The ethics of AI in health care: A mapping review

Jessica Morleya,∗, Caio C.V. Machadoa,1, Christopher Burra, Josh Cowlsta,b, Indra Joshid, Mariarosaria Taddeoa,b,c, Luciano Floridia,b,c
a Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, OX1 3JS, UK
b Alan Turing Institute, British Library, 96 Euston Rd, London, NW1 2DB, UK
c Department of Computer Science, University of Oxford, 15 Parks Rd, Oxford, OX1 3QD, UK
d NHSX, Skipton House, 80 London Road, SE1 6LH, UK

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 effects; 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.

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 and Upshur, 2019). The argument stems not from the belief that all healthcare needs will soon be taken care of by “robot doctors” (Chin-Yee and 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 diagnose Diabetes Mellitus (Barakat

ARTICLE INFO

Keywords:
Artificial intelligence
Ethics
Healthcare
Health policies
Machine learning

Received in revised form 22 June 2020; Accepted 23 June 2020

☆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.

∗Corresponding author.
E-mail address: jessica.morley@oii.ox.ac.uk (J. Morley).
1These authors contributed equally to the writing of this paper.
et al., 2010); ensemble learning methods can predict outcomes for cancer patients (Kouroú et al., 2015); and neural networks can be used to diagnose stroke (Jiang 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, 2019a, 2019b), 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 and 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 and 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 (Lu and Wang, 2019; Nelson et al., 2019). However, as Ngiam and 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 (Nebeke 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 et al., 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 and 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 et al., 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 and 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, evidence of efficacy) to remain focused on mapping the current debate about the ethics of AI specifically. Records 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 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.

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 in the appendix illustrates our methodology. Also, a summary map of our findings (Table 2) is provided at the end of the section.

3.1. Epistemic concerns: inconclusive, inscrutable, and misguided evidence

Many factors are encouraging the development of AI-Health (Chin-Yee and 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 and 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 and 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); (3) population health monitoring (F. S. Lu et al., 2019; Zacher and Czogi, 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 et al., 2018a, 2018b), raising

<table>
<thead>
<tr>
<th>Ethical Concern</th>
<th>Explanation</th>
<th>Medical Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epistemic concerns</td>
<td>Inconclusive Evidence</td>
<td>Algorithmic outcomes (e.g. classification) are probabilistic and not infallible. They are rarely sufficient to posit the existence of a causal relationship.</td>
</tr>
<tr>
<td></td>
<td>Inscrutable Evidence</td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Misguided Evidence</td>
<td>Algorithmic outcomes can only be as reliable (but also as neutral) as the data they are based on.</td>
</tr>
<tr>
<td>Normative Concerns</td>
<td>Unfair outcomes</td>
<td>An action can be found to having more of an impact (positive or negative) on one group of people</td>
</tr>
<tr>
<td></td>
<td>Transformative effects</td>
<td>Algorithmic activities, like profiling, re-conceptualise reality in unexpected ways.</td>
</tr>
<tr>
<td></td>
<td>Traceability</td>
<td>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.</td>
</tr>
</tbody>
</table>
serious questions about the scientific rigor of AI-Health and its safety (Vayena et al., 2018a, 2018b). 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 and 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 and 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 impersonalisation of care provision (Jueungst 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’ (Jueungst 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 and 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 et al., 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 and 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 and 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 and 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 and 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
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., 2018a, 2018b). 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 and 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 and 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 and 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 and 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., 2018a, 2018b). 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 and 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 and 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 well-being (Burr et al., 2019)? 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 significant damage to public trust.

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 et al., 2018a, 2018b; Turilli and Floridi, 2009). This is making it increasingly challenging to 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 (Wardrop, 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 and 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 provides 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 and Miah, 2014) risks creating an elaborate mechanism for victim-blaming (Danis and 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 preempt 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 sociodemographic issues which entail harmful habits, and cannot be oversimplified to a matter of delivering the adequate information to the patient (Owens and 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 (Hallu, 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 system 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 and 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 and 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., 2018a, 2018b), 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 and 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.

### 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.

<table>
<thead>
<tr>
<th>A. Individual</th>
<th>B. Interpersonal</th>
<th>C. Group</th>
<th>D. Institutional</th>
<th>E. Sectoral</th>
<th>F. Societal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misdiagnosis or missed diagnosis</td>
<td>Surveillance &amp; profiling of personal data</td>
<td>'Bad Users' could come to be blamed for their own ill-health</td>
<td>Power public healthcare provision and resulting health outcomes for society</td>
<td>Inequalities in outcomes stemming from inaccurate coding or poor-quality training data</td>
<td>Scope of society must decide through regulation whether to have safety and effectiveness reviews at scale—impacting the ‘good care’ &amp; ‘bad users’ debate</td>
</tr>
<tr>
<td>Loss of trust in HCP-Patient relationship, de-personalisation</td>
<td>Profiling and discrimination on AI-tools, and undermining of autonomy and integrity of self</td>
<td>Specific groups framed as being more morally responsible for their health than others</td>
<td>Transforming care pathways &amp; redefining ‘good care’</td>
<td>Lack of clarity over liability with regards to who has the responsibility for the safety and effectiveness of an AI-powered medical product that causes harm to patients</td>
<td></td>
</tr>
<tr>
<td>Lack of trust in government, de-professionalisation of care</td>
<td>Specific groups framed as being more morally responsible for their health than others</td>
<td>See institutional</td>
<td>Sees institutional</td>
<td>See institutional</td>
<td></td>
</tr>
</tbody>
</table>

Table 2
4. The need for an ethically-blind 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 development lifecycle, 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 and 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 and 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 and 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, 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).

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) (Policy for Device Software Functions and Mobile Medical Applications, 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 et al., 2019). 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; Rahman, 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 and Deetjen, 2019), fair distribution of benefits (Balthazar et al., 2018) (Kohli and 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.

Funding

Taddeo’s and Floridi’s 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 Machado’s, Taddeo’s and Floridi’s work was also partially supported by a Microsoft grant and a Google grant.

Appendix A. Methodology

<table>
<thead>
<tr>
<th>Phase 1: Exploratory Reading and Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 papers (academic, News and policy-based) were identified using unstructured search and reviewed to assess vocabulary and topics for a structured search.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 2: Database search</th>
</tr>
</thead>
<tbody>
<tr>
<td>using pre-selected search inputs (Scopus, Google Scholar, Philpapapers, Web of Science, PubMed). Papers were first screened by the relevance of title.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 3: Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 4: Analysing the Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>
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.

Table 4

<table>
<thead>
<tr>
<th>Database</th>
<th>Search Query</th>
<th>Results</th>
<th>Titles Selected</th>
<th>Titles Downloaded</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCOPUS</td>
<td>ethic AND algorithm AND health</td>
<td>596</td>
<td>39</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>(ethic AND &quot;Artificial Intelligence&quot; OR ai) AND health*</td>
<td>259</td>
<td>37</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>(moral) AND &quot;Artificial Intelligence&quot; OR ai) AND health*</td>
<td>46</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(fair) AND &quot;Artificial Intelligence&quot; OR ai) AND health*</td>
<td>122</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>(moral) OR ethic* AND &quot;machine learning&quot; AND health*</td>
<td>91</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>(fair) AND &quot;machine learning&quot; AND health*</td>
<td>70</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Web of Science</td>
<td>(fair) OR moral* OR ethic* AND (&quot;machine learning&quot; OR &quot;Artificial Intelligence&quot; OR AI OR algorithm) AND health*</td>
<td>668</td>
<td>45</td>
<td>26</td>
</tr>
<tr>
<td>Philpapers</td>
<td>&quot;machine learning&quot; AND health*</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Artificial Intelligence AND health* AND ethic* algorithms AND health* ether*</td>
<td>1000+</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>ethics AND &quot;Artificial Intelligence&quot; AND health* AI or Artificial Intelligence or Fair AND ethic or moral or health AND health*</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>ethics algorithms health</td>
<td>15,400</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>ethics of machine learning in health</td>
<td>21,300</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>ETHICS &amp; HEALTH and at least one of: algorithm OR machine learning OR artificial intelligence OR AI</td>
<td>716,000</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ETHICS &amp; HEALTH and at least one of: algorithm OR AI</td>
<td>105,000</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>MORAL &amp; HEALTH and at least one of: algorithm OR AI</td>
<td>26,900</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>FAIR &amp; HEALTH And at least one of: algorithm OR AI</td>
<td>38,000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PubMed</td>
<td>ETHICS &amp; ARTIFICIAL INTELLIGENCE OR MACHINE LEARNING</td>
<td>34,193</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>958,645</td>
<td>223</td>
<td>147</td>
</tr>
</tbody>
</table>

References


Danis, M., Solomon, M., 2013. Providers, payers, the community, and patients are all stakeholders. In: Medical Machine Ethics, vol. 74. Springer.


