Routledge
Taylor & Francis Group

The New Bioethics

A Multidisciplinary Journal of Biotechnology and the Body

ISSN: 2050-2877 (Print) 2050-2885 (Online) Journal homepage: www.tandfonline.com/journals/ynbi20

Ethical Challenges to the Adoption of AI in Healthcare: A Review

Michal Pruski

To cite this article: Michal Pruski (2024) Ethical Challenges to the Adoption of AI in Healthcare: A Review, The New Bioethics, 30:4, 251-267, DOI: 10.1080/20502877.2025.2541438

To link to this article: https://doi.org/10.1080/20502877.2025.2541438

THE NEW BIOETHICS, Vol. 30 No. 4, 2024, 251–267

Ethical Challenges to the Adoption of AI in Healthcare: A Review

MICHAL PRUSKI
Faculty of Biology, Medicine and Health, The University of Manchester, UK
CEDAR, PIER, Cardiff and Value University Health Board, UK

There are few comprehensive summaries of the ethical challenges associated with the adoption of artificial intelligence in healthcare. This review utilizes a systematic search focused on identifying the barriers and facilitators to the implementation of artificial intelligence in healthcare, highlighting the diversity of ethical challenges and the complex interactions between practical challenges and ethics issues. For example, the quality of the data upon which artificial intelligence models are developed relates to several ethics principles, as does the issue of gaining user trust. Importantly, there is also the difficulty of achieving the right balance between the discussed principles, since one might not be able to maximize one principle without having to sacrifice another. For example, maximizing privacy might require minimizing data collection from patients, which might negatively affect beneficence. As such, this review highlights the variety and complexity of ethical issues associated with artificial intelligence implementation in healthcare.

KEYWORDS Artificial intelligence, digital health, machine learning, digital ethics

Introduction

This review was undertaken as part of an ongoing mixed-methods project looking at opinions of stakeholders on barriers and facilitators to the adoption of machine learning (ML) in Wales with regards to Value Based Healthcare (VBH). VBH is sometimes known as Prudent Healthcare (Bevan Commission 2022). Within this context, the empirical component of the project will focus on the use of Patient Reported Outcome Measures (PROMs) in healthcare decision-making. The aim of the review was to provide a global picture of the ethical challenges associated with the application of artificial intelligence (AI) in healthcare, so that the findings

DOI 10.1080/20502877.2025.2541438

of the review could be used as a basis for later parts of the project. The scope of the review was chosen to be broader than that of the empirical component of the study because: (1) it was anticipated that few studies would have addressed ethical challenges associated with the implementation of ML in VBH, (2) a narrow scope would risk the empirical study producing confirmation bias where only those phenomena identified by past researchers would be explored, while a wider scope had the potential to expand the range of potentially applicable phenomena, (3) applying a wider review scope makes it more useful for future studies and other researchers.

Methods

While a systematic review was not necessary to achieve the aim of this review, it was decided to utilize a literature search strategy which focused on identifying publications discussing barriers and facilitators to the adoption of AI in healthcare, and extract the subset of publications which mentioned ethics. This was done not to limit the content of the review, but to ensure a comprehensive picture of the issues mentioned in the field. The original search strategy and search results can be found in Supplementary File 1.

Using EndNote, the titles and abstracts of all publications containing the word 'ethic' (to account for such variants as 'ethics' or 'ethical') were identified. The publications were then assessed for suitability for inclusion at title / abstract level, and then at a full-text level; decisions for full-text inclusions and exclusions can be found in Supplementary File 2. During the full-text assessment key references were identified and their texts included. When writing the review, additional references were included to provide examples of the relevant issues, provide additional evidence or to provide background to the ethical issues mentioned. The issues discussed in each paper were extracted and grouped together.

Ethical issues for the adoption of AI in healthcare

The following discussion is based on the 11 ethics principles identified by Jobin *et al.* (2019) who reviewed 84 AI ethics guidelines and was further utilized by Goirand *et al.* (2021) who conducted a scoping review of frameworks for ethical AI implementation in healthcare which identified 33 such frameworks; Zhou *et al.* (2020) additionally identified an 85th framework, when reviewing Jobin and colleagues' work. These frameworks pertained to the whole lifecycle of AI (requirement capture, design, development, testing and quality assurance, deployment, evaluation, commercialization and procurement), which is crucial in ensuring that AI is truly compliant with ethics principles (Zhou *et al.* 2020), but there was little coverage about how the implementation of ethics principles in AI could be evaluated with no clear measure of success stated (Goirand *et al.* 2021).

Goirand *et al.* (2021) highlight that the first four principles are taken from Beauchamp and Childress (2013), which is probably the most popular bioethics

framework used in healthcare, but not the only one. Importantly, there might be conflict between identified principles, such as when ensuring privacy and autonomy might negatively affect the potential beneficence of an application, and there might be conflict between how carers, clinicians and patients, as well as other stakeholders, envision the application of these principles (European Commission: High-Level Expert Group on Artificial Intelligence 2019, Jobin *et al.* 2019, Goirand *et al.* 2021).

Finally, it is important to keep in mind from the outset that several of these principles are interlinked, and sometimes the presented issues do not neatly fit into one theme. For example, data itself is not an ethics principle, but several of the principles will be relevant to the issues of responsible data stewardship. Hence part of the complexity of applying ethics principles to AI involves the overlap and divergence between the principles, as well as disagreements and subjectivity relating to the content and interpretation of each principle (Zhou *et al.* 2020).

Autonomy & freedom

This theme relates to the choices that patients and clinicians have over the actions that they enact or which affect them, and so relates to such concepts as self-determination, consent and empowerment (Jobin *et al.* 2019). Autonomy is facilitated by patient consent and clinician freedom, which might be negatively affected when humans are removed from the decision loop or by the black-box nature of some AI models (Olczak *et al.* 2021, Tranter-Entwistle *et al.* 2021, Stogiannos *et al.* 2023).

Fears regarding privacy and data security, such as that consent for data use in clinical AI will have the consequence of the data being sent to a third party for commercial gain, might constrain autonomy (Thieme *et al.* 2020, Reddy *et al.* 2021, Tranter-Entwistle *et al.* 2021). To facilitate autonomy, the consent process needs to be dynamic, ensuring that data subjects have a right to access their data and decide how it can be accessed by third parties (IEEE 2019, Stogiannos *et al.* 2023).

AI might possess a risk to clinician autonomy if practitioners rationalize their decisions to match AI predictions rather than act in ways that they believe to be in the best interest of their patients (Thieme *et al.* 2020). Such a situation risks turning the AI into the master and the clinician into the servant, yet simultaneously AI has the power to correct clinician's misconceptions, making it difficult to delineate the role AI should play in healthcare. Consequently, it is important to mitigate risks associated with the socio-technical systems associated with AI (Zhou *et al.* 2020), to prevent detrimental effects of AI systems on autonomy.

Non-maleficence

Non-maleficence was mentioned in over 80% of healthcare AI adoption frameworks making it the most frequently mentioned principle (Goirand *et al.* 2021). It relates to the Hippocratic principle of 'primum non nocere' and to the concepts of security, safety and prevention (Jobin *et al.* 2019), which all those involved in the development and implementation of an AI must bear in mind to mitigate against foreseeable AI facilitated harms (Reddy *et al.* 2021, MacKay *et al.* 2023).

A key source of harm is misdiagnosis. Problems relating to AI development and implementation can result in under – or over – diagnosis which can occur due to issues of fairness and quality of the underlying data (e.g. because the data was not annotated to provide the appropriate context for the diagnosis), staff were not trained how to use the AI and adversarial attacks (Thieme *et al.* 2020, Harvey and Gowda 2021, Olczak *et al.* 2021, Tranter-Entwistle *et al.* 2021, Saw and Ng 2022, MacKay *et al.* 2023, Stogiannos *et al.* 2023). Harms can also originate from distractions which AI algorithms might generate through prompts, which can contribute to 'alarm fatigue' or from inappropriately shared data (Hickman *et al.* 2021). These harms may be of physical, psychological or economic nature and affect individuals as well as healthcare systems through, e.g. insurance eligibility or costs (Harvey and Gowda 2021, Hickman *et al.* 2021, Reddy *et al.* 2021).

Importantly, some of these issues cannot be easily picked up during the statistical evaluation of AI models (Olczak *et al.* 2021), though before implementation it is imperative that AI results are reproducible if not replicable (MacKay *et al.* 2023). Noteworthy, some issues such as potential increases in workload due to AI facilitated changes to working routines, are not even related to the model's performance (Saw and Ng 2022). Consequently, it is important to be aware of the limitations of an AI models and not to over-trust AI predictions, especially in grave matters such as deprivation of liberty in psychiatric patients (Thieme *et al.* 2020).

Due to calibration drift, AI models are at a risk of declining function which can result in patient harm, but safeguards need to be employed to ensure that if the AI is learning in a continuous and dynamic fashion, that quality and safety are also safeguarded (Hickman *et al.* 2021, Mazaheri *et al.* 2021, MacKay *et al.* 2023, Pruski 2023, Stogiannos *et al.* 2023). Nevertheless, this might be problematic in regions where clinical AI only gains regulatory approval in a static form, like in the U.S.A. (Nair *et al.* 2022). There is also an issue with validating and assessing such dynamic AI technologies (Tranter-Entwistle *et al.* 2021). Fundamentally, continuous evaluation of AI should be the motor for use reassessment, re-design, and future development of these technologies (European Commission: High-Level Expert Group on Artificial Intelligence 2019, The Institute for Ethical Ai & Machine Learning n.d.). Importantly, employing human oversight in monitoring AI performance will require staff input and increase costs, but having a human in the loop might also benefit patient autonomy and provide a way to minimize the harm from mistaken AI predictions (The Institute for Ethical Ai & Machine Learning n.d., Hickman *et al.* 2021, Hine *et al.* 2022).

An area that is particularly at risk of AI generated harm is that of rare events, where it might be more difficult to obtain high-quality data, which is key to developing safe and effective AI models (Reddy *et al.* 2021). In such situations there might be conflicts between current guidance and the AI's recommendations (Monteith *et al.* 2023). Since clinicians are generally trained to interpret raw data, it might be difficult for them to interpret the reasonability of AI recommendations in such situations, especially if the AI model is of a black-box nature (Monteith *et al.* 2023). In such situations it is important that clinicians do not fall into automation bias or complacency, but that they also do not disregard potentially

beneficial AI advice, since the AI might see features in the data which humans cannot (Nair *et al.* 2022, Monteith *et al.* 2023).

The risks of such harms can be minimized by stringent technology development process and regulations, which focus on identifying the potential harms and their sources as well as taking action to decrease the potential extent and likelihood of harm, for example by stating staff training requirements (IEEE 2019, MacKay *et al.* 2023). Yet, such process should not stifle innovation in clinical AI, which might lead to patient benefits (Sutton and Rushlow 2011, Hine *et al.* 2022). Moreover, it is important to have multidisciplinary teams involved in ensuring that data is appropriately annotated and can be used to answer the relevant clinical questions the AI is trying to address, as well as that the data is representative of the local population (Wiens *et al.* 2019, Tranter-Entwistle *et al.* 2021, Saw and Ng 2022). Such processes help ensure that ethics are built into AI models (Zhou *et al.* 2020). Furthermore, clear guidance on the process of AI implementation is needed to avoid harms during the implementation stage, such as misdiagnosis, and the process of implementation (and monitoring) should be driven by healthcare providers rather than technology vendors (Thieme *et al.* 2020, Tranter-Entwistle *et al.* 2021). Finally, it is important to consider if patients should be informed about these potential harms and give consent to these potential new sources of harm (Olczak *et al.* 2021, Pruski 2024a).

Beneficence

Beneficence means that the AI should be effective in delivering maximal benefits to those who it serves, both in clinical and administrative terms, and includes considerations of the wider societal or common good (Jobin *et al.* 2019, Thieme *et al.* 2020, Reddy *et al.* 2021). In the context of AI this is a complex concept, as an AI can be seen as serving the patient to improve their health, the clinician to make their job easier and the organization to improve its efficiency. Consequently (at least partially) beneficence, can relate to improving performance metrics. Isbanner *et al.* (2022) found that the accuracy of an AI was the aspect of its functioning that was voted most important, out of seven principles, by their respondents, while AI speed and cost reduction were the least highly rated attributes.

The lack of datasets gathered to address well-defined clinical questions, use of unsuitable data (e.g. collected for billing purposes), as well as the development of technically well performing models that address questions of little clinical relevance have hampered the beneficence of clinical AI (Wiens *et al.* 2019). For example, a model predicting patient mortality from the fact that a patient is receiving palliative care might be very accurate but is not very useful (Hine *et al.* 2022, MacKay *et al.* 2023). To deliver good AI facilitated outcomes, the robustness of the data used to develop the AI model is key, as well as the quality of the data pertaining to our patient which we will feed into the model needs to be at least as good as information collected during a 'normal' clinical encounter (Fleming 2021). Having clear standards about how to harmonize data collection globally, could ease the building of data models using large

datasets representative of the global population, and help with the global implementation of these models (Wiens *et al.* 2019), ensuring generalisability and minimizing biases.

The beneficence of AI technologies can be further increased by tailoring them to the needs of particular patient populations and clinicians. If AI models are developed to address pertinent questions, AI can be used together with precision medicine, with the hope of moving towards truly personalized medicine. In this respect, it is clear that metrics are needed to monitor the accuracy of AI predictions and the cost impact of the implemented technologies (The Institute for Ethical Ai & Machine Learning n.d.).

There is a risk with using AI to direct staff to make decisions which will improve unit performance metrics or maximize profit but will not positively affect patient experience (Char *et al.* 2018, Zhou *et al.* 2020). Moreover, best practice and true diagnosis are not always the same, especially in emerging fields. AI can prevent patient experience improvement by perpetuating the application of suboptimal guidelines, which then become even more established as the standard of care and so the AI might prevent improvement by promoting stagnation (Char *et al.* 2018). This does not mean that AI should not be evaluated against the gold standards which are acknowledged at the time of the AI's development (Kernbach *et al.* 2022), but that system goals need to be set according to clinical priorities and that the impact of such AI technologies is evaluated (Reddy *et al.* 2021, Hine *et al.* 2022).

Fairness & justice

This topic consisted of such themes as non-discrimination, equality, equity and remedy of harms (Jobin *et al.* 2019, Thieme *et al.* 2020, Pruski 2024b). Surprisingly, fairness was the third least important principle in the study carried out by Isbanner *et al.* (2022).

Algorithms have the potential to perpetuate human biases, amplify them and create new ones, which is known as the phenomenon of 'algorithmic bias', which when caused by historical data is known as 'innate latent bias' (Char *et al.* 2018, Wiens *et al.* 2019, Harvey and Gowda 2021, Hickman *et al.* 2021, Mazaheri *et al.* 2021, Filipow *et al.* 2022, Saw and Ng 2022, MacKay *et al.* 2023). These biases might relate to ethnicity or disability status, due to the history of rationing of healthcare resources and inconsistencies of access to healthcare for members of underrepresented groups (e.g. due to financial or communication barriers). These issues have historically affected the health outcomes of various groups and hence it will affect the algorithms predicting their health outcomes if an AI is trained on data originating from centres where such inequalities occurred (Char *et al.* 2018, Wiens *et al.* 2019, Hickman *et al.* 2021, Filipow *et al.* 2022). It can also occur from the lack of data for patients from specific groups or the poor quality of data (e.g. due to poor data handling or coding practices), as well as decisions made during the design and development of the AI or post-deployment monitoring (Olczak *et al.* 2021, Tranter-Entwistle *et al.* 2021, Filipow *et al.* 2022, MacKay *et al.* 2023, Monteith *et al.* 2023, Stogiannos *et al.* 2023).

Ensuring diversity and inclusion should be considered key to the development of effective AI models. Datasets need to be assessed for their coverage, consistency and the reproducibility of the data collection methodology, to ensure that the datasets are valid for the intended application (Reddy *et al.* 2021), and techniques should be established to monitor the likelihood of AI bias during the model's development stage (The Institute for Ethical AI & Machine Learning n.d.).

There are often no clear answers as to how to ensure algorithmic fairness. For example, the inclusion of protected characteristics in AI models can perpetuate biases in some cases (Tranter-Entwistle *et al.* 2021, Filipow *et al.* 2022), while in other circumstances it might be appropriate; adversarial debiasing has been proposed as a method to deal with biases originating from the inclusion of such characteristics (Kernbach *et al.* 2022). Yet, there are other complexities which might affect the fairness and justice of AI applications, affecting their usability across a general population, such as the geographical level of the data on which an AI was developed (Nair *et al.* 2022, MacKay *et al.* 2023, Pruski 2024b). This might affect the AI's efficacy across all possible social groups and the potential beneficence towards the local population. Hence, it is important to contextualize the target population of the AI to ensure that the potential trade-off between generalisability of the model and benefit to specific groups is appropriately balanced (Reddy *et al.* 2021, Pruski 2024b).

Fleming (2021) and Thieme *et al.* (2020) also noted that there might be inequalities in access to digital technologies by different social groups (the so called 'digital divide'), which can affect the benefit these groups reap from the use of AI. Moreover, when AI technologies rely on text mining data, e.g. from patients' social media accounts, such technologies might have limited benefit for multilingual speakers if the AI is only trained on the dominant language. Hence it is key that the data upon which such AI applications are built represent the diversity of patients and clinical environments in which such applications will be used and that process are present which cater to patients who do not use digital technologies, so as to minimize such AI facilitated harms (Harvey and Gowda 2021, Hine *et al.* 2022, Stogiannos *et al.* 2023).

AI technologies have the scope to deepen inequities if they are not designed in a way that allows disadvantaged centres to implement these technologies (Hine *et al.* 2022). The price of AI solutions and the inability to recruit the staff necessary to use and maintain AI applications might disadvantage poorer or remote centres (Harvey and Gowda 2021, Nair *et al.* 2022). Yet, benefits that have been gathered from, e.g. giving private developers access to NHS data, must be distributed across the whole patient population and not contribute to increasing regional inequalities (Hickman *et al.* 2021).

Fairness demands that the benefits, risks and cost of AI development and implementation are proportionally distributed across the population (Thieme *et al.* 2020, Stogiannos *et al.* 2023, Pruski 2024b). Such a distribution of AI benefits and cost should help to maintain social cohesion and uphold the dignity of potentially vulnerable groups (Jobin *et al.* 2019).

Dignity

Dignity is a complex concept (Killmister 2010), which relates to the fundamental worth of people and the respect that they are due, and is often likened to autonomy and the rights of individuals (Macklin 2003, European Commission: High-Level Expert Group on Artificial Intelligence 2019). To maintain dignity, it is important to negotiate the role of AI in care processes across all relevant actors (e.g. patients, healthcare staff, institutions, families), so as to attend to any feelings of worry, provide reassurance and promote patient choice (Hine *et al.* 2022).

It is unclear whether using publicly available patient data (e.g. from social media) without patient consent might affect patient dignity (Thieme *et al.* 2020). Similarly, just because we can make accurate predictions using AI does not mean that we should. For example, if we could impute a patient's smoking or HIV status (even with high accuracy), it does not mean that we should, especially if the patient refused to provide this information in the first instance (Wiens *et al.* 2019). As such, to maintain dignity, it is key to ensure that the AI design and implementation processes embody human rights principles.

Dignity can also relate to the integrity of healthcare staff, which might be particularly affected by societal, management or governmental pressures to defer decision-making to AI (while potentially still bearing the liability burden for AI mistakes), even when this goes against clinical judgement. A lawsuit in the U.S.A. alleges that this happened with elderly care recipients (Olczak *et al.* 2021, Saw and Ng 2022, Napolitano 2023). Therefore, it is important to allow healthcare staff to contest the decisions made by a clinical AI (Goirand *et al.* 2021).

Responsibility

Responsibility was one of the three most important traits of an AI in Isbanner *et al.*'s (2022) survey and is a concept that relates to liability, accountability for the actions of the AI and integrity (Jobin *et al.* 2019). Clinical AI provide a challenge to this principle since traditionally it was the clinician who made the recommendations from which any harm occurred. Issues can now arise because of problems with AI recommendations, e.g. due to training data issues, and clinicians might not be able to scrutinize these recommendations due to the black-box nature of many AI technologies (Olczak *et al.* 2021, Nair *et al.* 2022, Monteith *et al.* 2023). Yet, it must be remember that clinicians are also often black-box decision-makers themselves, acting on their intuitions or accepting advice from colleagues whose reasoning they might not fully understand (Olczak *et al.* 2021, Pruski 2024a).

Generally, there are three key players with respect to whom the potential responsibility for clinical AI failure might be attributed: clinicians, healthcare organizations, and AI developers or technology companies (Sutton and Rushlow 2011, Harvey and Gowda 2021). Liability could occur potentially from clinicians deferring to or overriding AI recommendations (MacKay *et al.* 2023). Legislators must develop tools which will allow AI application developers to be held accountable for how their apps functions (Ben Ali *et al.* 2021, Stogiannos *et al.* 2023). Indeed, the current trend is moving away from the learned intermediary standard where clinicians assumed any responsibilities for issues caused by the devices they used during

patient care, while other parties were absolved from any responsibility (Saw and Ng 2022). This is not to say that clinicians should not be responsible for implementing inappropriate AI tools in their clinical practice (Stogiannos *et al.* 2023). Procurement procedures should guard against the undue influence of staff who have participated in the development of an AI applications or have financial links to it (Nair *et al.* 2022). Whistleblowing can be an important tool in holding those developing and implementing AI accountable for their actions (Jobin *et al.* 2019).

Ensuring responsible data use might be difficult when privacy is being protected, as e.g. it is harder to track anonymised data and ensure that it does not end in data sets which are used both for training and testing of AI algorithms (Hickman *et al.* 2021). Similarly, it might be difficult to assign responsibility when AI is used not as anticipated by its designers and develops beyond its initial intent. While scenarios where AI develops its own consciousness might largely belong to the realm of science fiction, instances of chat-bots being coerced into using inappropriate language do highlight that such malicious learning can occur (Kernbach *et al.* 2022, McKendrick and Thurai 2022).

There is need for clear AI regulation, similar to medical device regulation, which would protect patients and staff from harm which can originate from the use of clinical AI, though tort law will also need to develop to handle any legal cases relating to AI facilitated harm (Kernbach *et al.* 2022, Stogiannos *et al.* 2023). Part of the complexity here originates from the fact that while sometimes software can directly harm patients (e.g. in AI controlled syringe drivers) in many cases it only contributes towards harmful situations (Kernbach *et al.* 2022).

Trust

Traditionally, the fiduciary relationship has been between patients and clinicians, yet with the implementation of AI technologies this might shift to a relationship between the patient and healthcare system (Char *et al.* 2018), since decision-making will be diffused between a multidisciplinary team and the technologies which the healthcare system implemented to support these teams. How much patients will trust such systems might depend on the accountability frameworks governing these processes. Moreover, with the ever increasing complexity of modern care the level of trust placed onto clinicians and AI might change, with AI potentially being regarded as the authority in best practice and the relegation of clinicians to a secondary role (Char *et al.* 2018). How much trust we are willing to place in such technologies might dictate how much we will be willing to move from AI acting in a supportive role to one of conditional autonomy (Sutton and Rushlow 2011).

Employing comprehensive and transparent laws, standards for regulatory AI technology approval, and guidelines for testing and on-site evaluation might foster robustness and prevent unintentional harms (European Commission: High-Level Expert Group on Artificial Intelligence 2019, Hine *et al.* 2022). Having clear guidance on clinical AI development oversight, and having patients involved in this process, can help in obtaining a social license for the use of healthcare AI (Hickman *et al.* 2021, MacKay *et al.* 2023). Public involvement can also help increase confidence that NHS data is used appropriately, and that if it is given to

private institutions, patients and the NHS will obtain commensurate benefits in exchange (Hickman *et al.* 2021, MacKay *et al.* 2023). The use of such public engagement can not only help to develop trust, but also help to ensure transparency and that the AI delivers the benefits expected by the public (MacKay *et al.* 2023). As such, it is key to consult stakeholders, especially those with whom asymmetries of power or information might be particularly evident, regarding both data use, as well as technology development and implementation, and to educate the stakeholders to present them with realistic expectations about the systems' abilities and limitations, so that trust can be built (European Commission: High-Level Expert Group on Artificial Intelligence 2019). Such a process might mitigate future problems associated with the use of publicly available social media data with AI technologies, for example, if stakeholders feedback that they are not always genuine in their social media depictions (Thieme *et al.* 2020).

When the AI is still in a developmental or research phase it is key that patients are informed about its use and appropriate consent is obtained (Kernbach *et al.* 2022). Importantly, stakeholders might be less averse to the introduction of AI where clear fall-back plans are in place in case AI malfunction and if rights of redress are established in cases when AI facilitated harms occur (European Commission: High-Level Expert Group on Artificial Intelligence 2019), and when technologies were developed by the governmental and academic sectors, rather than private sector companies (Saw and Ng 2022). Trust can be further strengthened when there is transparency as to the use of the technology and the evidence for the AI's effectiveness (IEEE 2019, Goirand *et al.* 2021). This means that that patients should be informed when AI is being used in their care, so that they can acknowledge the risks and benefits of AI involvement (Hickman *et al.* 2021) and potentially request human input into this process. Ongoing consent processes to both data and AI utilization, might help to ensure public trust in these technologies (Hine *et al.* 2022). Although, it is unclear to what extent it is necessary to inform patients about this and how practical it will be when clinical AI becomes ubiquitous (Pruski 2024a). Paradoxically, an otherwise laborious ongoing consent process could be supported by digital technologies themselves.

Privacy

Privacy, how personal or private information is treated, is core to clinical practice as it is a manifestation of the Hippocratic Oath's promise of confidentiality (Jobin *et al.* 2019, Ben Ali *et al.* 2021), but touches also on such concepts like information security, data ownership and secondary use of data (Harvey and Gowda 2021). Out of all the implementation frameworks reviewed by Goirand *et al.* (2021), over 60% mentioned it. This is not surprising considering that data breaches are not uncommon in healthcare (Uwizeyemungu *et al.* 2019, Ben Ali *et al.* 2021). Such cybersecurity breaches do not only affect privacy but can have a negative effect on patient treatment if data for decision-making is not at hand, highlighting the criticality of data privacy legislation (Saw and Ng 2022, Stogiannos *et al.* 2023) and cybersecurity in protecting patient and institutions from harms originating from cyber-attacks (Nair *et al.* 2022, The Institute for Ethical Ai & Machine Learning n.d.).

Privacy is one of the principles that helps to foster trust and dignity. As such, data privacy might need to be discussed with patients when soliciting their consent for the delivery of digitally enhanced healthcare and ensured throughout the whole product lifecycle (Fleming 2021, Reddy *et al.* 2021, The Institute for Ethical Ai & Machine Learning n.d.). Consequently, considerations of privacy need to be embedded in AI design and legislation, to ensure that AI is non-maleficent (Hine *et al.* 2022).

It is important that healthcare institutions support independent quality research by health informatics staff to ensure data protection and it's appropriate use to facilitate the development of trustworthy AI applications (Ben Ali *et al.* 2021). Moreover, there is a need for a clear regulation that will specify the responsibility for the data used by AI applications and ensure that it is secure (Olczak *et al.* 2021). Clinicians and healthcare organizations are responsible for the data they utilize during patient care (Fleming 2021), and in the UK the GDPR (legislation.gov.uk n.d.) designates clear responsibilities for data controllers and processors as well as provides a framework for responsible sharing of data (Nair *et al.* 2022), but such regulation is still not internationally ubiquitous.

There can be a clear conflict between owners of AI technologies and users in terms of profit (Char *et al.* 2018). For example, AI applications nested within a wider digital ecosystem might execute hidden tasks to increase the profit of some stakeholders, such as the designers, by intercepting information from other applications within such an ecosystem (Fleming 2021). While data can and should be shared for legitimate purposes, data must not be sold to insurance companies, for example, and there might be little accountability for companies which engage in such practices (Ben Ali *et al.* 2021, Saw and Ng 2022). Importantly, the de-identification of patient data might not be the ultimate defence against such practices (Ben Ali *et al.* 2021). Pseudonymisation still links all the data to a specific patient, and when more modalities of data are available it becomes easier to triangulate a specific individual from anonymised datasets (Hickman *et al.* 2021).

In the wake of learning healthcare systems true confidentiality might require not noting some information in the medical record, which might prevent patients from receiving appropriate care and from reaping the benefits of any implemented AI technologies (Char *et al.* 2018). For example, patients might not wish some embarrassing symptoms to be recorded in their electronic health record, but this means that the presence of these symptoms cannot be input into an AI model. Similarly, it is unclear if we sometimes do not collect too much information about patients and whether we should only collect and keep the information that is strictly necessary (Fleming 2021) or whether such an approach would entail that we are potentially discarding data which will be essential for future patient care. Importantly, collecting information, especially from a patient's social (as opposed to medical) history or the patient's social media (even with their permission) risks inadvertently collecting data that pertains to other people, such as the patient's family and friends, creating further privacy challenges (Thieme *et al.* 2020, Fleming 2021).

The notion of data ownership is both important and complex in an era when secondary use of public data might involve selling this data for use in the private industry with the aim of developing profit generating technologies (Sutton and Rushlow

2011). Here privacy interlinks with the concepts of beneficence, fairness, autonomy and trust, as exemplified in previous remarks on NHS utilization, and distributes any gains from data sharing, and patient consent to this sharing. This raises questions as to whether patients give consent to the use of their data for specific purposes or do they give broad consent, and what are the ownership implications for the patients themselves from such commercialized used of their data (Sutton and Rushlow 2011).

Transparency

Transparency is the final principle that Goirand *et al.* (2021) classified as one of the most frequently (over 60%) mentioned among the frameworks they identified. It relates closely to the principle of trust, since transparency can generate trust (Stogiannos *et al.* 2023). It covers such aspects of AI as its explainability (or interpretability), openness about AI data flows and development, and communicating this to patients and staff (IEEE 2019, Jobin *et al.* 2019, Mazaheri *et al.* 2021, Olczak *et al.* 2021, Reddy *et al.* 2021, MacKay *et al.* 2023, Stogiannos *et al.* 2023). Explainability was, nevertheless, only the fourth most important principle out of the seven explored by Isbanner *et al.* (2022), perhaps because people might care more that things work reliably rather than be concerned with how they work (London 2019). Yet, knowing how a model makes its decisions might help with the progress of science (Pruski 2024a).

The presence of property clauses can perpetuate the black-box nature of AI algorithms by legally safeguarding information which could help realize transparency (Hickman *et al.* 2021). While there have been calls for the development of tools and processes that would facilitate AI explainability (The Institute for Ethical Ai & Machine Learning n.d.), the use of 'model fact labels' (similar to those used for medication) has been proposed as a solution that balances the rights of developers and needs of users (Olczak *et al.* 2021).

Transparency makes it easier for AI model evaluation (Kernbach *et al.* 2022). This can be greatly enhanced when there is traceability as to the AI's development process and its clinical use, facilitating AI audit and dialogue with stakeholders (European Commission: High-Level Expert Group on Artificial Intelligence 2019, Jobin *et al.* 2019, Stogiannos *et al.* 2023). Yet, interpretability can make AI technologies more susceptible to external manipulation (Stogiannos *et al.* 2023).

By means such as making end-users aware of the limitations of the AI's training data, transparency might help clinicians to be vigilant against unreasonable decisions made by AI and not neglect their own intuitions, preventing automation bias (Mazaheri *et al.* 2021, Nair *et al.* 2022). By being open-source and using interoperability standards (MacKay *et al.* 2023, Monteith *et al.* 2023) applications can be scrutinized by the expert community, and prevent issues originating from vendor lock-in.

Importantly, some authors have noted that it is a myth that more accurate AI technologies are more complex and hence less interpretable, as complex models might be more prone to overfitting (Kernbach *et al.* 2022). Hence, transparency is not necessarily a barrier to the development of clinically beneficial AI models.

Sustainability

This principle relates to the potential environmental costs of AI development and implementation. Maximizing sustainability might be beneficial for people, though it might also prevent the maximization of other ethical principles (European Commission: High-Level Expert Group on Artificial Intelligence 2019, Jobin *et al.* 2019, Zhou *et al.* 2020). Consequently, it is important to assess the impact of developing or implementing an AI on natural resources, and that environmental costs are appropriately distributed amongst the communities that benefit from the AI (Pruski 2024b).

Solidarity

Solidarity largely reflects the impact of AI on the labour market and the workforce (Jobin *et al.* 2019), but has wider implications regarding citizen rights, social cohesion and security (European Commission: High-Level Expert Group on Artificial Intelligence 2019, IEEE 2019, Jobin *et al.* 2019, Zhou *et al.* 2020). For example, algorithms might become the repository of collective knowledge, which can result in the de-skilling of clinicians and poorer developmental opportunities for trainee healthcare staff, even leading to job loss (Char *et al.* 2018, Jobin *et al.* 2019, Zhou *et al.* 2020, Hickman *et al.* 2021, Nair *et al.* 2022, Monteith *et al.* 2023). Even the introduction of AI to tackle automatable tasks will require adaptation of clinical working practices (Nair *et al.* 2022). This might decrease the market worth of healthcare staff, and deny patients a back-up of a well skilled clinician in cases of clinical AI failure (Hickman *et al.* 2021). As such, there needs to be a clear commitment to buffer the effects of AI on worker displacement (The Institute for Ethical Ai & Machine Learning n.d.).

It is important to remember that AI models are not always the correct tools for dealing with clinical problems and hence maintaining a skilled workforce is important (Reddy *et al.* 2021). Human contact is still highly sought after among patients in their clinical encounters (Isbanner *et al.* 2022) and is perhaps a phenomenon that highlights the need for human interaction in a technology focused age. Hence, AI's role is being advocated as an adjunct to human skills, which can take over repetitive tasks, and not a replacement for empathy, compassion and care (MacKay *et al.* 2023).

Ensuring that AI is used for the common good and upholding people's rights, might require debate between competing visions of humane societies, e.g. with respect to individualistic and communitarian oriented data sharing frameworks (Future of Life Institute 2017, Jobin *et al.* 2019, Kernbach *et al.* 2022).

Conclusions

The identified literature has shown a range of ethical issues associated with the adoption of AI into healthcare, and since there was a saturation of healthcare related themes as data extraction proceeded, this review is likely to present a comprehensive picture of these issues.

While the emerging issues have been categorized into the 11 themes defined by Jobin *et al.* (2019), many of the identified issues do not fit neatly into any one category. This highlights that there is a clear interplay between various ethics principles and technical challenges. For example, the cases used in the section on dignity also

relate to privacy, and issues of privacy raise issues of responsibility when there is a failure in ensuring privacy. The section on transparency also demonstrated that it can foster responsibility by identifying where problems might occur. These illustrate overlap between principles, but the principle with the most overlap is likely to be trust, as the enactment of all the other principles might foster trust. Similarly, the theme of data reappeared under the headings of several principles, as e.g. exemplified by the challenges relating to the use of patients' social media data. Consequently, these are not standalone principles, but rather convenient descriptors for complex and nuanced phenomena.

While there might not be any inherent order of importance of the 11 aforementioned principles, surveying potential healthcare AI users to understand what is important to them can help those developing and implementing healthcare AI to develop technologies with high user acceptability, and aid policy makers. For example, the findings from this review are being currently used in Wales as part of a study to understand the importance of these principles and other factors to members of the public and healthcare professionals with respect to healthcare AI.

Acknowledgments

I thank Kathleen Withers for comments on the manuscript draft, Meg Kiseleva and Simone Willis for their advice on developing and executing the original literature search strategy, as well as Robert Palmer, Andrew Brass and Frances Hooley for advice on this project.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This review was carried out as part of MP's DClinSci thesis /HSST research project in Health Informatics for which he is funded through Health Education and Improvement Wales. It is part of a project 'Identifying the barriers and facilitators for implementing Machine learning to achieve Value-Based Healthcare in Wales', which is being carried out with the support from the Welsh Value in Health Centre.

Supplemental data

Supplemental data for this article can be accessed online at https://doi.org/10.1080/20502877.2025.2541438.

Notes on contributor

Dr. Michal Pruski is a Senior Clinical Scientist at the University Hospital of Wales. He works in both clinical patient-facing and research roles and is involved in

providing clinical ethics support. Michal is also a Higher Specialist Scientist Trainee and a DClinSci candidate in health informatics.

ORCID

Michal Pruski http://orcid.org/0000-0001-7582-1418

References

Beauchamp, T.L., and Childress, J.F., 2013. *Principles of biomedical ethics*. 7th ed. New York: OUP USA.

Ben Ali, W., *et al*., 2021. Implementing machine learning in interventional cardiology: the benefits are worth the trouble. *Frontiers in cardiovascular medicine*, 8, 711401.

Bevan Commission. 2022. Prudent healthcare principles. *Bevan Commission*. [online]. Available from: https://www.bevancommission.org/about/prudent-principles/ [Accessed 4 September 2022].

Char, D.S., Shah, N.H., and Magnus, D., 2018. Implementing machine learning in health care – addressing ethical challenges. *The New England journal of medicine*, 378 (11), 981–983.

European Commission: High-Level Expert Group on Artificial Intelligence. 2019. Ethics guidelines for trustworthy AI.

Filipow, N. *et al*., 2022. Implementation of prognostic machine learning algorithms in paediatric chronic respiratory conditions: a scoping review. *BMJ open respiratory research*, 9(1). [online]. Available from: https://urldefense.com/v3/__http://ovidsp.ovid.com/ovidweb.cgi?T = JS&PAGE = reference&D = med21&NEWS = N&AN = 35297371__;!!PDiH4ENfjr2_Jw!CLyru8EmRUPnMcbuTEReC8dBowl8rZcYxvA1OYJ9ZC4W1VSt8a8HR90mP1uGPg0PQyoQeB10VpeFXkRp9UTf_j9NvQDovV4bSPT2cGtCnuyZLA$ [ovidsp[.]ovid[.]com].

Fleming, M.N., 2021. Considerations for the ethical implementation of psychological assessment through social media via machine learning. *Ethics & behavior*, 31 (3), 181–192.

Future of Life Institute. 2017. Asilomar AI Principles. *Future of Life Institute*. [online]. Available from: https://futureoflife.org/2017/08/11/ai-principles/ [Accessed April 9, 2022].

Goirand, M., Austin, E., and Clay-Williams, R., 2021. Implementing ethics in healthcare AI-based applications: A scoping review. *Science and engineering ethics*, 27 (5), 61.

Harvey, H.B., and Gowda, V., 2021. Regulatory issues and challenges to artificial intelligence adoption. *Radiologic clinics of North America*, 59 (6), 1075–1083.

Hickman, S.E., Baxter, G.C., and Gilbert, F.J., 2021. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. *British journal of cancer*, 125 (1), 15–22.

Hine, C., Nilforooshan, R., and Barnaghi, P., 2022. Ethical considerations in design and implementation of home-based smart care for dementia. *Nursing ethics*, 29 (4), 1035–1046.

IEEE. 2019. Ethically aligned design; First Edition. [online]. Available from: http://link.springer.com/10.1007978-3-030-12524-0_2 [Accessed 8 December 2023].

The Institute for Ethical Ai & Machine Learning, n.d. The Institute for Ethical AI & Machine Learning. [online]. Available from: https://ethical.institute [Accessed 8 December 2023].

Isbanner, S., *et al*., 2022. The adoption of artificial intelligence in health care and social services in Australia: findings from a methodologically innovative national survey of values and attitudes (the AVA-AI study). *Journal of medical internet research*, 24 (8), e37611.

Jobin, A., Ienca, M., and Vayena, E., 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1 (9), 389–399.

Kernbach, J.M., *et al*., 2022. The artificial intelligence doctor: considerations for the clinical implementation of ethical AI. *Acta neurochirurgica. supplement*, 134, 257–261.

Killmister, S., 2010. Dignity: not such a useless concept. *Journal of medical ethics*, 36 (3), 160–164.

legislation.gov.uk, n.d. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement

of such data (United Kingdom General Data Protection Regulation). https://webarchive.nationalarchives.gov.uk/eu-exit/https://eur-lex.europa.eu/legal-content/EN/TXT/?uri = CELEX:02016R0679-20160504. [online]. Available from: https://www.legislation.gov.uk/eur/2016/679/contents [Accessed April 10, 2022].

London, A.J., 2019. Artificial intelligence and black-Box medical decisions: accuracy versus explainability. *Hastings center report*, 49 (1), 15–21.

MacKay, C., *et al*., 2023. A framework for implementing machine learning in healthcare based on the concepts of preconditions and postconditions. *Healthcare analytics*, 3, 100155.

Macklin, R., 2003. Dignity is a useless concept. *British medical journal*, 327 (7429), 1419–1420.

Mazaheri, S., *et al*., 2021. Challenges of implementing artificial intelligence in interventional radiology. *Seminars in interventional radiology*, 38 (5), 554–559.

McKendrick, J., and Thurai, A. 2022. AI isn't ready to make unsupervised decisions. *Harvard business review*. [online]. Available from: https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions [Accessed 22 January 2024].

Monteith, S., *et al*. 2023. Challenges and ethical considerations to successfully implement artificial intelligence in clinical medicine and neuroscience: a narrative review. *Pharmacopsychiatry*. [online]. Available from: https://urldefense.com/v3/__http://ovidsp.ovid.com/ovidweb.cgi?T = JS&PAGE = reference&D = medp&NEWS = N&AN = 37643732__;!!PDiH4ENfjr2_Jw!CLyru8EmRUPnMcbuTEReC8dBowl8rZcYxvA1OYJ9ZC4W1VSt8a8HR90mP1uGPg0PQyoQeB10VpeFXkRp9UTf_j9NvQDovV4bSPT2cGs9z2716w$ [ovidsp[.]ovid[.]com].

Nair, A.V., *et al*., 2022. Barriers to artificial intelligence implementation in radiology practice: what the radiologist needs to know. *Radiologia*, 64 (4), 324–332.

Napolitano, E. 2023. UnitedHealth uses faulty AI to deny elderly patients medically necessary coverage, lawsuit claims. *CBS News*. [online]. Available from: https://www.cbsnews.com/news/unitedhealth-lawsuit-ai-deny-claims-medicare-advantage-health-insurance-denials/ [Accessed January 22, 2024].

Olczak, J., *et al*., 2021. Presenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a clinical AI research (CAIR) checklist proposal. *Acta orthopaedica*, 92 (5), 513–525.

Pruski, M., 2023. Ethics framework for predictive clinical AI model updating. *Ethics and information technology*, 25 (3), 48.

——., 2024a. AI-enhanced healthcare: not a new paradigm for informed consent. *Journal of bioethical inquiry*, 21, 475–489.

——., 2024b. What does it mean for a clinical AI to be just: conflicts between local fairness and being fit-for-purpose? *Journal of medical ethics*. [online]. Available from: https://jme.bmj.com/content/early/2024/02/29/jme-2023-109675 [Accessed 29 February 2024].

Reddy, S., *et al*., 2021. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ health & care informatics*, 28 (1), e100444.

Saw, S.N., and Ng, K.H., 2022. Current challenges of implementing artificial intelligence in medical imaging. *Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian association of biomedical physics (AIFB)*, 100, 12–17.

Stogiannos, N., *et al*., 2023. Black box no more: a scoping review of AI governance frameworks to guide procurement and adoption of AI in medical imaging and radiotherapy in the UK. *The British journal of radiology*, 20221157.

Sutton, L.P., and Rushlow, W.J., 2011. Regulation of Akt and Wnt signaling by the group II metabotropic glutamate receptor antagonist LY341495 and agonist LY379268. *Journal of neurochemistry*, 117 (6), 973–983.

Thieme, A., Belgrave, D., and Doherty, G., 2020. Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM transactions on computer-human interaction*, 27 (5),). [online]. Available from: https://urldefense.com/v3/__https://www.scopus.com/inward/record.uri?eid = 2-s2.0-85090445793&doi = 10.1145*2f3398069&partnerID = 40&md5 = ab0bc8c892659a57e1539d9a6e0eb5a3__;JQ!!PDiH4ENfjr2_Jw!CLyru8EmRUPnMcbuTEReC8dBowl8rZcYxvA1OYJ9ZC4W1VSt8a8HR90mP1uGPg0PQyoQeB10VpeFXkRp9UTf_j9NvQDovV4bSPT2cGvN8JZjSQ$ [scopus[.]com].

Tranter-Entwistle, I., *et al*., 2021. The challenges of implementing artificial intelligence into surgical practice. *World journal of surgery*, 45 (2), 420–428.

Uwizeyemungu, S., Poba-Nzaou, P., and Cantinotti, M., 2019. European hospitals' transition toward fully electronic-based systems: Do information technology security and privacy practices follow? *JMIR medical informatics*, 7 (1), e11211.

Wiens, J., *et al*., 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25 (9), 1337–1340.

Zhou, J., *et al*. 2020. A survey on ethical principles of AI and implementations. In 2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020. pp. 3010–3017. [online]. Available from: https://urldefense.com/v3/__https://www.scopus.com/inward/record.uri?eid = 2-s2.0-85099711716&doi = 10.1109*2fSSCI47803.2020.9308437&partnerID = 40&md5 = 7dea9e1e9ba849eb24591fa6486df4bd__;JQ!!PDiH4ENfjr2_Jw!CLyru8EmRUPnMcbuTEReC8dBowl8rZcYxvA1OYJ9ZC4W1VSt8a8HR90mP1uGPg0PQyoQeB10VpeFXkRp9UTf_j9NvQDovV4bSPT2cGvvy40T58$ [scopus[.]com].