in input or discriminatory output. Traceability[1] arises from the complexity of the system when all of the pieces are put together. This typology will be cross-referenced with each level of abstraction (LoA) we propose below.

First, the selected literature was reviewed to identify healthcare examples of each of the concerns highlighted in the original map, as shown in Table 1, and then reviewed more thoroughly to identify how the ethical issues may vary depending on whether the analysis was being conducted at: (i) individual, (ii) interpersonal, (iii) group (e.g. family or population), (iv) institutional, (v) sectoral, and/or (v) societal levels of abstraction (LoAs) (Floridi, 2008). An LoA can be imagined as an interface that enables one to observe some aspects of a system analysed, while making other aspects opaque or indeed invisible. LoAs are common in computer science, where systems are described at different LoAs (computational, hardware, user-centred etc.). Note that LoAs can be combined in more complex sets, and can be, but are not necessarily hierarchical, with higher or lower 'resolution' or granularity of information. This helped the review avoid the narrow focus on individual-level impacts highlighted in the introduction. This approach is not intended to imply that there is no overlap between the levels.

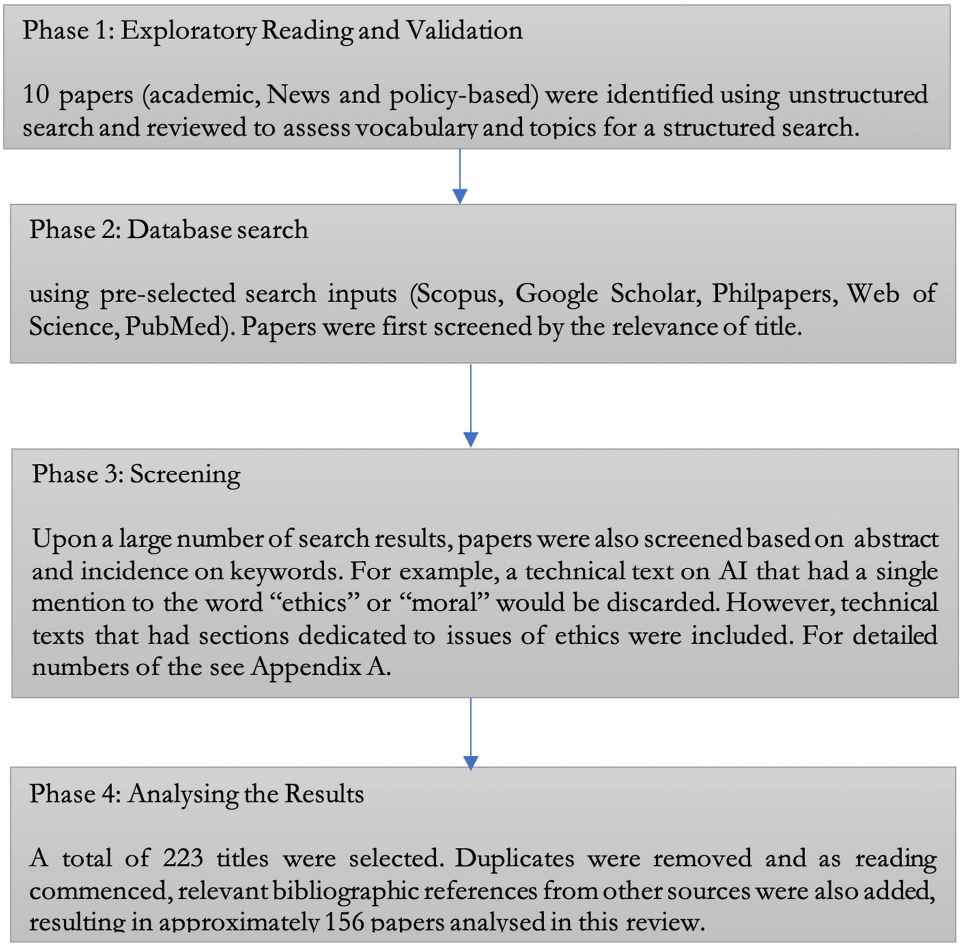
[1] Traceability is introduced as an overarching ethical concern by Mittelstadt et al., (2016). It is used to summarise concerns that arise from the fact that potential algorithmic harms result from the actions of multiple actors. This makes it hard to find the 'cause' of the harm and hard to identify who should be held responsible and/or accountable for the harm caused. It is an overarching concern as it encompasses ethical risks that are both normative and epistemic, and can be applied at any LoA.

	Ethical Concern	Explanation	Medical Example
Epistemic concerns	Inconclusive Evidence	Algorithmic outcomes (e.g. classification) are probabilistic and not infallible. They are rarely sufficient to posit the existence of a causal relationship.	*EKG readers in smartwatches may 'diagnose' a patient as suffering from arrhythmia when it may be due to a fault with the watch not being able to accurately read that user's heartbeat (for example due to the colour of their skin) or the 'norm' is inappropriately calibrated for that individual (Hailu, 2019)*
	Inscrutable Evidence	Recipients of an algorithmic decision very rarely have full oversight of the data used to train or test an algorithm or the data points used to reach a specific decision.	*A clinical decision support system deployed in a hospital may make a treatment recommendation, but it may not be clear on what basis it has made that 'decision' raising the risk that it has used data that are inappropriate for the individual in question or that there is a bug in the system leading to issues with over or under prescribing (Wachter, 2015).*
	Misguided Evidence	Algorithmic outcomes can only be as reliable (but also as neutral) as the data they are based on.	*Watson for Oncology is in widespread use in China for 'diagnosis' via image recognition but has primarily been trained on a Western data set leading to issues with concordance and poorer results for Chinese patients than their Western counterparts (Liu et al., 2018).*
Normative Concerns	Unfair outcomes	An action can be found to having more of an impact (positive or negative) on one group of people	An algorithm *'learns' to prioritise patients it predicts to have better outcomes for a particular disease. This turns out to have a discriminatory effect on people within the Black and minority ethnic communities (Garattini et al., 2019).*
	Transformative effects	Algorithmic activities, like profiling, re-conceptualise reality in unexpected ways.	*An individual using personal health app has limited oversight over what passive data it is collecting and how that is being transformed into a recommendation to improve, limiting their ability to challenge any recommendations made and a loss of personal autonomy and data privacy (Kleinpeter, 2017).*
Overarching	Traceability	Harm caused by algorithmic activity is hard to debug (to detect the harm and find its cause), and it is hard to identify who should be held responsible for the harm caused.	*If a decision made by clinical decision support software leads to a negative outcome for the individual, it is unclear who to assign the responsibility and or liability to and therefore to prevent it from happening again (Racine et al., 2019)..*

Table 1: A summary of the epistemic, normative and overarching ethical concerns related to algorithmic use in healthcare based on Mittelstadt et al (2016) from (Morley & Floridi, 2020b) .

3. Findings

What follows is a detailed discussion of the issues uncovered. A total of 223 titles were selected, duplicates were removed and, as reading commenced, relevant bibliography references were also added, resulting in approximately 147 papers to be read and included in the review. The flowchart below illustrates our methodology. Also, a summary map of our findings (Table 2) is provided at the end of the section.



Phase 1: Exploratory Reading and Validation

10 papers (academic, News and policy-based) were identified using unstructured search and reviewed to assess vocabulary and topics for a structured search.

Phase 2: Database search

using pre-selected search inputs (Scopus, Google Scholar, Philpapers, Web of Science, PubMed). Papers were first screened by the relevance of title.

Phase 3: Screening

Upon a large number of search results, papers were also screened based on abstract and incidence on keywords. For example, a technical text on AI that had a single mention to the word "ethics" or "moral" would be discarded. However, technical texts that had sections dedicated to issues of ethics were included. For detailed numbers of the see Appendix A.

Phase 4: Analysing the Results

A total of 223 titles were selected. Duplicates were removed and as reading commenced, relevant bibliographic references from other sources were also added, resulting in approximately 156 papers analysed in this review.

Figure 1: Flowchart offering and overview of the steps taken in our literature review, filtering from several thousand titles to identified abstracts and selecting 156 papers to read**.**

3.1. Epistemic Concerns: Inconclusive, Inscrutable, and Misguided Evidence

Many factors are encouraging the development of AI-Health (Chin-Yee & Upshur, 2019). One of the main driving forces is the belief that algorithms can make more objective, robust and evidence-based clinical decisions (in terms of diagnosis, prognosis or treatment recommendations) than a human healthcare practitioner (HCP) can (Kalmady et al., 2019). This is not an unfounded position. Machine learning methods, especially ensemble and unsupervised methods (Harerimana et al., 2018), can take into account a far greater range of evidence (data) than a Healthcare Provider (HCP) when making a clinical decision, including five of the seven dimensions of healthcare data provided by the US Department of Health and Human services: (1) demographic and socioeconomic data; (2) symptom and existing diagnosis data; (3) treatment data; (4) outcome data; and (5) other omic data

(Holzinger et al., 2019). If designed taking into account the multiple epistemic concerns, this ability enables clinical algorithms to act as digital companions (Morley & Floridi, 2019a), reducing the information asymmetry that exists between a HCP and the individual seeking care by making available information accessible to both parties and helping ensure that the most informed decision possible is made by the person who has the right to make it (Morley & Floridi, 2019b).

It is at least in part due to this ability to make 'evidence-based' decisions that, as AI-health research has shown, AI techniques can considerably augment or surpass human capabilities when it comes to tasks including: (1) analysis of risk factors (De Langavant et al., 2018; Deng et al., 2018); (2) prediction of disease (Moscoso et al., 2019); (3) prediction of infection (Barton et al., 2019)(López-Martínez et al., 2019); (4) population health monitoring (F. S. Lu et al., 2019; Zacher & Czogiel, 2019); (5) prediction of adverse effects (Ding et al., 2019; Mortazavi et al., 2017); (6) prediction of outcome and/or likelihood of survival (Dong et al., 2019; Popkes et al., 2019; Topuz et al., 2018); and (7) analysing electronic health records (Shickel et al., 2018). These capabilities should not be underestimated, particularly as AI-Health solutions can operate at scale, diagnosing or predicting outcomes for multiple people at once—something that an HCP could never do. Yet in many ways this almost unwavering faith in the truth-telling power of AI-Health is flawed.

As has been highlighted multiple times in the wider ethical AI literature, the belief that algorithms are more objective than humans is a 'carefully crafted myth' (Gillespie et al., 2014), and just because an algorithm can recognise a pattern, for example, does not necessarily make it meaningful (Floridi, 2014). In the context of healthcare, existing methods and studies (potentially including those referenced) suffer from overfitting due to small numbers of samples, meaning that the majority of results (e.g. patterns of disease risk factors, or presence of disease) are inconclusive (Holzinger et al., 2019). This is a problem that is further magnified by the lack of reproducibility, and external validity, of results. AI-Health solutions are often untranslatable between different settings and rarely work in settings different to those in which the initial result was obtained (Vollmer, Mateen, Bohner, Király, Ghani, et al., 2018), raising serious questions about the scientific rigor of AI-Health and its safety (Vayena, Blasimme, & Cohen, 2018). Furthermore, the results can often be heavily value-laden, based on the definition of 'healthy' by influential people or powerful companies (McLaughlin, 2016). This raises a number of significant ethical concerns.

At the **individual LoA** there is considerable risk of misdiagnosis. This can happen in at least two ways: either by an individual using a wearable device that has a bug or is inappropriately calibrated for them (e.g. they could be 'told' that they are suffering from a health condition when they are not, or vice versa), or, an HCP relying on clinical decision support software (CDSS)

(Ruckenstein & Schüll, 2017) could be given an inaccurate diagnosis or recommendation which they do not question due to a tendency to uncritically accept the decisions of automated systems (Challen et al., 2019). Moreover, this can have impacts in medical practice, causing overreliance on the machine diagnostics and deskilling of practitioners (Cabitza et al., 2017). Not only is this a risk for individuals, but it also reverses the advantage of AI-Health solutions being able to operate at scale by introducing the **group LoA** ethical concern of misdiagnosis or missed diagnosis happening repeatedly. Whilst an HCP might give one person the wrong diagnosis and then be corrected, a faulty algorithm, based on the misguided, inscrutable or inconclusive evidence could give the same wrong diagnosis to hundreds or thousands of people at a time (Topol, 2019). The scale of the problems is as large as the scale of the solutions.

Building on this, there are also ethical implications at the **interpersonal LoA.** HCP-patient relationships are primarily based on trust and empathy, and whilst AI-Health solutions can take over tasks that are more routine and standardised, they cannot reproduce the emotional virtues of which human HCPs are capable (Ngiam & Khor, 2019). Consequently, an over-reliance on the 'quantitative' and objective evidence that fuels clinical algorithms (Cabitza et al., 2017) could discredit other forms of diagnosis and treatment (Rosenfeld et al., 2019)—a prominent concern in the case of clinical psychiatry (Burns, 2015). This could lead to the de-humanisation or *im*personalisation of care provision (Juengst et al., 2016), from a service based on *listening* and *theory* to one based purely on *categorisation* (an issue that could again lead to a **group LoA** harm of group-profiling and associated discrimination by providers including insurers; see section 3.2). Not only is this effectively 'paternalism in disguise' (Juengst et al., 2016) but it could also lead to poorer health outcomes due to the disconnect between pure medical evidence and actual behaviour change (Emanuel & Wachter, 2019).

Finally, scaling up to the **institutional, sectoral and societal LoAs,** there is the concern that public health decisions are increasingly made on predictive AI-Health algorithms, which too often rely on the same flawed assumptions as outlined above. Regarding these assumptions, consider the example of Google Flu Trends monitoring the influenza virus. The initial algorithm distorted the spread of the virus in the US (Vayena, Salathé, Madoff, & Brownstein, 2015) making it appear that there were a greater number of influenza cases than there were by mis-classifying influenza-like-illnesses as confirmed cases of influenza (Ortiz et al., 2011). This study carries obvious limitations: the healthcare-seeking behaviour from the population, for example searching for information on the outbreak of a Flu, make this research susceptible to distortion. Such distortion would be affected, for example, if there is large media coverage of an epidemic, or by demographic factors, such as digital divides.

If policy decisions about where to deploy health resources are based on such poor-quality evidence, this could result in the waste of public funds (e.g., promoting vaccination campaigns where they are not needed), damage local economies (e.g., scaring away tourists from a region)—which would result in a positive feedback loop of less money available for public expenditure—and lead to poorer quality public healthcare provision, and thus worse health outcomes for society at large. This worry is particularly paramount when it is considered that the ultimate ambition of AI-Health is to create a learning healthcare system where the 'system' is constantly learning from the data it receives on the performance of its interventions (Faden et al., 2013). Furthermore, it is worth noting that, at this juncture, the example offered above of Flu Trends does not represent the limits of Google's interest—and that of its subsidiaries and its siblings under parent company Alphabet—in public health. As we discuss below, the engagement between Alphabet's AI subsidiary DeepMind and a major UK hospital has attracted the attention of data protection regulators, the press, and academics (Information Commissioner, 2018; Powles & Hodson, 2017). The challenge of ensuring that AI-Health systems function accurately has in turn sparked debates about the appropriateness of sharing data between public and private entities. In response to claims that patient data transferred from the Royal Free Hospital to DeepMind was "far in excess of the requirements of those publicly stated needs" (Powles & Hodson, 2017), DeepMind representatives argued that "data processed in the application have been defined by and are currently being used by clinicians for the direct monitoring and care of AKI [acute kidney injury] patients" (King et al., 2018). Powles and Hodson responded in turn that it is a "statement of fact that the data transferred is broader than the requirements of AKI" (Powles & Hodson, 2018). As this series of claims and counter-claims demonstrates, the quality and quantity of data required for a particular AI-Health application is likely to be a matter of dispute in the context of the collection and sharing of patient data in training AI-Health.

Ultimately, data is necessary for medical practice and thus so are AI-Health solutions that can take in greater volumes of data. But data collected and used in this way is insufficient to inform medical practice; it must be transformed to be useful (Car et al., 2019) and if this transformation process is flawed the results could be hugely damaging, resulting in either wasted funds and poorer health provision, or undue sharing of patient data with private sector actors under the guise of AI-Health.

3.2. Normative Concerns: Unfair Outcomes and Transformative Effects

As referenced in the introduction, healthcare systems across the globe are struggling with increasing costs and decreasing outcomes (Topol, 2019) and their administrators increasingly believe that the answer may well lie in making healthcare systems more informationally mature and able to capitalise on the opportunities presented by AI-Health significantly to improve outcomes for patients, and to reduce the burdens on the system (Cath et al., 2017). Whilst it would be ethically remiss to ignore these opportunities (Floridi, 2019a), it would be equally ethically problematic to ignore the fact that these opportunities are not created by AI-Health technologies *per se* but by their ability to fundamentally change the intrinsic nature of the ways in which healthcare is delivered by coupling, re-coupling and de-coupling different parts of the system. This changes the affordances and constraints of different governing bodies, regulators, and system agents, undermining the mechanisms in place to hold those delivering care accountable and thus introducing new risks (Floridi, 2017a). For example(Morley & Floridi, 2020a):

- Coupling: patients and their data are so strictly and interchangeably linked that the patients *are* their genetic profiles, latest blood results, personal information, allergies etc. (Floridi, 2017a). What the legislation calls 'data subjects" become "data patients";
- Re-Coupling: research and practice have been sharply divided since the publication of the National Commission for the Protection of Human Subjects in the 1970s, but in the digital scenario described above, they are re-joined as one and the same again (Petrini, 2015) (Faden et al., 2013);
- De-Coupling: presence of Healthcare Provider (HCP) and location of Patient become independent, for example because of the introduction of online consultations (NHS England, 2019).

As a result of these transformations a number of ethical concerns arise.

Starting once again with the **individual LoA**: as more diagnostic and therapeutic interventions become based on AI-Health solutions, individuals may be encouraged to share more and more personal data about themselves (Racine et al., 2019)—data that can then be used in opaque ways (Sterckx et al., 2016). This means that the ability for individuals to be meaningfully involved in shared decision making is considerably undermined (Vayena et al., 2018). As a result, the increasing use of algorithmic decision-making in clinical settings can have negative implications for individual autonomy, as for an individual to be able to exert agency over the AI-Health derived clinical decision, they would need to have a good understanding of the underlying data, processes and technical possibilities that were involved in it being reached (DuFault & Schouten, 2018) and be able to ensure their own values are taken into consideration (McDougall, 2019). The vast majority of the population do not have the level of eHealth literacy necessary for this (Kim & Xie, 2017), and those that do (including HCPs) are prevented from gaining this understanding due to the black-box nature of AI-Health algorithms (Watson et al., 2019). In extreme instances, this could undermine an individual's confidence in their ability to refuse treatment (Ploug & Holm, 2019). Such issues pose a substantial threat to an individual's integrity of self (the ability of an individual to understand the forces acting on them) (Cheney-Lippold, 2017). Given that damage to a person's psychological integrity can be perceived as a 'harm', not accounting for this potentiality poses the risk of creating a system that violates the first principle of medical ethics: *primum non nocere* ("first, do no harm") (Andorno, 2004; Morley & Floridi, 2019a).

It is not necessarily the case that harmful impacts will primarily be felt by the patients. At the **interpersonal LoA,** HCPs may themselves feel increasingly left 'out of the loop' as decisions are made by patients and their 'clinical advice' algorithm in a closed digital loop (Nag et al., 2017). As a result, HCPs may too feel unable to exert their own agency over the decision-making capacity of AI-Health solutions. Though the use of algorithmic decision-making makes diagnostics seem like a straightforward activity of identifying symptoms and fitting them into textbook categories, medical practice is much less clear-cut than it seems (Cabitza et al., 2017). Clinical practice involves a series of evaluations, trial and error, and a dynamic interaction with the patient and the medical literature. As a result, formal treatment protocols should be seen more as evaluative guidelines than well-defined, isolated categories. AI-Health solutions may not be in accordance with current best practice, which is necessary to handle the great degree of uncertainty and can only be fully evaluated by physicians (Cabitza et al., 2017). Therefore, AI-Health solutions need to allow HCPs to exert influence in the decision-making process.

At the **group LoA** the concern is that AI-Health systems may well be able to better identify illnesses and injuries that have well-established and fairly set (and therefore automatable) treatment protocols. These are more likely to exist for afflictions most commonly suffered by white men as there is a greater volume of medical trials data for this group than there is for almost any other group. Algorithms trained on such biased datasets could make considerably poorer predictions for, for example, younger black women (Vayena et al., 2018). If HCPs are left out of the loop completely and learning healthcare systems primarily rely on automated decisions, there is considerable potential to exacerbate existing inequalities between the "haves" and the "have-nots" of the digital healthcare ecosystem, i.e. those that generate enough data on themselves to ensure accurately trained algorithms and those that do not (Topol, 2019).

To mitigate these and associated risks, **institutions** need to be asking the crucial question: how much clinical decision-making should we be delegating to AI-Health solutions (Di Nucci, 2019)? If it is known that algorithms which enable profiling (e.g. those that determine genetic risk profiles) can ignore outliers and provide the basis for discrimination (Garattini et al., 2019), deciding whether healthcare is also seen as a means of promoting social justice is crucial in order to establish: what type of data services will be embedded in the system (Voigt, 2019); what data should be collected; and which values should be embedded in algorithmic decision-making services (McDougall, 2019). This decision also determines what sort of population-level behavioural change the health system should be able to aim for depending on cost management, data collection and fairness in data-driven systems (Department of Health and Social Care, 2018.). If not carefully considered, this process of transforming the provision of care risks over-fitting the system to a specific set of values that may not represent those of society at large (McDougall, 2019).

Another, more subtle yet pervasive transformative effect arises at the **sectoral** level. Powles & Hodson (2017) argue that one risk that may arise from collaboration between public and private sector entities such as that between the Royal Free London hospital and DeepMind is that the positive benefits of AI-Health "solutions" will be siloed within private entities. They note that in the Royal Free case, "DeepMind [was given] a lead advantage in developing new algorithmic tools on otherwise privately-held, but publicly-generated datasets" (Powles & Hodson, 2017, p. 362). This, they suggest, may mean that the only feasible way that future advances may be developed is "via DeepMind on DeepMind's terms". This interpretation was contested by DeepMind, who called it "unevidenced and untrue" and claimed that the Information Commissioner agreed with their stance in her 2018 ruling (King et al., 2018). Whatever the circumstances of this particular case, the broader risk of privately held AI-Health solutions—trained on datasets that have been generated *about* the public *by* public actors but then (lawfully) shared with private companies—is worthy of caution going forward and a worthwhile topic of ongoing discussion in public health ethics.

As may now be clear, these transformative effects also have significant ethical implications at the **societal LoA.** Before institutions can establish where and how (and, from the sectoral perspective, whether) AI-Health solutions can improve care, society itself must make difficult decisions about what care *is* and what constitutes *good* care (Coeckelbergh, 2014). To offer a simplistic example, does it mean purely providing a technical diagnosis and an appropriate prescription or does it involve contemplating a series of human necessities that revolve around wellbeing (Burr, Taddeo, & Floridi, 2020)? If it is the former, then it is relatively easy to automate the role of non-surgical clinicians through AI (although this does not imply that doctors *should* be substituted by AI systems). However, if is the latter, then providing good healthcare means encompassing psychological wellbeing and other elements related to quality of life, which would make human interaction an essential part of healthcare provision, as a machine does not have the capability to make emotionally-driven decisions. Consequently, certain decisions may completely exceed the machine's capabilities and thus delegating these tasks to AI-Health would be ethically concerning (Matthias, 2015).

Consider, for example, a situation where an AI-Health solution decides which patients are sent to the Intensive Care Unit (ICU). Intensive care is a limited resource and only people who are at risk of losing their lives or suffering grave harms are sent there. Triage decisions are currently made by humans with the aim of maximising well-being for the greatest number of people. Doctors weigh different factors when making this decision, including the likelihood of people surviving if they are sent to the ICU. These situations often involve practitioners (implicitly) taking moral stances, by prioritising individuals based on their age or health conditions. These cases are fundamentally oriented by legal constraints and medical norms (e.g. adherence to bioethical principles or codes of best practice), yet personal expertise, experience and values also inevitably play a role. Having the support of AI-Health in the ICU screening increases the number of agents and complicates the norms involved in these decisions, since the doctor may follow his or her professional guidelines, while the algorithm will be oriented by the values embedded in its code. Unless there is a transparent process for society to be involved in the weighing of values embedded in these decision-making tools (for instance, how is 'fair' provision of care defined?) (Cohen et al., 2014), then the use of algorithms in such scenarios could result in the overfitting of the health system to a specific set of values that are not representative of society at large.

In response to this risk, some attempts have already been made to involve the public at large in decisions over the design and deployment of AI systems. In early 2019, the UK's Information Commissioner's Office and the National Institute for Health Research staged a series of "citizens' juries" to obtain the opinions of a representative cross-section of British society regarding the use of AI in health (Information Commissioner, 2019). The "juries" were presented with four scenarios, two relating to health—using AI to diagnose strokes, and using it to find potential matches for a kidney transplant—and another two relating to criminal justice. Notably, the juries "strongly favoured accuracy over explanation" in the two scenarios involving AI in health (National Institute for Health Research, 2019). This is just one example of research which attempted to obtain public opinion data regarding AI in health, and there are reasons to suppose that the apparent preference among participants for accurate over explainable AI systems reflects the high-stakes and fast-moving scenarios that were presented (as opposed to, say, the more routine illnesses and injuries we are focusing on here). Nonetheless, it demonstrates the plausibility and preferability of involving the public in designing AI-Health systems.

To conclude this sub-section, the notion that AI-Health technologies are ethically neutral is unrealistic, and having them perform moral decision-making and enforcement may provoke immoral and unfair results (Rajkomar et al., 2018). The direct involvement of the public in the design of AI-Health may help mitigate these risks. This should be borne in mind by all those involved in the AI-driven transformation of healthcare systems.

3.3 Overarching Concerns: Traceability

The previous sub-section outlined how the increasing use of AI-Health is fundamentally transforming the delivery of healthcare and the ethical implications of this process, particularly in terms of potentially unfair outcomes. This transformation process means that healthcare systems now rely on a dynamic, cyclical and intertwined series of interactions between human, artificial and hybrid agents (Vollmer, Mateen, Bohner, Király, & Ghani, 2018; Turilli & Floridi, 2009). This is making it increasingly challenging identify interaction-emerging risks and allocate liability, raising ethical concerns with regards to moral responsibility.

Moral responsibility involves both looking forward, where an individual, group or organisation is perceived as being in charge of guaranteeing a desired outcome, and looking backwards to appropriate blame and possibly redress, when a failure has occurred (Wardrope, 2015). In a well-functioning healthcare system, this responsibility is distributed evenly and transparently across all nodes so that the causal chain of a given outcome can be easily replicated in the case of a positive outcome, or prevented from repeating in the case of a negative outcome (Floridi, 2013, 2016). In an algorithmically-driven healthcare system, a single AI diagnostic tool might involve many people organising, collecting and brokering data, and performing analyses on it, making this transparent allocation of responsibility almost impossible. In essence, not only is the decision-making process of a single algorithm a black-box, but the entire chain of actors that participate in the end product of AI-Health solutions is extremely complex. This makes the entire AI-Health ecosystem inaccessible and opaque, making responsibility and accountability difficult.

To clearly outline the ethical implications of this at-scale lack of traceability, let us take the example of a digital heart-rate monitor that 'intelligently' processes biological and environmental data to signal to its user their risk of developing a heart condition.

At the **individual LoA** this process relies on what can be termed the 'digital medical gaze' (Morley & Floridi, 2019a) and is based on this micro-cycle of self-reflection adapted from (Garcia et al., 2014):

1. Gaining Knowledge: Algorithm reads multi-omic dataset to determine risk of heart attack and providers individual with a 'heart health score'
2. Gaining Awareness: on the advice of the algorithm, individual starts monitoring their activity level and becomes aware of how active they are
3. Self-reflection: as directed by the algorithm the individual reflects on how much high fat food they are eating in a day and compares this to their optimal diet based on their genomic profile and their level of activity
4. Action: individual takes the advice of the algorithm and takes specific actions to improve their heart-health score e.g. starts regular exercise.

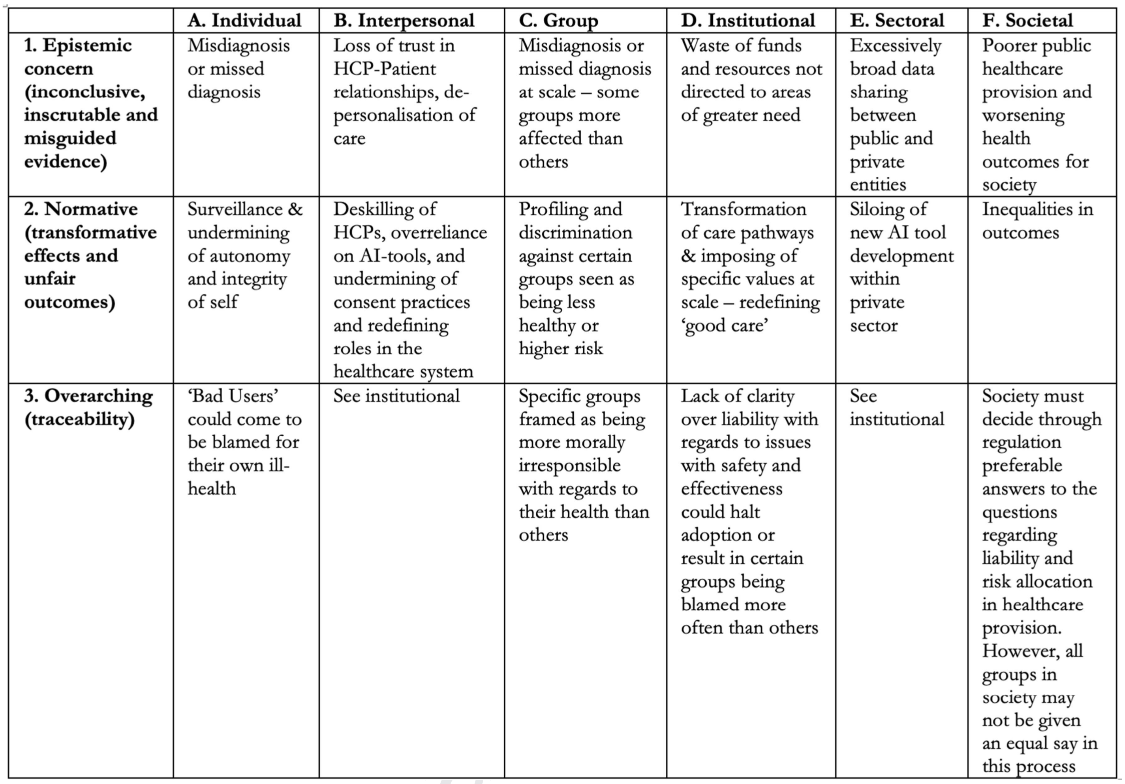
If this process of self-reflection does not 'work' in the sense that it does not result in a person taking appropriate action to improve their heart-health, for any number of reasons, including data inaccuracy, and the individual still ends up experiencing heart failure, this process of algorithmic surveillance (Rich & Miah, 2014) risks creating an elaborate mechanism for victim-blaming (Danis & Solomon, 2013; McLaughlin, 2016). The individual may be seen as being a 'bad user' for failing to act upon the allegedly objective and evidence-based advice of the algorithm (see section 3.1), and may therefore be framed as being morally responsible for their poor health and not deserving of state-provided healthcare. Yet, due to the lack of traceability, there can be no certainty that the poor outcome was due to the lack of action by the individual: it could be a faulty device, buggy code, or the result of biased datasets (Topol, 2019). Moreover, even if a negative outcome were to result purely from an individual disregarding the guidance, the adoption of digital infrastructure that enables failure to be ascribed to a morally 'culpable' individual is itself a matter of ethical concern. These new insights may enable lives to be saved and quality of life to be drastically improved, yet they also shift the ethical burden of 'living well' squarely onto newly accountable individuals. The ontological shift that this new infrastructure permits—from individuals-as-patients deserving quality healthcare, regardless of their prior choices as fallible humans, to individuals-as-agents expected to take active steps to pre-empt negative outcomes—raises stark questions for bioethics, which has traditionally been seen as an "ethics of the receiver" (Floridi 2008). Moreover, these technological changes might prompt a shift in the ethical framework, burdening the individuals, while not providing *de facto* means of behavioural change. Many concerns stem from socio-demographic issues which entail harmful habits, and cannot oversimplified to a matter of delivering the adequate information to the patient (Owens & Cribb, 2019).

Due to issues of bias (discussed further in section 3.2), there is, further, a **group LoA** ethical risk that some groups may come to be seen as being more morally irresponsible about their healthcare than others. Heart-rate monitors, for example, are notoriously less accurate for those with darker skin (Hailu, 2019), meaning that they could give considerably less accurate advice to people of colour than to those with light skin. If this results in people of colour being less able to use AI-Health advice to improve their heart-health, then these groups of people may be seen as morally reprehensible when it comes to their health. Furthermore, the healthcare could then 'learn' to predict that people of colour have worse heart-health, potentially resulting in these groups of individuals being discriminated against by, for example, insurers (Martani et al., 2019).

At the **interpersonal**, **institutional** and **sectoral LoAs,** this moral responsibility translates into liability. If for example, instead of the heart-health algorithm providing the advice back to the individual, it provides the data to the individual's HCP and the HCP provides advice that either fails to prevent an adverse event or directly causes an adverse event, this could be the basis of a medical malpractice suit (Price, 2018). In this scenario, it remains unclear where the liability will eventually sit (Ngiam & Khor, 2019). Current law implies that the HCP would be at fault, and therefore liable, for an adverse event as the algorithm in this scenario would be considered a diagnostic support tool—just like a blood test—with no decision making capacity, so it is the HCP's responsibility to act appropriately based on the information provided (Price et al., 2019; Schönberger, 2019; Sullivan & Schweikart, 2019). However, the supply chain for any clinical algorithm is considerably more complex and less transparent than that of a more traditional diagnostic tool meaning that many are questioning whether this is actually how the law will be interpreted in the future. For example, does the liability really sit with the HCP for not questioning the results of the algorithm, even if they were not able to evaluate the quality of the diagnostic against other sources of information, including their own personal knowledge of the patient due to the black-box nature of the algorithm itself? And what about the role of the hospital or care facility: does it have a responsibility to put in place a policy allowing HCPs to overrule algorithmic advice when this seems indicated? Similarly, what role do commissioners or retailers of the device that contains the algorithm play? Do they not carry some responsibility for not checking its accuracy, or do they assume that this responsibility sits with the regulator (for example, MHRA in the UK, the FDA in the US or the CFDA in China) who should, therefore, carry the burden for not appropriately assessing the product before it was deployed in the market? What if the problem is further back in the chain, stemming from inaccurate coding or poor-quality training data? There is a clear lack of distributed responsibility (Floridi, 2013, 2016)—a problem that is exacerbated by a lack of transparency—making it hard to hold individual parts of the chain accountable for poor outcomes which poses a significant ethical risk.

In their overview of patient-safety issues with AI in healthcare, He et al. (2019) state that those working in the field are trying to establish a systems-wide approach that does not attribute blame to individuals or individual companies, but conclude that where liability will ultimately rest remains to be seen. This is problematic because, as Hoffman et al. (2019) stress, uptake of algorithmic-decision-making tools by the clinical community is highly unlikely until this liability question is resolved (Vayena et al., 2018), which could result in the overarching ethical concern raised in the introduction—that of a significant missed opportunity. Many, including Holzinger et al., (2019), believe that explainability is the answer to solving this problem and that, if HCPs can understand how a decision was reached, then reflecting on the output of an algorithm is no different from any other diagnostic tool. Indeed Schönberger (2019) argues that legally this is the case and that as long as it can be proven that the duty of care was met, then harm caused to a patient by an erroneous prediction of an AI-Health system would not yet constitute medical negligence but that it *might* in the near future constitute negligence to *not* rely on the algorithmic output, which brings us back to the issues outlined in section 3.1.

Overall, this lack of clarity will continue to persist for some time (Schönberger, 2019), making it once again a social issue. Society will ultimately dictate what the socially acceptable and socially preferable (Floridi & Taddeo, 2016) answers are to these pressing questions. The ethical issue is whether all parts of society will have an equal say in this debate, as in the example of citizens' juries above, or whether it will be those individuals or groups with the loudest voices that get to set the rules. As Beer (2017) attests, when thinking about the power of an algorithm, we need to think beyond the impact and consequences of the code, to the powerful ways in which notions and ideas about the algorithm circulate throughout the social world.



	A. Individual	B. Interpersonal	C. Group	D. Institutional	E. Sectoral	F. Societal
1. Epistemic concern (inconclusive, inscrutable and misguided evidence)	Misdiagnosis or missed diagnosis	Loss of trust in HCP-Patient relationships, de-personalisation of care	Misdiagnosis or missed diagnosis at scale – some groups more affected than others	Waste of funds and resources not directed to areas of greater need	Excessively broad data sharing between public and private entities	Poorer public healthcare provision and worsening health outcomes for society
2. Normative (transformative effects and unfair outcomes)	Surveillance & undermining of autonomy and integrity of self	Deskilling of HCPs, overreliance on AI-tools, and undermining of consent practices and redefining roles in the healthcare system	Profiling and discrimination against certain groups seen as being less healthy or higher risk	Transformation of care pathways & imposing of specific values at scale – redefining 'good care'	Siloing of new AI tool development within private sector	Inequalities in outcomes
3. Overarching (traceability)	'Bad Users' could come to be blamed for their own ill-health	See institutional	Specific groups framed as being more morally irresponsible with regards to their health than others	Lack of clarity over liability with regards to issues with safety and effectiveness could halt adoption or result in certain groups being blamed more often than others	See institutional	Society must decide through regulation preferable answers to the questions regarding liability and risk allocation in healthcare provision. However, all groups in society may not be given an equal say in this process

Table 2: Summary of the epistemic, normative and overarching ethical concerns associated with AI-Health at the six different LoAs as identified by the literature review

4.The Need for an Ethically-Mindful and Proportionate Approach

The literature surveyed in this review clearly indicates the need for an agreed standard for AI-Health ethical evaluation. While these issues are all connected, they cannot be treated under the blanket discussion of "Ethics of AI" when discussing specific recommendations and solutions. For example, handling privacy at the individual LoA, considering design issues, is different from handling privacy at a group level, where the concern is raised from the ways in which the aggregate data is treated. For these reasons, we need to consider the epistemic, normative and traceability ethical concerns at the six different LoAs to set the different fields of discussion. Protecting people from the harms of AI-Health goes beyond protecting data collection and ensuring that the algorithmic models have been validated. The discussion needs to discern how these issues present differently at the different stages of the algorithmic developmet lifecyle, the ethical issues present at the data collection stage are likely to be different to those present at the deployment stage.

An example of an issue that shows up often is the legal challenge of liability allocation in cases of medical error. This legislative and regulatory discussion is directly dependent on understanding the ethical issues of each stage an AI implementation (e.g. data collection, training, deployment) and making a normative decision on how these risks and burdens are going to distributed through society. Therefore, the ethical discussion needs to be plotted before engaging in any sort of policy, regulatory or legislative discussion.

Similarly, many challenges will not be addressed through rules. Much of the risk of handling data and algorithms stems from professionals not adopting measures to protect privacy and support cybersecurity. In these cases, policy-makers can use our framework to identify at which levels they can best tackle issues. For example, one issue can be *individual* capacitation, where a solution would be promoting doctors' and patients' understanding and control over AI tools; educating about how AI-Health produces predictions or recommendations that are used in treatment plans, and access to and protection of patient data (Ngiam & Khor, 2019). The issue, however, could occur at an organisational (*group*) scale, so better control over how the interface and design of AI-Health products influences HCP-patient-artificial-agent interactions (Cohen et al., 2014) could address the issue. Finally, some cases could be handled at an institutional level, organising campaigns and creating certifications for professionals seeking to use AI-Health tools is also necessary for the adequate implementation and use of AI (Kluge et al., 2018).

To tackle these challenges, regulators will have to consider hard and soft mechanisms, meaning what *ought* to be done and what *may* be done based on the existing moral obligations (Floridi, 2018). These mechanisms will have to consider the different stakeholders involved in each issue and LoA, to balance the need to protect individuals from harm, whilst still supporting innovation that can deliver genuine system and patient benefit (Morley & Joshi, 2019). In short, healthcare systems should not be overly cautious about the adoption of AI-Health solutions, but should be mindful of the potential ethical impacts (Floridi, 2019a) so that proportionate governance models can be developed (Sethi & Laurie, 2013). These governance models can, in turn, help ensure that those responsible for ensuring that healthcare systems are held accountable for the delivery of high-quality equitable and safe care.

What these regulations, standards and policies should cover and how they should be developed remain open questions (Floridi, 2017b), which will likely be 'solved' multiple times over by different healthcare systems operating in different settings. However, in order to lend a more systematic approach to addressing these outstanding questions, enabling greater coherence and speed in addressing these challenges, in Table 3 below we have assembled a list of essential cross-cutting considerations that emerge from our mapping review. The table indicates from which aspect of our mapping review (ethical concern × LoA, corresponding to a cell in Table 2) each consideration is assigned by an increasing Level of Abstraction: Individual (A), Interpersonal (B), Group (C), Institutional (D), Sectoral (E) and Societal (F).

Consideration	**Key supporting literature**	**Relevant aspects (ascending LoA[2])**	**Example body responsible for answering this question based on the English National Health System**
What skills will the professional healthcare workforce require in order to make safe and effective use of AI-Health solutions in the future?	(Kluge et al., 2018)	Epistemic (A, B, C, F) Normative (B, C, D, E) Overarching (A, C)	Health Education England should survey the skills currently available in the workforce and conduct a gap analysis of the skills that will be needed
Which tasks should be delegated to AI-Health solutions, and which should not?	(Di Nucci, 2019)	Epistemic (A, B, C, D, F) Normative (B, C, D, F) Overarching (A, C, D)	Department of Health and Social Care should conduct a multi-stakeholder engagement process to understand which tasks are socially acceptable to be delegated to AI-health and make this offficial policy,.
What evidence is needed to 'prove' clinical effectiveness of an AI-Health solution?	(Greaves et al., 2018)	Epistemic (A, B, C, E, F) Normative (E) Overarching (A, C, D, F)	Medicines and Healthcare Regulators Medicines and Healthcare products Regulatory Agency should update the medical device regulations to include a minimum required standard of accuracy and a minimum standard of evidence to demonstrate that the AI-health product is genuinely capable of performing at this level
What mechanisms should be put in place to enable people to report and seek redress for AI-Health associated harms?	(Schönberger, 2019)	Epistemic (A, C, E, F) Normative (A, C, E, F) Overarching (A, C, D)	The Care Quality Commission should update its inspection framework to regularly check that AI-health products in use are continuing to operate safely. Medicines and Healthcare Regulators Medicines and Healthcare products Regulatory Agency should introduce a 'yellow card' scheme for AI-health products so that users can report errors and be assured that they are being

[2] Denoted by an increasing Level of Analysis: Individual (A), Interpersonal (B), Group (C), Institutional (D), Sectoral (E) and Societal (F).

			taken care of.
What mechanisms should be put in place to ensure all relevant stakeholder views are included in the development of AI-Health solutions?	(Aitken et al., 2019)	Epistemic (C, E, F) Overarching (A, C, D, F)	The Health Research Authority should update its guidance on ethical approval for AI-health research and product development to set out the minimum participation requirements for diverse stakeholders
How can the explainability of AI-Health solutions be guaranteed?	(Watson et al., 2019)	Epistemic (A, C) Normative (A, C, E) Overarching (A, D)	Medicines and Healthcare Regulators Medicines and Healthcare products Regulatory Agency should update the regulations governing medical devices to set out the minimum standards for 'explainability' of AI-health products
What mechanisms can be put in place to ensure reliability, replicability and safety of AI-Health solutions?	(Challen et al., 2019)	Epistemic (A, C, F) Normative (C, E, F) Overarching (A, C, D)	The National Institute for Health and Care Excellence should make make it a requirement of formal health technology assessment that, within the bounds of technical feasibility and respecting intellectual property, developers make code open to enable reproducibility and error checking.
How can transparency over how algorithmic tools are integrated into the healthcare workflow, how it shapes decisions, and how it affects process optimization within medical services, be guaranteed?	(Vayena et al., 2015)	Epistemic (A, B, C, D, E, F) Normative (A, B, D, F) Overarching (A, D, F)	NHS England should make it a requirement of all NHS trusts, hospitals and providers of care to declare when an AI-health solution is being used in a specific care pathway and to be clear about how its safety and quality is being regularly assessed.
How can traditional and non-traditional sources of health data be incorporated into AI-Health decision making? And how can it be appropriately protected and how can it be harmonised?	(e.g. Maher et al., 2019; Ploug & Holm, 2016; Richardson, Milam, & Chrysler, 2015; Townend, 2018)	Epistemic (A, C, D, E, F) Normative (A, C, D, E, F) Overarching (A, C, D, E)	The Health Research Authority and NHS Digital should update guidance and regulations governing secondary uses of health data to incorporate the specific considerations of AI-Health as we have outlined in this paper/
How are bioethical concepts (beneficence, non-maleficence, autonomy and justice (Beauchamp & Childress, 2013) challenged by AI-Health?	(Mittelstadt, 2019)	Epistemic (B, F) Normative (A, C, D, F) Overarching (A, F)	The Nuffield Council on Bioethics should update its guidance on the bioethical principles for data initiatives to incorporate AI-health specific considerations.
How can concepts such as fairness, accountability and transparency can be maintained at scale (Morley & Floridi, 2020a)?	(Rosenfeld et al., 2019)	Epistemic (C, D, E, F) Normative (D, E, F) Overarching (F)	The Care Quality Commission, should develop a mechanism for monitoring these impacts at scale as part of its regular review process that

			is designed to ensure safe and high-quality care. This may require the Department of Health and Social Care extending its regulatory powers.

Table 3: Eleven key considerations for policymakers that arose from the literature review, denoted by an increasing Level of Abstraction: Individual (A), Interpersonal (B), Group (C), Institutional (D), Sectoral (E) and Societal (F).

There are steps being taken towards regulation and legislation, however, these discussions often fail to address broader ethical questions such as "what constitutes good healthcare?" (Coeckelbergh, 2014), "what services should be contemplated in our standard of ´care´?", and others. Without addressing these larger questions, it is hard to orient greater normative frameworks and produce coherency across stakeholders in each LoA. For these reasons, their development is progressing slowly (which is why the relevant literature is unlikely to reflect all current developments) and almost all focus solely on interventions positioning themselves as being health-related in the medical sense, not in the wider, wellbeing sense (e.g., healthy exercise, diet, sleeping habits).

Awareness of the need to consider these questions is increasing, and efforts are being made at both a national and international level to adapt existing regulations so that they remain fit for purpose (The Lancet Digital Health, 2019). The American Food and Drug Administration (FDA) is now planning on regulating Software as a Medical Devices (SaMD) (Food and Drug Administration (FDA), 2019) and in both the EU and the UK Regulation 2017/745 on medical devices comes into effect in April 2020 and significantly increases the range of software and non-medical products that will need to be classed (and assessed) as medical devices. This practical, normative debate necessarily needs to go through the discussion about what is expected of a medical device, and therefore what is considered to be treatment.

Similarly, there has been moves to pass ethical codes without considering these multi-layered interests and challenges. However, some changes are worth noting. The UK has published its Code of Conduct for data-driven health and care technologies, standards for evidence of clinical effectiveness for digital health technologies (Greaves et al., 2018)—a digital assessment questionnaire standards for apps—and is currently developing a 'regulation as a service' model to ensure that there are appropriate regulatory checks at all stages of the AI development cycle. The World Health Organisation has a number of projects under way to develop guidance for member states (Aicardi et al., 2016) (World Health Organisation, 2019). In China, several norms provide specific and detailed instructions to ensure health data security and confidentiality (Wang, 2019) to ensure that health and medical big data sets can be used as a national resource to develop algorithms (Zhang et al., 2018) for the improvement of public health (Li et al., 2019).

The ethical questions involved in the use of AI for healthcare trickle down to issues of which matters can or should be regulated within the scope of healthcare, against what is considered simply a wellbeing service. Therefore, thinking in the terms of the proposed framework helps policymakers also understand and delineate the scope of their regulation. For example, some algorithmic tools potentially enable people to bypass formal and well-regulated healthcare systems entirely by accessing technology directly, either by using a wearable device or consulting online databases (Burr, Morley, Taddeo, & Floridi, 2020). There must be a discussion, considering the LoAs and concerns, on whether these services have *de facto* overstepped the boundaries into healthcare in any of those levels.

Similarly, although some technical solutions have been put forward for mitigating issues with data bias (Gebru et al., 2018; Holland et al., 2018) and data quality (Dai et al., 2018) and ensuring social inclusion in decision-making (Balthazar et al., 2018; Friedman et al., 2017; Rahwan, 2018), these remain relatively untested. Unless a competitive advantage of taking such pro-ethical steps becomes clear without these approaches being made mandatory, it is unlikely that they will have a significant impact on the ethical impacts of AI-Health in the near future. As a result, there is still little control over the procedures followed and quality control mechanisms (Cohen et al., 2014) involved in the development, deployment and use of AI-Health.

As these comparatively easier to tackle problems do not yet have adequate solutions, it is unsurprising that the bigger issues regarding the protection of equality of care (Powell & Deetjen, 2019), fair distribution of benefits (Balthazar et al., 2018) (Kohli & Geis, 2018) and the protection and promotion of societal values (Mahomed, 2018) have barely even been considered. Given that healthcare systems in many ways act as the core of modern societies this is concerning. If mistakes are made too early in the adoption and implementation of AI in healthcare, the fall-out could be significant enough to undermine public trust, resulting in significant opportunity costs, and potentially encouraging individuals to seek their healthcare from outside of the formal systems where they may be presented with even greater risks. A coherent approach is needed and urgently, hopefully this systematic overview of the issues to be considered can help speed up its development.

5. Conclusion

This thematic literature review has sought to map out the ethical issues around the incorporation of data-driven AI technologies into healthcare provision and public health systems. In order to make this overview more useful, the relevant topics have been organised into themes and six different levels of abstraction (LoAs) have been highlighted. The hope is that by encouraging a discussion of the ethical implications of AI-Health at individual, interpersonal, group, institutional and societal LoAs, policymakers and regulators will be able to segment a large and complex conversation into tractable debates around specific issues, stakeholders, and solutions. This is important, as Topol (2019) states 'there cannot be exceptionalism for AI in medicine,' especially not when there is potentially so much to gain (Miotto et al., 2018).

With this in mind, the review has covered a wide range of topics while also venturing into the specificity of certain fields. This approach has enabled a fuller and more nuanced understanding of the ethical concerns related to the introduction of AI into healthcare systems than has been previously seen in the literature. Inevitably, there are limitations to this approach. Firstly, it is important to note that the selection of articles and policy documents was restricted to those written in English. This means that some ethical issues will have been overlooked (e.g. those in Spanish-speaking countries or in China). Second, academic literature, much like regulation, tends to struggle to keep pace with technological development. This literature review did not seek to identify ethical issues associated with specific use cases of AI first-hand, for example, by reviewing recently published studies available on pre-print servers such as arXiv, but instead focused on providing an overview of the ethical issues already identified and becoming mature. As a result, there may well be ethical concerns that are associated with more emergent use cases of AI for healthcare that we have not identified as they have not yet been discussed in formal peer-reviewed publications.

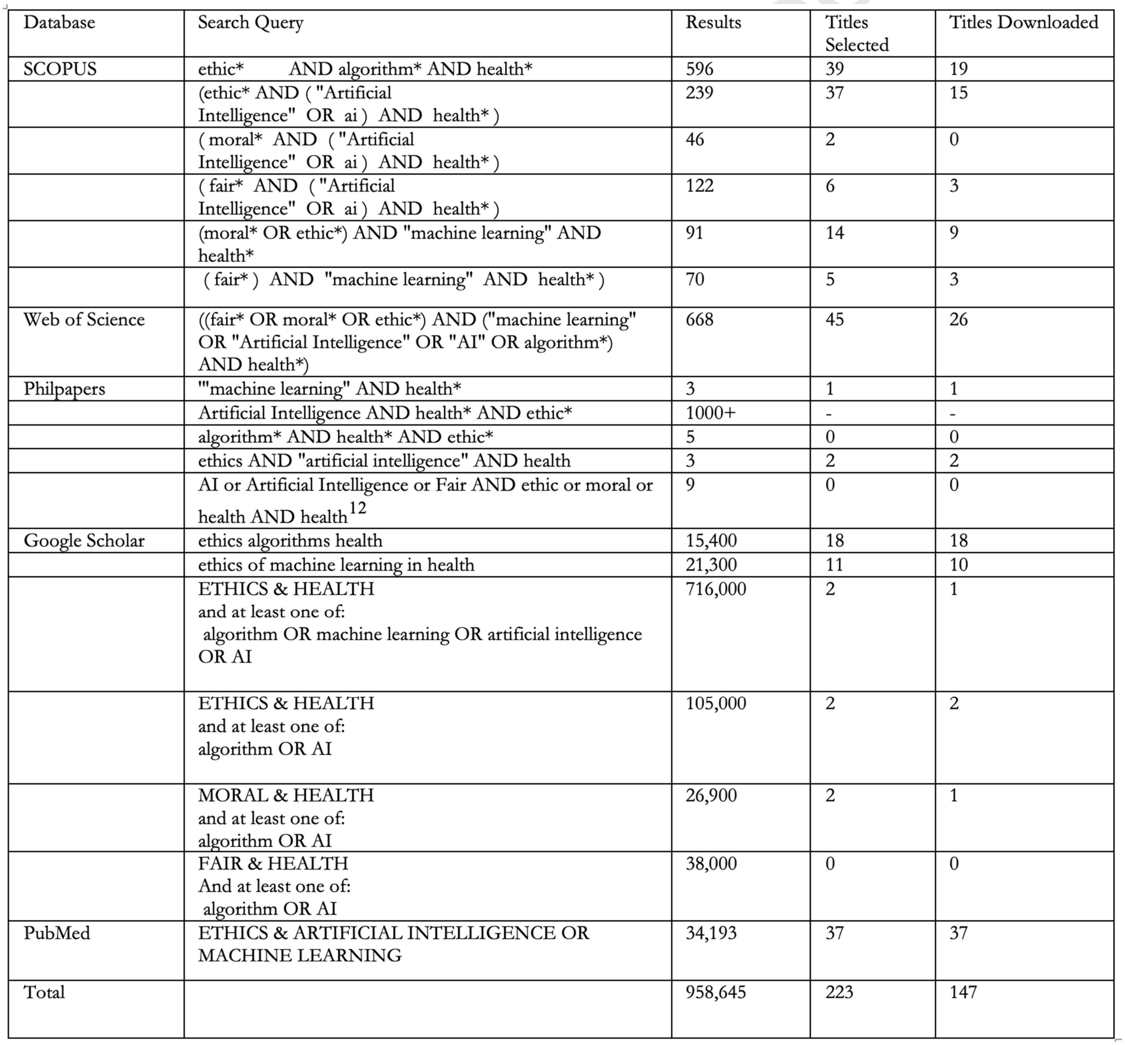
To overcome these limitations, further research could seek to expand the literature review by including a wider range of search queries, and by taking a case-study approach to analysing the ethical issues of specific practices and then aggregating these. This could be complemented by a comprehensive review of the policies, standards and regulations in development in different healthcare systems across the globe to assess the extent to which these are likely to be effective at mitigating these ethical concerns.

In this article, we hope to have provided a sufficiently comprehensive and detailed analysis of the current debates on ethical issues related to the introduction of AI into healthcare systems. The aim is to help policymakers and legislators develop evidence-based and proportionate policy and regulatory interventions. In particular, we hope to encourage the development of a system of transparent and distributed responsibility, where all those involved in the clinical algorithm supply chain can be held proportionately and appropriately accountable for the safety of the patient at the end, not just the HCP. It is only by ensuring such a system is developed that policymakers and legislators can be confident that the inherent risks we have described are appropriately mitigated (as far as possible) and only once this is the case will the medical community at large feel willing and able to adopt AI technologies.

Appendix – Methodology

This review process resulted in 156 papers suitable for analysis and inclusion in the initial review. Subsequent relevant papers that met the criteria were added at a later date during the writing up of the results.

This literature review also included accessory readings and case studies that were encountered during the research process. This includes bibliography obtained from the references of the papers analysed, and case studies identified in the readings (e.g. the Deep Mind case study). It is our belief that these exploratory readings enrich our systematic approach by developing on interesting findings and topics identified throughout our investigation.



Database	Search Query	Results	Titles Selected	Titles Downloaded
SCOPUS	ethic* AND algorithm* AND health*	596	39	19
	(ethic* AND ("Artificial Intelligence" OR ai) AND health*)	239	37	15
	(moral* AND ("Artificial Intelligence" OR ai) AND health*)	46	2	0
	(fair* AND ("Artificial Intelligence" OR ai) AND health*)	122	6	3
	(moral* OR ethic*) AND "machine learning" AND health*	91	14	9
	(fair*) AND "machine learning" AND health*)	70	5	3
Web of Science	((fair* OR moral* OR ethic*) AND ("machine learning" OR "Artificial Intelligence" OR "AI" OR algorithm*) AND health*)	668	45	26
Philpapers	'"machine learning" AND health*	3	1	1
	Artificial Intelligence AND health* AND ethic*	1000+	-	-
	algorithm* AND health* AND ethic*	5	0	0
	ethics AND "artificial intelligence" AND health	3	2	2
	AI or Artificial Intelligence or Fair AND ethic or moral or health AND health[12]	9	0	0
Google Scholar	ethics algorithms health	15,400	18	18
	ethics of machine learning in health	21,300	11	10
	ETHICS & HEALTH and at least one of: algorithm OR machine learning OR artificial intelligence OR AI	716,000	2	1
	ETHICS & HEALTH and at least one of: algorithm OR AI	105,000	2	2
	MORAL & HEALTH and at least one of: algorithm OR AI	26,900	2	1
	FAIR & HEALTH And at least one of: algorithm OR AI	38,000	0	0
PubMed	ETHICS & ARTIFICIAL INTELLIGENCE OR MACHINE LEARNING	34,193	37	37
Total		958,645	223	147

Table 4: Showing the final results from all searches. It is important to note that multiple search queries were made to cover all the combinations and the numbers in the table thus represent the sum of results, titles evaluated and downloaded (not all found files were accessible for download). It is also important to note that only the first 500 most relevant results from Google Scholar were reviewed and anything written before 2014 was excluded to make the number of results more manageable.

Funding
XXX work was partially supported by Privacy and Trust Stream – Social lead of the PETRAS Internet of Things research hub – PETRAS is funded by the UK Engineering and Physical Sciences Research Council (EPSRC), grant agreement no. EP/N023013/1 XXX work was also partially supported by a Microsoft grant and a Google grant.

References

Aicardi, C., Del Savio, L., Dove, E. S., Lucivero, F., Tempini, N., & Prainsack, B. (2016). Emerging ethical issues regarding digital health data. On the World Medical Association Draft Declaration on Ethical Considerations Regarding Health Databases and Biobanks. *Croatian Medical Journal*, *57*(2), 207–213. https://doi.org/10.3325/cmj.2016.57.207

Aitken, M., Tully, M. P., Porteous, C., Denegri, S., Cunningham-Burley, S., Banner, N., Black, C., Burgess, M., Cross, L., Van Delden, J., Ford, E., Fox, S., Fitzpatrick, N., Gallacher, K., Goddard, C., Hassan, L., Jamieson, R., Jones, K. H., Kaarakainen, M., … Willison, D. J. (2019). Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research. *International Journal of Population Data Science*, *4*(1). https://doi.org/10.23889/ijpds.v4i1.586

Álvarez-Machancoses, Ó., & Fernández-Martínez, J. L. (2019). Using artificial intelligence methods to speed up drug discovery. *Expert Opinion on Drug Discovery*, *14*(8), 769–777. https://doi.org/10.1080/17460441.2019.1621284

Andorno, R. (2004). The right not to know: An autonomy based approach. *Journal of Medical Ethics*, *30*(5), 435–439. https://doi.org/10.1136/jme.2002.001578

Arieno, A., Chan, A., & Destounis, S. V. (2019). A review of the role of augmented intelligence in breast imaging: From automated breast density assessment to risk stratification. *American Journal of Roentgenology*, *212*(2), 259–270. https://doi.org/10.2214/AJR.18.20391

Balthazar, P., Harri, P., Prater, A., & Safdar, N. M. (2018). Protecting Your Patients' Interests in the Era of Big Data, Artificial Intelligence, and Predictive Analytics. *Journal of the American College of Radiology*, *15*(3), 580–586. https://doi.org/10.1016/j.jacr.2017.11.035

Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. *IEEE Transactions on Information Technology in Biomedicine*, *14*(4), 1114–1120. https://doi.org/10.1109/TITB.2009.2039485

Bartoletti, I. (2019). AI in healthcare: Ethical and privacy challenges. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11526 LNAI*, 7–10. https://doi.org/10.1007/978-3-030-21642-9_2

Barton, C., Chettipally, U., Zhou, Y., Jiang, Z., Lynn-Palevsky, A., Le, S., Calvert, J., & Das, R. (2019). Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs. *Computers in Biology and Medicine*, *109*, 79–84. https://doi.org/10.1016/j.compbiomed.2019.04.027

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed). Oxford University Press.

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society*, *20*(1), 1–13. https://doi.org/10.1080/1369118X.2016.1216147

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, *97*(1), 262–267. https://doi.org/10.1073/pnas.97.1.262

Burns, T. (2015). *Our necessary shadow: The nature and meaning of psychiatry*. Pegasus Books.

Burr, C., Mariarosaria, T., & Floridi, L. (2019, February 1). *The Ethics of Digital Well-Being: A Thematic Review*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3338441&download=yes

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended Consequences of Machine Learning in Medicine. *JAMA*, *318*(6), 517. https://doi.org/10.1001/jama.2017.7797

Car, J., Sheikh, A., Wicks, P., & Williams, M. S. (2019). Beyond the hype of big data and artificial intelligence: Building foundations for knowledge and wisdom. *BMC Medicine*, *17*(1). https://doi.org/10.1186/s12916-019-1382-x

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

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, *28*(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care—Addressing Ethical Challenges. *The New England Journal of Medicine*, *378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Cheney-Lippold, J. (2017). *We are data: Algorithms and the making of our digital selves*. New York University Press.

Chin-Yee, B., & Upshur, R. (2019). Three Problems with Big Data and Artificial Intelligence in Medicine. *Perspectives in Biology and Medicine*, *62*(2), 237–256. https://doi.org/10.1353/pbm.2019.0012

Coeckelberg, M. (2014). Good Healthcare Is in the "How": The Quality of Care, the Role of Machines, and the Need for New Skills. In *Machine medical ethics* (Vol. 74). Springer.

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

Cookson, C. (2018, September 6). Artificial intelligence faces public backlash, warns scientist. *Financial Times*. https://www.ft.com/content/0b301152-b0f8-11e8-99ca-68cf89602132

Cowie, J., Calveley, E., Bowers, G., & Bowers, J. (2018). Evaluation of a digital consultation and self-care advice tool in primary care: A multi-methods study. *International Journal of Environmental Research and Public Health*, *15*(5). https://doi.org/10.3390/ijerph15050896

Dai, W., Yoshigoe, K., & Parsley, W. (2018). Improving Data Quality through Deep Learning and Statistical Models. *ArXiv:1810.07132 [Cs]*, *558*, 515–522. https://doi.org/10.1007/978-3-319-54978-1_66

Danis, M., & Solomon, M. (2013). Providers, Payers, The Community, And Patients Are All Obliged To Get Patient Activation And Engagement Ethically Right. *Health Affairs*, *32*(2), 401–407. https://doi.org/10.1377/hlthaff.2012.1081

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, *24*(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

De Langavant, L. C., Bayen, E., & Yaffe, K. (2018). Unsupervised machine learning to identify high likelihood of dementia in population-based surveys: Development and validation study. *Journal of Medical Internet Research*, *20*(7). https://doi.org/10.2196/10493

Deng, X., Luo, Y., & Wang, C. (2018). Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods. *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 631–635. https://doi.org/10.1109/CCIS.2018.8691126

Department of Health and Social Care. (n.d.-a). *Annual Report of the Chief Medical Office 2018: Health 2040—Better Health Within Reach*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/767549/Annual_report_of_the_Chief_Medical_Officer_2018_-_health_2040_-_better_health_within_reach.pdf

Department of Health and Social Care. (n.d.-b). *Health Secretary announces £250 million investment in artificial intelligence* [Gov.uk]. Retrieved 8 August 2019, from https://www.gov.uk/government/news/health-secretary-announces-250-million-investment-in-artificial-intelligence

Di Nucci, E. (2019). Should we be afraid of medical AI? *Journal of Medical Ethics*. https://doi.org/10.1136/medethics-2018-105281

Ding, Y., Tang, J., & Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing*, *325*, 211–224. https://doi.org/10.1016/j.neucom.2018.10.028

Dong, R., Yang, X., Zhang, X., Gao, P., Ke, A., Sun, H., Zhou, J., Fan, J., Cai, J., & Shi, G. (2019). Predicting overall survival of patients with hepatocellular carcinoma using a three-category method based on DNA methylation and machine learning. *Journal of Cellular and Molecular Medicine*, *23*(5), 3369–3374. https://doi.org/10.1111/jcmm.14231

Dudley, J. T., Listgarten, J., Stegle, O., Brenner, S. E., & Parts, L. (2015). *Personalized medicine: From genotypes, molecular phenotypes and the quantified self, towards improved medicine*. 342–346.

DuFault, B. L., & Schouten, J. W. (2018). Self-quantification and the datapreneurial consumer identity. *Consumption Markets & Culture*, 1–27. https://doi.org/10.1080/10253866.2018.1519489

Emanuel, E. J., & Wachter, R. M. (2019). Artificial Intelligence in Health Care: Will the Value Match the Hype? *JAMA*, *321*(23), 2281–2282. https://doi.org/10.1001/jama.2019.4914

Faden, R. R., Kass, N. E., Goodman, S. N., Pronovost, P., Tunis, S., & Beauchamp, T. L. (2013). An Ethics Framework for a Learning Health Care System: *A Departure from Traditional Research Ethics and Clinical Ethics*. *Hastings Center Report*, *43*(s1), S16–S27. https://doi.org/10.1002/hast.134

Fleming, N. (2018). How artificial intelligence is changing drug discovery. *Nature*, *557*(7707), S55–S57. https://doi.org/10.1038/d41586-018-05267-x

Floridi, L. (2008). The Method of Levels of Abstraction. *Minds and Machines*, *18*(3), 303–329. https://doi.org/10.1007/s11023-008-9113-7

Floridi, L. (2013). Distributed Morality in an Information Society. *Science and Engineering Ethics*, *19*(3), 727–743. https://doi.org/10.1007/s11948-012-9413-4

Floridi, L. (2014). *The 4th revolution: How the infosphere is reshaping human reality*. Oxford Univ. Press.

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L. (2017a). Digital's Cleaving Power and Its Consequences. *Philosophy & Technology*, *30*(2), 123–129. https://doi.org/10.1007/s13347-017-0259-1

Floridi, L. (2017b). The Logic of Design as a Conceptual Logic of Information. *Minds and Machines*, *27*(3), 495–519. https://doi.org/10.1007/s11023-017-9438-1

Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *376*(2133). https://doi.org/10.1098/rsta.2018.0081

Floridi, L. (2019a). AI opportunities for healthcare must not be wasterd. *Health Management*, *19*.

Floridi, L. (2019b). What the Near Future of Artificial Intelligence Could Be. *Philosophy & Technology*, *32*(1), 1–15. https://doi.org/10.1007/s13347-019-00345-y

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

Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction*, *11*(2), 63–125. https://doi.org/10.1561/1100000015

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

Garcia, J., Romero, N., Keyson, D., & Havinga, P. (2014). Reflective healthcare systems: Mirco-cylce of self-reflection to empower users. *Interaction Design and Architecture(s)*, *23*(1), 173–190.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). Datasheets for Datasets. *ArXiv:1803.09010 [Cs]*. http://arxiv.org/abs/1803.09010

Gillespie, T., Boczkowski, P. J., & Foot, K. A. (2014). *Media technologies: Essays on communication, materiality, and society*. The MIT Press.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies: A typology of reviews, *Maria J. Grant & Andrew Booth. Health Information & Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Greaves, F., Joshi, I., Campbell, M., Roberts, S., Patel, N., & Powell, J. (2018). What is an appropriate level of evidence for a digital health intervention? *The Lancet*, *392*(10165), 2665–2667. https://doi.org/10.1016/S0140-6736(18)33129-5

Hailu, R. (2019). Fitbits and other wearables may not accurately track heart rates in people of color. *STAT*. https://www.statnews.com/2019/07/24/fitbit-accuracy-dark-skin/

Harerimana, G., Jang, B., Kim, J. W., & Park, H. K. (2018). Health Big Data Analytics: A Technology Survey. *IEEE Access*, *6*, 65661–65678. https://doi.org/10.1109/ACCESS.2018.2878254

Hay, S. I., George, D. B., Moyes, C. L., & Brownstein, J. S. (2013). *Big data opportunities for global infectious disease surveillance*.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0

Hoffman, L., Benedetto, E., Huang, H., Grossman, E., Kaluma, D., Mann, Z., & Torous, J. (2019). Augmenting Mental Health in Primary Care: A 1-Year Study of Deploying Smartphone Apps in a Multi-site Primary Care/Behavioral Health Integration Program. *Frontiers in Psychiatry*, *10*, 94. https://doi.org/10.3389/fpsyt.2019.00094

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *ArXiv:1805.03677 [Cs]*. http://arxiv.org/abs/1805.03677

Holzinger, A., Haibe-Kains, B., & Jurisica, I. (2019). Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging*. https://doi.org/10.1007/s00259-019-04382-9

Information Commissioner. (2018, June 6). *Royal Free—Google DeepMind trial failed to comply with data protection law*. https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2017/07/royal-free-google-deepmind-trial-failed-to-comply-with-data-protection-law/

Information Commissioner. (2019, June 3). *Project ExplAIn interim report*. https://ico.org.uk/about-the-ico/research-and-reports/project-explain-interim-report/

Jiang, W., & Yin, Z. (2015). Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks. *Proceedings of the 23rd ACM International Conference on Multimedia - MM '15*, 1307–1310. https://doi.org/10.1145/2733373.2806333

Juengst, E., McGowan, M. L., Fishman, J. R., & Settersten, R. A. (2016). From "Personalized" to "Precision" Medicine: The Ethical and Social Implications of Rhetorical Reform in Genomic Medicine. *Hastings Center Report*, *46*(5), 21–33. https://doi.org/10.1002/hast.614

Kalmady, S. V., Greiner, R., Agrawal, R., Shivakumar, V., Narayanaswamy, J. C., Brown, M. R. G., Greenshaw, A. J., Dursun, S. M., & Venkatasubramanian, G. (2019). Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. *Npj Schizophrenia*, *5*(1), 2. https://doi.org/10.1038/s41537-018-0070-8

Kim, H., & Xie, B. (2017). Health literacy in the eHealth era: A systematic review of the literature. *Patient Education and Counseling*, *100*(6), 1073–1082.

King, D., Karthikesalingam, A., Hughes, C., Montgomery, H., Raine, R., Rees, G., & On behalf of the DeepMind Health Team. (2018). Letter in response to Google DeepMind and healthcare in an age of algorithms. *Health and Technology*, *8*(1), 11–13. https://doi.org/10.1007/s12553-018-0228-4

Kluge, E.-H., Lacroix, P., & Ruotsalainen, P. (2018). Ethics Certification of Health Information Professionals. *Yearbook of Medical Informatics*, *27*(01), 037–040. https://doi.org/10.1055/s-0038-1641196

Kohli, M., & Geis, R. (2018). Ethics, Artificial Intelligence, and Radiology. *Journal of the American College of Radiology*, *15*(9), 1317–1319. https://doi.org/10.1016/j.jacr.2018.05.020

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Kunapuli, G., Varghese, B. A., Ganapathy, P., Desai, B., Cen, S., Aron, M., Gill, I., & Duddalwar, V. (2018). A Decision-Support Tool for Renal Mass Classification. *Journal of Digital Imaging*, *31*(6), 929–939. https://doi.org/10.1007/s10278-018-0100-0

Kuo, W.-J., Chang, R.-F., Chen, D.-R., & Lee, C. C. (2001). Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. *Breast Cancer Research and Treatment*, *66*(1), 51–57. https://doi.org/10.1023/A:1010676701382

Li, B., Li, J., Jiang, Y., & Lan, X. (2019). Experience and reflection from China's Xiangya medical big data project. *Journal of Biomedical Informatics*, *93*. https://doi.org/10.1016/j.jbi.2019.103149

López-Martínez, F., Núñez-Valdez, E. R., Lorduy Gomez, J., & García-Díaz, V. (2019). A neural network approach to predict early neonatal sepsis. *Computers & Electrical Engineering*, *76*, 379–388. https://doi.org/10.1016/j.compeleceng.2019.04.015

Lu, F. S., Hattab, M. W., Clemente, C. L., Biggerstaff, M., & Santillana, M. (2019). Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-08082-0

Lu, H., & Wang, M. (2019). RL4health: Crowdsourcing Reinforcement Learning for Knee Replacement Pathway Optimization. *ArXiv:1906.01407 [Cs, Stat]*. http://arxiv.org/abs/1906.01407

Mahomed, S. (2018). Healthcare, artificial intelligence and the Fourth Industrial Revolution: Ethical, social and legal considerations. *South African Journal of Bioethics and Law*, *11*(2), 93. https://doi.org/10.7196/SAJBL.2018.v11i2.00664

Martani, A., Shaw, D., & Elger, B. S. (2019). Stay fit or get bit—Ethical issues in sharing health data with insurers' apps. *Swiss Medical Weekly*, *149*, w20089. https://doi.org/10.4414/smw.2019.20089

Matthias, A. (2015). Robot Lies in Health Care: When Is Deception Morally Permissible? *Kennedy Institute of Ethics Journal*, *25*(2), 169–162. https://doi.org/10.1353/ken.2015.0007

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

McLaughlin, K. (2016). *Empowerment: A critique*. http://public.eblib.com/choice/publicfullrecord.aspx?p=4332655

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Mittelstadt, B. (2019). The Ethics of Biomedical 'Big Data' Analytics. *Philosophy & Technology*, *32*(1), 17–21. https://doi.org/10.1007/s13347-019-00344-z

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679

Morley, J, & Joshi, I. (2019). Developing effective policy to support Artificial Intelligence in Health and Care. *Eurohealth*, *25*(2). https://doi.org/Forthcoming

Morley, J, & Floridi, L. (2019a). The Limits of Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-019-00115-1

Morley, J, & Floridi, L. (2019b). Enabling digital health companionship is better than empowerment. *The Lancet Digital Health*, S2589750019300792. https://doi.org/10.1016/S2589-7500(19)30079-2

Morley, J, & Floridi, L. (2020a). How to design a governable digital health ecosystem. *Digital Ethics Lab Yearbook 2020*. https://doi.org/10.13140/rg.2.2.28320.74244/1

Morley, J, & Floridi, L. (2020b). An ethically mindful approach to AI for health care. *The Lancet*, *395*(10220), 254–255. https://doi.org/10.1016/S0140-6736(19)32975-7

Morley, J, Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*. http://arxiv.org/abs/1905.06876 (Pre-Print)

Mortazavi, B. J., Desai, N., Zhang, J., Coppi, A., Warner, F., Krumholz, H. M., & Negahban, S. (2017). Prediction of Adverse Events in Patients Undergoing Major Cardiovascular Procedures. *IEEE Journal of Biomedical and Health Informatics*, *21*(6), 1719–1729. https://doi.org/10.1109/JBHI.2017.2675340

Moscoso, A., Silva-Rodríguez, J., Aldrey, J. M., Cortés, J., Fernández-Ferreiro, A., Gómez-Lado, N., Ruibal, Á., & Aguiar, P. (2019). Prediction of Alzheimer's disease dementia with MRI beyond the short-term: Implications for the design of predictive models. *NeuroImage: Clinical*, *23*, 101837. https://doi.org/10.1016/j.nicl.2019.101837

Nag, N., Pandey, V., Oh, H., & Jain, R. (2017). Cybernetic Health. *ArXiv:1705.08514 [Cs]*. http://arxiv.org/abs/1705.08514

National Institute for Health Research. (2019, June 14). *Involving the public in complex questions around artificial intelligence research*. https://www.nihr.ac.uk/blog/involving-the-public-in-complex-questions-around-artificial-intelligence-research/12236

Nebeker, C., Torous, J., & Bartlett Ellis, R. J. (2019). Building the case for actionable ethics in digital health research supported by artificial intelligence. *BMC Medicine*, *17*(1). https://doi.org/10.1186/s12916-019-1377-7

Nelson, A., Herron, D., Rees, G., & Nachev, P. (2019). Predicting scheduled hospital attendance with artificial intelligence. *Npj Digital Medicine*, *2*(1), 26. https://doi.org/10.1038/s41746-019-0103-3

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, *20*(5), e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4

NHS England. (2019). *The NHS Long Term Plan*. NHS. https://www.longtermplan.nhs.uk/wp-content/uploads/2019/01/nhs-long-term-plan.pdf

Ortiz, J. R., Zhou, H., Shay, D. K., Neuzil, K. M., Fowlkes, A. L., & Goss, C. H. (2011). Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends. *PLoS ONE*, *6*(4), e18687. https://doi.org/10.1371/journal.pone.0018687

Owens, J., & Cribb, A. (2019). 'My Fitbit Thinks I Can Do Better!' Do Health Promoting Wearable Technologies Support Personal Autonomy? *Philosophy & Technology*, *32*(1), 23–38. https://doi.org/10.1007/s13347-017-0266-2

Panch, T., Mattie, H., & Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. *Npj Digital Medicine*, *2*(1), 77. https://doi.org/10.1038/s41746-019-0155-4

Petrini, C. (2015). On the 'pendulum' of bioethics. *Clinica Terapeutica*, *166*(2), 82–84. https://doi.org/10.7417/CT.2015.1821

Ploug, T., & Holm, S. (2019). The right to refuse diagnostics and treatment planning by artificial intelligence. *Medicine, Health Care, and Philosophy*. https://doi.org/10.1007/s11019-019-09912-8

Popkes, A.-L., Overweg, H., Ercole, A., Li, Y., Hernández-Lobato, J. M., Zaykov, Y., & Zhang, C. (2019). Interpretable Outcome Prediction with Sparse Bayesian Neural Networks in Intensive Care. *ArXiv:1905.02599 [Cs, Stat]*. http://arxiv.org/abs/1905.02599

Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. *Health and Technology*, 1–17. https://doi.org/10.1007/s12553-017-0179-1

Powles, J., & Hodson, H. (2018). Response to DeepMind. *Health and Technology*, *8*(1), 15–29. https://doi.org/10.1007/s12553-018-0226-6

Price, W. N. (2018). Medical Malpractice and Black-Box Medicine. In I. G. Cohen, H. F. Lynch, E. Vayena, & U. Gasser (Eds.), *Big Data, Health Law, and Bioethics* (1st ed., pp. 295–306). Cambridge University Press. https://doi.org/10.1017/9781108147972.027

Price, W. N., Gerke, S., & Cohen, I. G. (2019). Potential Liability for Physicians Using Artificial Intelligence. *JAMA*. https://doi.org/10.1001/jama.2019.15064

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

Rahwan, I. (2018). Society-in-the-Loop: Programming the Algorithmic Social Contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, *169*(12), 866. https://doi.org/10.7326/M18-1990

Rich, E., & Miah, A. (2014). Understanding Digital Health as Public Pedagogy: A Critical Framework. *Societies*, *4*(2), 296–315. https://doi.org/10.3390/soc4020296

Rosenfeld, A., Benrimoh, D., Armstrong, C., Mirchi, N., Langlois-Therrien, T., Rollins, C., Tanguay-Sela, M., Mehltretter, J., Fratila, R., Israel, S., Snook, E., Perlman, K., Kleinerman, A., Saab, B., Thoburn, M., Gabbay, C., & Yaniv-Rosenfeld, A. (2019). Big Data Analytics and AI in Mental Healthcare. *ArXiv:1903.12071 [Cs]*. http://arxiv.org/abs/1903.12071

Ruckenstein, M., & Schüll, N. D. (2017). The Datafication of Health. *Annual Review of Anthropology*, *46*(1), 261–278. https://doi.org/10.1146/annurev-anthro-102116-041244

Schönberger, D. (2019). Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology*, *27*(2), 171–203. https://doi.org/10.1093/ijlit/eaz004

Sethi, N., & Laurie, G. T. (2013). Delivering proportionate governance in the era of eHealth: Making linkage and privacy work together. *Medical Law International*, *13*(2–3), 168–204. https://doi.org/10.1177/0968533213508974

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

Sterckx, S., Rakic, V., Cockbain, J., & Borry, P. (2016). "You hoped we would sleep walk into accepting the collection of our data": Controversies surrounding the UK care.data scheme and their wider relevance for biomedical research. *Medicine, Health Care and Philosophy*, *19*(2), 177–190. https://doi.org/10.1007/s11019-015-9661-6

Sullivan, H. R., & Schweikart, S. J. (2019). Are current tort liability doctrines adequate for addressing injury caused by AI? *AMA Journal of Ethics*, *21*(2), 160–166. https://doi.org/10.1001/amajethics.2019.160

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good: *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., & Yildirim, M. B. (2018). Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decision Support Systems*, *106*, 97–109. https://doi.org/10.1016/j.dss.2017.12.004

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105–112. https://doi.org/10.1007/s10676-009-9187-9

Vayena, E., Tobias, H., Afua, A., & Allesandro, B. (2018). Digital health: Meeting the ethical and policy challenges. *Swiss Medical Weekly*, *148*(34). https://doi.org/10.4414/smw.2018.14571

Vayena, Effy, Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Vayena, Effy, Salathé, M., Madoff, L. C., & Brownstein, J. S. (2015). Ethical Challenges of Big Data in Public Health. *PLOS Computational Biology*, *11*(2), e1003904. https://doi.org/10.1371/journal.pcbi.1003904

Voigt, K. (2019). Social Justice, Equality and Primary Care: (How) Can 'Big Data' Help? *Philosophy & Technology*, *32*(1), 57–68. https://doi.org/10.1007/s13347-017-0270-6

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., & Ghani, R. (2018). *Machine learning and AI research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics and Effectiveness*. 25.

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., Granger, D., Birse, M., Branson, R., Moons, K. G., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H. (2018). Machine learning and AI research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics and Effectiveness. *ArXiv:1812.10404 [Cs, Stat]*. http://arxiv.org/abs/1812.10404

Wang, Z. (2019). Data integration of electronic medical record under administrative decentralization of medical insurance and healthcare in China: A case study. *Israel Journal of Health Policy Research*, *8*(1). https://doi.org/10.1186/s13584-019-0293-9

Wardrope, A. (2015). Relational Autonomy and the Ethics of Health Promotion. *Public Health Ethics*, *8*(1), 50–62. https://doi.org/10.1093/phe/phu025

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: Beyond the black box. *BMJ*, *364*, l886. https://doi.org/10.1136/bmj.l886

World Health Organisation. (n.d.). *Big data and artificial intelligence*. Retrieved 29 June 2019, from https://www.who.int/ethics/topics/big-data-artificial-intelligence/en/

Zacher, B., & Czogiel, I. (2019). Supervised learning improves disease outbreak detection. *ArXiv:1902.10061 [Cs, Stat]*. http://arxiv.org/abs/1902.10061

Zhang, L., Wang, H., Li, Q., Zhao, M.-H., & Zhan, Q.-M. (2018). Big data and medical research in China. *BMJ*, j5910. https://doi.org/10.1136/bmj.j5910

The Ethics of AI in Health Care: a Mapping Review

	Ethical Concern	Explanation	Medical Example
Epistemic concerns	Inconclusive Evidence	Algorithmic outcomes (e.g. classification) are probabilistic and not infallible. They are rarely sufficient to posit the existence of a causal relationship.	*EKG readers in smartwatches may 'diagnose' a patient as suffering from arrhythmia when it may be due to a fault with the watch not being able to accurately read that user's heartbeat (for example due to the colour of their skin) or the 'norm' is inappropriately calibrated for that individual (Hailu, 2019)*
	Inscrutable Evidence	Recipients of an algorithmic decision very rarely have full oversight of the data used to train or test an algorithm or the data points used to reach a specific decision.	*A clinical decision support system deployed in a hospital may make a treatment recommendation, but it may not be clear on what basis it has made that 'decision' raising the risk that it has used data that are inappropriate for the individual in question or that there is a bug in the system leading to issues with over or under prescribing (Wachter, 2015).*
	Misguided Evidence	Algorithmic outcomes can only be as reliable (but also as neutral) as the data they are based on.	*Watson for Oncology is in widespread use in China for 'diagnosis' via image recognition but has primarily been trained on a Western data set leading to issues with concordance and poorer results for Chinese patients than their Western counterparts (Liu et al., 2018).*
Normative Concerns	Unfair outcomes	An action can be found to having more of an impact (positive or negative) on one group of people	An algorithm *'learns' to prioritise patients it predicts to have better outcomes for a particular disease. This turns out to have a discriminatory effect on people within the Black and minority ethnic communities (Garattini et al., 2019).*
	Transformative effects	Algorithmic activities, like profiling, re-conceptualise reality in unexpected ways.	*An individual using personal health app has limited oversight over what passive data it is collecting and how that is being transformed into a recommendation to improve, limiting their ability to challenge any recommendations made and a loss of personal autonomy and data privacy (Kleinpeter, 2017).*
Overarching	Traceability	Harm caused by algorithmic activity is hard to debug (to detect the harm and find its cause), and it is hard to identify who should be held responsible for the harm caused.	*If a decision made by clinical decision support software leads to a negative outcome for the individual, it is unclear who to assign the responsibility and or liability to and therefore to prevent it from happening again (Racine et al., 2019)..*

Table 1: A summary of the epistemic, normative and overarching ethical concerns related to algorithmic use in healthcare based on Mittelstadt et al (2016) from (Redacted for anonymity) .

	A. Individual	**B. Interpersonal**	**C. Group**	**D. Institutional**	**E. Sectoral**	**F. Societal**
1. Epistemic concern (inconclusive, inscrutable and misguided evidence)	Misdiagnosis or missed diagnosis	Loss of trust in HCP-Patient relationships, de-personalisation of care	Misdiagnosis or missed diagnosis at scale – some groups more affected than others	Waste of funds and resources not directed to areas of greater need	Excessively broad data sharing between public and private entities	Poorer public healthcare provision and worsening health outcomes for society
2. Normative (transformative effects and unfair outcomes)	Surveillance & undermining of autonomy and integrity of self	Deskilling of HCPs, overreliance on AI-tools, and undermining of consent practices and redefining roles in the healthcare system	Profiling and discrimination against certain groups seen as being less healthy or higher risk	Transformation of care pathways & imposing of specific values at scale – redefining 'good care'	Siloing of new AI tool development within private sector	Inequalities in outcomes
3. Overarching (traceability)	'Bad Users' could come to be blamed for their own ill-health	See institutional	Specific groups framed as being more morally irresponsible with regards to their health than others	Lack of clarity over liability with regards to issues with safety and effectiveness could halt adoption or result in certain groups being blamed more often than others	See institutional	Society must decide through regulation preferable answers to the questions regarding liability and risk allocation in healthcare provision. However, all groups in society may not be given an equal say in this process

Table 2: Summary of the epistemic, normative and overarching ethical concerns associated with AI-Health at the six different LoAs as identified by the literature review

Consideration	**Key supporting literature**	**Relevant aspects (ascending LoA[1])**	**Example body responsible for answering this question based on the English National Health System**
What skills will the professional healthcare workforce require in order to make safe and effective use of AI-Health solutions in the future?	(Kluge et al., 2018)	Epistemic (A, B, C, F) Normative (B, C, D, E) Overarching (A, C)	Health Education England should survey the skills currently available in the workforce and conduct a gap analysis of the skills that will be needed
Which tasks should be delegated to AI-Health solutions, and which should not?	(Di Nucci, 2019)	Epistemic (A, B, C, D, F) Normative (B, C, D, F) Overarching (A, C, D)	Department of Health and Social Care should conduct a multi-stakeholder engagement process to understand which tasks are socially acceptable to be delegated to AI-health and make this offficial policy,.
What evidence is needed to 'prove' clinical effectiveness of an AI-Health solution?	(Greaves et al., 2018)	Epistemic (A, B, C, E, F) Normative (E) Overarching (A, C, D, F)	Medicines and Healthcare Regulators Medicines and Healthcare products Regulatory Agency should update the medical device regulations to include a minimum required standard of accuracy and a minimum standard of

[1] Denoted by an increasing Level of Analysis: Individual (A), Interpersonal (B), Group (C), Institutional (D), Sectoral (E) and Societal (F).

			evidence to demonstrate that the AI-health product is genuinely capable of performing at this level
What mechanisms should be put in place to enable people to report and seek redress for AI-Health associated harms?	(Schönberger, 2019)	Epistemic (A, C, E, F) Normative (A, C, E, F) Overarching (A, C, D)	The Care Quality Commission should update its inspection framework to regularly check that AI-health products in use are continuing to operate safely. Medicines and Healthcare Regulators Medicines and Healthcare products Regulatory Agency should introduce a 'yellow card' scheme for AI-health products so that users can report errors and be assured that they are being taken care of.
What mechanisms should be put in place to ensure all relevant stakeholder views are included in the development of AI-Health solutions?	(Aitken et al., 2019)	Epistemic (C, E, F) Overarching (A, C, D, F)	The Health Research Authority should update its guidance on ethical approval for AI-health research and product development to set out the minimum participation requirements for diverse stakeholders
How can the explainability of AI-Health solutions be guaranteed?	(Watson et al., 2019)	Epistemic (A, C) Normative (A, C, E) Overarching (A, D)	Medicines and Healthcare Regulators Medicines and Healthcare products Regulatory Agency should update the regulations governing medical devices to set out the minimum standards for 'explainability' of AI-health products
What mechanisms can be put in place to ensure reliability, replicability and safety of AI-Health solutions?	(Challen et al., 2019)	Epistemic (A, C, F) Normative (C, E, F) Overarching (A, C, D)	The National Institute for Health and Care Excellence should make make it a requirement of formal health technology assessment that, within the bounds of technical feasibility and respecting intellectual property, developers make code open to enable reproducibility and error checking.
How can transparency over how algorithmic tools are integrated into the healthcare workflow, how it shapes decisions, and how it affects process optimization within medical services, be guaranteed?	(Vayena et al., 2015)	Epistemic (A, B, C, D, E, F) Normative (A, B, D, F) Overarching (A, D, F)	NHS England should make it a requirement of all NHS trusts, hospitals and providers of care to declare when an AI-health solution is being used in a specific care pathway and to be clear about how its safety and quality is being regularly assessed.
How can traditional and non-traditional sources of health data be incorporated into AI-Health decision making? And how can it be appropriately protected and how can it be harmonised?	(e.g. Maher et al., 2019; Ploug & Holm, 2016; Richardson, Milam, & Chrysler, 2015; Townend, 2018)	Epistemic (A, C, D, E, F) Normative (A, C, D, E, F) Overarching (A, C, D, E)	The Health Research Authority and NHS Digital should update guidance and regulations governing secondary uses of health data to incorporate the specific considerations of AI-Health as we have outlined in this paper/
How are bioethical concepts	(Mittelstadt, 2019)	Epistemic (B, F)	The Nuffield Council on

(beneficence, non-maleficence, autonomy and justice (Beauchamp & Childress, 2013) challenged by AI-Health?		Normative (A, C, D, F) Overarching (A, F)	Bioethics should update its guidance on the bioethical principles for data initiatives to incorporate AI-health specific considerations.
How can concepts such as fairness, accountability and transparency can be maintained at scale (Redacted for anonymity)?	(Rosenfeld et al., 2019)	Epistemic (C, D, E, F) Normative (D, E, F) Overarching (F)	The Care Quality Commission, should develop a mechanism for monitoring these impacts at scale as part of its regular review process that is designed to ensure safe and high-quality care. This may require the Department of Health and Social Care extending its regulatory powers.

Table 3: Eleven key considerations for policymakers that arose from the literature review, denoted by an increasing Level of Abstraction: Individual (A), Interpersonal (B), Group (C), Institutional (D), Sectoral (E) and Societal (F).

	Ethical Concern	Explanation	Medical Example
Epistemic concerns	Inconclusive Evidence	Algorithmic outcomes (e.g. classification) are probabilistic and not infallible. They are rarely sufficient to posit the existence of a causal relationship.	*EKG readers in smartwatches may 'diagnose' a patient as suffering from arrhythmia when it may be due to a fault with the watch not being able to accurately read that user's heartbeat (for example due to the colour of their skin) or the 'norm' is inappropriately calibrated for that individual (Hailu, 2019)*
	Inscrutable Evidence	Recipients of an algorithmic decision very rarely have full oversight of the data used to train or test an algorithm or the data points used to reach a specific decision.	*A clinical decision support system deployed in a hospital may make a treatment recommendation, but it may not be clear on what basis it has made that 'decision' raising the risk that it has used data that are inappropriate for the individual in question or that there is a bug in the system leading to issues with over or under prescribing (Wachter, 2015).*
	Misguided Evidence	Algorithmic outcomes can only be as reliable (but also as neutral) as the data they are based on.	*Watson for Oncology is in widespread use in China for 'diagnosis' via image recognition but has primarily been trained on a Western data set leading to issues with concordance and poorer results for Chinese patients than their Western counterparts (Liu et al., 2018).*
Normative Concerns	Unfair outcomes	An action can be found to having more of an impact (positive or negative) on one group of people	An algorithm *'learns' to prioritise patients it predicts to have better outcomes for a particular disease. This turns out to have a discriminatory effect on people within the Black and minority ethnic communities (Garattini, Raffle, Aisyah, Sartain, & Kozlakidis, 2019).*
	Transformative effects	Algorithmic activities, like profiling, re-conceptualise reality in unexpected ways.	*An individual using personal health app has limited oversight over what passive data it is collecting and how that is being transformed into a recommendation to improve, limiting their ability to challenge any recommendations made and a loss of personal autonomy and data privacy (Kleinpeter, 2017).*
Overarching	Traceability	Harm caused by algorithmic activity is hard to debug (to detect the harm and find its cause), and it is hard to identify who should be held responsible for the harm caused.	*If a decision made by clinical decision support software leads to a negative outcome for the individual, it is unclear who to assign the responsibility and or liability to and therefore to prevent it from happening again (Racine, Boehlen, & Sample, 2019)..*

Table 1: A summary of the epistemic, normative and overarching ethical concerns related to algorithmic use in healthcare based on Mittelstadt et al (2016) from (Jessica Morley & Floridi, 2020b) .

	A. Individual	B. Interpersonal	C. Group	D. Institutional	E. Sectoral	F. Societal
1. Epistemic concern (inconclusive, inscrutable and misguided evidence)	Misdiagnosis or missed diagnosis	Loss of trust in HCP-Patient relationships, de-personalisation of care	Misdiagnosis or missed diagnosis at scale – some groups more affected than others	Waste of funds and resources not directed to areas of greater need	Excessively broad data sharing between public and private entities	Poorer public healthcare provision and worsening health outcomes for society
2. Normative (transformative effects and unfair outcomes)	Surveillance & undermining of autonomy and integrity of self	Deskilling of HCPs, overreliance on AI-tools, and undermining of consent practices and redefining roles in the healthcare system	Profiling and discrimination against certain groups seen as being less healthy or higher risk	Transformation of care pathways & imposing of specific values at scale – redefining 'good care'	Siloing of new AI tool development within private sector	Inequalities in outcomes
3. Overarching (traceability)	'Bad Users' could come to be blamed for their own ill-health	See institutional	Specific groups framed as being more morally irresponsible with regards to their health than others	Lack of clarity over liability with regards to issues with safety and effectiveness could halt adoption or result in certain groups being blamed more often than others	See institutional	Society must decide through regulation preferable answers to the questions regarding liability and risk allocation in healthcare provision. However, all groups in society may not be given an equal say in this process

Table 2: Summary of the epistemic, normative and overarching ethical concerns associated with AI-Health at the six different LoAs as identified by the literature review

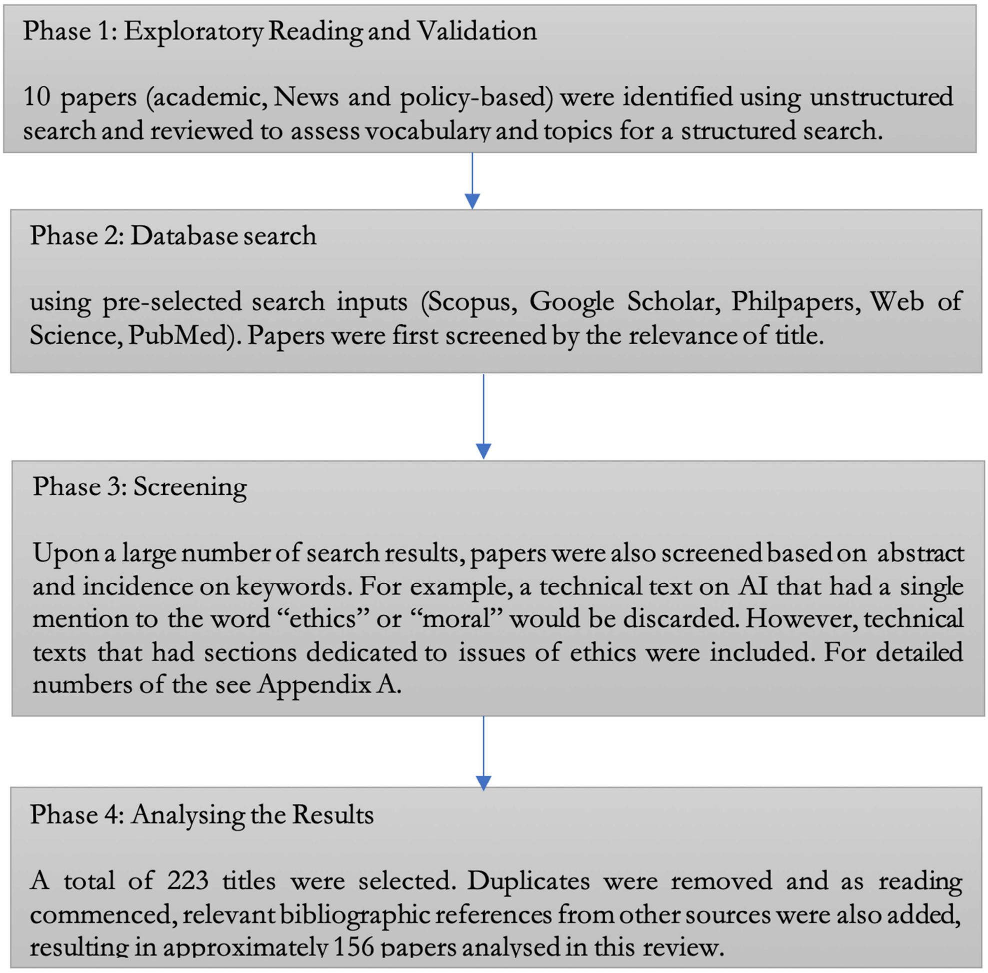
Consideration	Key supporting literature	Relevant aspects (ascending	Example body

		LoA[1])	**responsible for answering this question based on the English National Health System**
What skills will the professional healthcare workforce require in order to make safe and effective use of AI-Health solutions in the future?	(Kluge et al., 2018)	Epistemic (A, B, C, F) Normative (B, C, D, E) Overarching (A, C)	Health Education England should survey the skills currently available in the workforce and conduct a gap analysis of the skills that will be needed
Which tasks should be delegated to AI-Health solutions, and which should not?	(Di Nucci, 2019)	Epistemic (A, B, C, D, F) Normative (B, C, D, F) Overarching (A, C, D)	Department of Health and Social Care should conduct a multi-stakeholder engagement process to understand which tasks are socially acceptable to be delegated to AI-health and make this offficial policy,.
What evidence is needed to 'prove' clinical effectiveness of an AI-Health solution?	(Greaves et al., 2018)	Epistemic (A, B, C, E, F) Normative (E) Overarching (A, C, D, F)	Medicines and Healthcare Regulators Medicines and Healthcare products Regulatory Agency should update the medical device regulations to include a minimum required standard of accuracy and a minimum standard of evidence to demonstrate that the AI-health product is genuinely capable of performing at this level
What mechanisms should be put in place to enable people to report and seek redress for AI-Health associated harms?	(Schönberger, 2019)	Epistemic (A, C, E, F) Normative (A, C, E, F) Overarching (A, C, D)	The Care Quality Commission should update its inspection framework to regularly check that AI-health products in use are continuing to operate safely. Medicines and Healthcare Regulators Medicines and Healthcare products Regulatory Agency should introduce a 'yellow card' scheme for AI-health products so that users can report errors and be assured that they are being taken care of.
What mechanisms should be put in place to ensure all relevant stakeholder views are included in the development of AI-Health solutions?	(Aitken et al., 2019)	Epistemic (C, E, F) Overarching (A, C, D, F)	The Health Research Authority should update its guidance on ethical approval for AI-health research and product development to set out the minimum participation requirements for diverse stakeholders
How can the explainability of AI-Health solutions be guaranteed?	(Watson et al., 2019)	Epistemic (A, C) Normative (A, C, E) Overarching (A, D)	Medicines and Healthcare Regulators Medicines and Healthcare products Regulatory Agency should update the regulations governing medical devices to set out the minimum standards for 'explainability' of AI-health products
What mechanisms can be put in	(Challen et al., 2019)	Epistemic (A, C, F)	The National Institute for

[1] Denoted by an increasing Level of Analysis: Individual (A), Interpersonal (B), Group (C), Institutional (D), Sectoral (E) and Societal (F).

place to ensure reliability, replicability and safety of AI-Health solutions?		Normative (C, E, F) Overarching (A, C, D)	Health and Care Excellence should make make it a requirement of formal health technology assessment that, within the bounds of technical feasibility and respecting intellectual property, developers make code open to enable reproducibility and error checking.
How can transparency over how algorithmic tools are integrated into the healthcare workflow, how it shapes decisions, and how it affects process optimization within medical services, be guaranteed?	(Vayena et al., 2015)	Epistemic (A, B, C, D, E, F) Normative (A, B, D, F) Overarching (A, D, F)	NHS England should make it a requirement of all NHS trusts, hospitals and providers of care to declare when an AI-health solution is being used in a specific care pathway and to be clear about how its safety and quality is being regularly assessed.
How can traditional and non-traditional sources of health data be incorporated into AI-Health decision making? And how can it be appropriately protected and how can it be harmonised?	(e.g. Maher et al., 2019; Ploug & Holm, 2016; Richardson, Milam, & Chrysler, 2015; Townend, 2018)	Epistemic (A, C, D, E, F) Normative (A, C, D, E, F) Overarching (A, C, D, E)	The Health Research Authority and NHS Digital should update guidance and regulations governing secondary uses of health data to incorporate the specific considerations of AI-Health as we have outlined in this paper/
How are bioethical concepts (beneficence, non-maleficence, autonomy and justice (Beauchamp & Childress, 2013) challenged by AI-Health?	(Mittelstadt, 2019)	Epistemic (B, F) Normative (A, C, D, F) Overarching (A, F)	The Nuffield Council on Bioethics should update its guidance on the bioethical principles for data initiatives to incorporate AI-health specific considerations.
How can concepts such as fairness, accountability and transparency can be maintained at scale (Morley & Floridi, 2020a)?	(Rosenfeld et al., 2019)	Epistemic (C, D, E, F) Normative (D, E, F) Overarching (F)	The Care Quality Commission, should develop a mechanism for monitoring these impacts at scale as part of its regular review process that is designed to ensure safe and high-quality care. This may require the Department of Health and Social Care extending its regulatory powers.

Table 3: Eleven key considerations for policymakers that arose from the literature review, denoted by an increasing Level of Abstraction: Individual (A), Interpersonal (B), Group (C), Institutional (D), Sectoral (E) and Societal (F).



Phase 1: Exploratory Reading and Validation

10 papers (academic, News and policy-based) were identified using unstructured search and reviewed to assess vocabulary and topics for a structured search.

Phase 2: Database search

using pre-selected search inputs (Scopus, Google Scholar, Philpapers, Web of Science, PubMed). Papers were first screened by the relevance of title.

Phase 3: Screening

Upon a large number of search results, papers were also screened based on abstract and incidence on keywords. For example, a technical text on AI that had a single mention to the word "ethics" or "moral" would be discarded. However, technical texts that had sections dedicated to issues of ethics were included. For detailed numbers of the see Appendix A.

Phase 4: Analysing the Results

A total of 223 titles were selected. Duplicates were removed and as reading commenced, relevant bibliographic references from other sources were also added, resulting in approximately 156 papers analysed in this review.

Figure 1: Flowchart offering and overview of the steps taken in our literature review, filtering from several thousand titles to identified abstracts and selecting 156 papers to read.

The Ethics of AI in Health Care: a Mapping Review

Research Highlights

- This article maps the ethics of artificial intelligence in healthcare
- Ethical issues can be epistemic, normative, or related to traceability
- Issues affect individuals, relationships, groups, institutions, sectors, societies
- An agreed standard for ethical analysis is needed; split by issue and level