AI and Ethics
https://doi.org/10.1007/s43681-023-00353-x

ORIGINAL RESEARCH

# The human role to guarantee an ethical AI in healthcare: a five-facts approach

**Raquel Iniesta**[1,2]

Received: 5 August 2023 / Accepted: 26 September 2023


**Abstract**
With the emergence of AI systems to assist clinical decision-making, several ethical dilemmas are brought to the general attention. AI systems are claimed to be the solution for many high-skilled medical tasks where machines can potentially surpass human ability as for example in identifying normal and abnormal chest X-rays. However, there are also warns that AI tools could be the basis for a human replacement that can risk dehumanisation in medicine. In recent years, important proposals in the domain of AI ethics in healthcare have identified main ethical issues, as for example fairness, autonomy, transparency, and responsibility. The human warranty, which implies human evaluation of the AI procedures, has been described to lower the ethical risks. However, as relevant these works have been, translating principles into action has proved challenging as existing codes were mostly a description of principles. There is a great need to produce *how-to* proposals that are specific enough to be action-guiding. We present five human-focussed facts designed into a framework of human action for an ethical AI in healthcare. Through the factors, we examine the role of medical practitioners, patients, and developers in designing, implementing, and using AI in a responsible manner that preserves human dignity. The facts encompass a range of ethical concerns that were commonly found in relevant literature. Given that it is crucial to bring as many perspectives as possible to the field, this work contributes to translate principles into human action to guarantee an ethical AI in health.

**Keywords** Ethics · AI · Human role · Healthcare · Principles

## 1 Introduction

Clinical practitioners and machines have had a *master–servant* relationship for years: physicians understood the machine's functioning, decided what the machine would do and when [1]. The machine produced outputs that needed further human translation and interpretation. The emergence of AI-based medical tools to assist clinical decision-making is leading to a completely new paradigm which resembles a more *symbiotic* relationship, in which humans and computers become *teammates* aiming to solve a common goal [1]. Even without being operated by a human, AI algorithms can provide information to aid practitioners in comprehension of a patient medical situation and can offer predictive capabilities, as for example, how a patient will progress or might respond to a given particular treatment [1, 2].

Governments all over the world, particularly in the US and China, are making big investments to integrate AI systems for healthcare [3–6] trusting the potential of AI technology to enhance health outcomes and help making cost-efficient clinical decisions [7–9]. Despite the big efforts that particularly the private sector has made to develop cutting-edge AI technologies [10, 11], the incorporation of AI systems in healthcare has been slower than hoped [12–14]. Important ethical challenges like the transparency, suitability, and adaptability of the tools, and the need of mutual collaboration between human agents have been named to be key reasons for that implementation gap [14, 15]. These and other severe ethical concerns of integrating AI models in medicine have been widely discussed in the literature with many academic and non-academic publications in the field. A global convergence about the main ethical principles for AI was described by Jobin et al. [16]. In 2021, a scoping review by Murphy et al. on ethical

✉ Raquel Iniesta
raquel.iniesta@kcl.ac.uk

[1] Biostatistics and Health Informatics Department, Institute of Psychiatry, Psychology and Neurosciences, King's College London, London, UK

[2] King's Institute for Artificial Intelligence, King's College London, London, UK

Published online: 25 October 2023

issues of integrating AI in healthcare involving 103 records identified four common ethical concerns [17]. The Ethics & AI: A Systematic Review on Ethical Concerns and Related Strategies for Designing with AI in Healthcare [18] systematically reviewed 45 documents and found 12 ethical challenges (Table 1). The Ethics and governance of artificial intelligence for health guidance provided by the World Health Organisation (WHO) [19] also aimed to identify the ethical challenges and risks with the use of AI for health and published 6 ethical consensus (Table 1). The EU is elaborating a regulation based in the union's values, with the purpose to promote the uptake of human centric and trustworthy artificial intelligence [20]. However, principles alone are not enough to guarantee trustworthy or ethical AI in medicine [15, 21]. There is an open debate on who is responsible and liable for an ethical AI in medicine and how principles should be translated into practice [22]. Existing codes contain abstract and vague concepts, as for example commitments to ensure that AI is 'fair', or respects 'human dignity', or enables 'human flourishing' which are not specific enough to be action-guiding' [21, 23].

A guidance developed by the WHO and other relevant works [15, 19, 24] reached consensus in considering that the ethical principles for AI are important for all stakeholders who seek guidance in the responsible development, deployment, use, and evaluation of AI technologies for health. From a broad perspective, this includes clinicians and primary care medical professionals, systems developers, health system administrators, policy-makers in health authorities, researchers, and local and national governments. Some works argued that a narrower focus should be put on elaborating strategies for clinicians, developers, and patients to effectively translate AI ethical principles into practice [15, 22, 25]. For example, accountability can be assured by application of "human warranty", which implies evaluation by patients and clinicians in the development and deployment of AI technologies [19].

Collaboration between medical doctors and AI designers was emphasised as critical to align algorithms with medical expertise, bioethics, and medical ethics [15]. Important ethical concerns like dehumanisation [26, 27], a consequence of deindividuating practices, or empathy reduction [13, 28] and disempowerment of both patients and clinicians could be alleviated by clinical decisions being shared between medical practitioners and patients [15]. Collaboration and shared decision-making between clinicians and patients are the basis for the Patient-Centered care (PCC) delivery model, highlighted by the WHO as a key dimension of personalised and comprehensive care [19, 29, 30]. The collaboration between stakeholders to reach a shared clinical decision is also considered as the key pillar of the Evidence-Based Medicine (EBM), a practice of medicine that integrates science, clinical experience, and the individual patient's unique circumstances [31–34]. Clinicians are increasingly required to base clinical decisions on the best available evidence [33].

Based on the idea of mutual collaboration and shared decision-making between physicians, patients, and designers, the present research characterises five facts that aim to contribute to translate ethical principles into human action—for clinicians, developers, and patients—that can ensure an ethical development, integration and deployment of AI systems in healthcare. The theoretical basis for the five-facts design lays on the integration of (1) the *collaborative model* [15], (2) the Patient-Centered practice [29], and (3) the Evidence-Based Medicine approach [32–34].

**Table 1** Ethical principles and issues included in (1) *Principles of Biomedical Ethics* by Beauchamp and Childress (1979) [52], (2) *The Ethics & AI: A Systematic Review on Ethical Concerns and Related Strategies for Designing with AI in Healthcare* by Li et al. (2023) [17], and (3) *The Ethics and governance of artificial intelligence for health guidance* provided by the WHO [18]. Ethical principles and ethical issues were matched when possible. Those principles that matched across sources were highlighted in bold

| (1) *Four ethical pillars* | (2) *Li* et al. *2023* | (3) *WHO guidance* |
|---|---|---|
| **Justice** | **Justice** and **fairness** | Ensure inclusiveness and **equity** |
| Respect for **autonomy** | Freedom and **autonomy** | Protect **autonomy** |
| | Privacy | |
| | **Transparency** | Ensure **transparency**, explainability, and intelligibility |
| **Non-maleficence** | **Patient safety** and cybersecurity | Promote human well-being, human **safety,** and the public interest |
| | Trust | |
| **Beneficence** | **Beneficence** | |
| | **Responsibility** | Foster **responsibility** and accountability |
| | Solidarity | |
| | **Sustainability** | Promote artificial intelligence that is responsive and **sustainable** |
| | Dignity | |
| | Conflicts | |

## 2 Methodology

This work analyzes the role of three types of human agents in enabling an ethical AI in medicine: clinicians, patients, and developers. A definition of what we understand by each character designation, as well as the equivalency between each category and other terms we used along the work to refer to them can be found in Table 2.

### 2.1 The patient-centered and evidence-based medicine perspectives

Health care institutions increasingly pursue to deliver care that is both evidence-based and patient-centered. Patient-Centered Care (PCC) focuses on the individual's particular health care needs. The goal of PCC is to empower patients to become active participants in their care [35, 36]. Defining the PCC pathway concept has proven difficult given a lack of consensus [29, 35, 36]. In a study analysing both observation of the clinical encounter and patient perceptions, the patients' perception of the patient centredness of the interaction, and not the experts', was the stronger predictor not only of health outcomes but also of efficiency of health care represented by fewer diagnostic tests and fewer referrals [37]. For our work, we will then consider a definition of PCC based on patients' perceptions on patient centredness [38]. Patients expressed their will on a PCC which (a) explores the patients' main reason for the visit, concerns, and need for information; (b) seeks an integrated understanding of the patients' world—that is, their whole person, emotional needs, and life issues; (c) finds common ground on what the problem is and mutually agrees on management; (d) enhances prevention and health promotion; and (e) enhances the continuing relationship between the patient and the doctor [29].

Evidence-Based Medicine (EBM) is a practice of medicine that integrates the best available science with the healthcare professional's clinical experience and the individual patient's values, preferences, and unique circumstances to arrive at the best medical decision shared with the patient [31–34]. The EBM perspective states that the "the unique preferences, concerns and expectations each patient brings to a clinical encounter must be integrated into clinical decisions if they are to serve the patient" [32] as well as with the best available scientific evidence. As explained by Sacket et al., under an EBM approach, clinicians should acquire an increased expertise from individual expertise and external clinical evidence that will be reflected "in more effective and efficient diagnosis and in the more thoughtful identification and compassionate use of individual patients' predicaments, rights, and preferences in making clinical decisions about their care".

Moving towards PCC and EBM has been a major trend in health care over the past 20 years [39–41]. Preserving both approaches in the AI era is a major challenge. Since 2001, there have been several claims to improve the quality and performance of healthcare services by national and international institutions, such as the Institute of Medicine of America, the National Academies of Sciences, Engineering, Medicine, and the World Health Organization [30, 36, 40–42]. Particularly, PCC was raised as a crucial aspect to conform the criteria needed to improve quality care, together with safe, effective, efficient, timely and equitable care [30, 40]. The Topol review on the premises that should guide the future application of AI in healthcare emphasised that the patient must be considered to be at the center upon implementation of any new technology [43]. The two approaches, PCC and EBM, are very often complementary as improvements in one will enhance performance in others [39].

### 2.2 Questions that motivated the facts design

The Mia software [44] is an AI-based tool developed by Kheiron Medical Technologies to analyse standard mammograms for breast cancer screening. In a survey to 87 doctors [45], these were asked about how comfortable would they be about the Mia software being routinely used in clinical care. The respondents approved AI replacing one of the initial two humans that usually read the scans but objected to AI replacing all human readers. Clinicians mostly preferred to base their clinical decisions on national guidelines (77%),

**Table 2** Definition of the terms we used to designate the three types of human agents involved in this research

| Human agent | Other terms we used |
|---|---|
| **Clinician**: A qualified person who works in a hospital or private practice and that is entitled to make clinical decisions about patients' health | • Medical practitioner<br>• Doctor<br>• Physician |
| **Patient**: A person who is receiving medical treatment from a doctor or hospital. A patient is also someone who is registered with a particular doctor | |
| **Developer**: A person who is responsible for developing, coding, installing, and maintaining software systems. In broad terms, we called developers to all humans involved into the AI integration process, from design to deployment. Inevitably developers also extend to their institutions as they are directly involved at the different stages | • Designer |

studies using a nationally representative dataset (65%), and independent prospective studies (60%) as the essential evidence to follow. They also expressed important concerns as the need for additional external and independent validation of the AI tool. Their answers were raising methodological concerns as clinicians mentioned the need of involving representative datasets in the AI system building or additional validation of the tool, this pointing towards the developers' responsibility. There was also the impression that clinicians did not fully trust the system and/or the developers as clinicians denied replacing the two human readers and asked to run extra independent studies. From their views, we might infer that practitioners were seeing a risk of human replacement and maybe of commercial opportunism. Other studies have suggested that clinicians' concerns about the AI use include the accuracy of advice given, potential legal liability if harm to a patient occurs [24, 46] and that medical practitioners fear that AI 'may reduce their professional autonomy or may be used against them in the event of medical-legal controversies' [46–50]. Many important questions arise: (1) What should be the role of clinicians to enable an ethical AI when an AI system is recommending clinical decisions? and (2) What should be the role of those developing AI systems to ensure an ethical AI in healthcare? These two questions intend to trigger reflection around how the interaction between medical doctors and AI systems can frame an ethical AI in medicine. However, under PCC and EBM perspectives, any clinical decision is to be shared with the patient, so the patient should be an active part of the decision [29, 31, 33, 34]. Hence, another important question to reflect on is (3) What should be the patients' role in guaranteeing that an AI system deploys clinical decisions ethically?

### 2.3 The "patient-extended" collaborative model

To reflect on the role that patients together with medical practitioners and developers may have to guarantee an ethical AI in health, an extension of the collaborative model was considered [15]. The original collaborative model presented by Gundersen and Bærøe comprises two main claims [15]. First, it states that there must be collaboration between designers and doctors, as well as expertise in ethics, in both the design and use of medical AI. Second, AI designers, bioethicists, and medical doctors must have the capacity to communicate meaningfully about the way algorithms work, their limitations, and the algorithmic risks that arise in clinical decision-making. A public deliberation model was also presented by the authors, this including designers, doctors, policy-makers, and the general public. This model is called for when the technology is recognised as fundamentally transforming the conditions for ethical shared decision-making [15].

In the present work, we propose a "patient-extended" collaborative model, an extension of the collaborative model that lies between the collaborative model and the public model. The "patient-extended" collaborative model states that there must be collaboration between designers, doctors, and also patients to allow for an ethical AI in healthcare. This extended model differs from the public deliberation model in the sense that it lies in a sphere closer to the design step and the doctor's visit, and not at the level of public debate. The "patient-extended" collaborative model is conceived as a model that enlightens an individualised and personalised PCC and EBM experience, that will contribute towards preventing existential risks as dehumanisation in medicine and disempowerment of both clinicians and patients [28, 51].The strategy to include patients, as presented through the factors definition, is conceived twofold: (1) A patient is educated on how the technology works, on the related ethical concerns and their own rights as a patient. The patient is invited to collaborate with clinicians and designers at different stages of the development of the AI algorithm, so that their views can be incorporated in the design. (2) Medical doctors and patients collaborate to reach a shared decision, for which both agents are responsible. The outputs from the AI system are made available to the patient by the doctor in an intelligible manner. If a patient can understand how an automatically deployed decision was made, this would enable an empowerment of the patient and a real shared decision-making process where the person of the patient, as a whole, is included.

### 2.4 Consensus on the ethical challenges of AI in healthcare

We investigated the most common ethical challenges of AI for health. We assumed the existence of an overlapping consensus around certain principles for AI in healthcare and focussed on the existing proposals to look for meaningful convergence between them [23]. In particular, we focussed on (1) the four ethical pillars that have been classically in use in medicine [52], (2) a recent academic publication that aimed to cover the core AI ethical issues in medicine existing in literature: The Ethics & AI: A Systematic Review on Ethical Concerns and Related Strategies for Designing with AI in Healthcare by Li et al. [18], and (3) The Ethics and governance of artificial intelligence for health guidance provided by the WHO [19] (Table 1). The classical principles in (1) have been extremely relevant in the field of medical ethics and have strongly influenced ethical assessment in health care. The ethical dilemmas that encompass the emergence of AI in medicine are not an exemption and the pillars can naturally apply to them [53]. The European Commission has recently published guidelines for ethical and trustworthy AI echoing the prima facie principles of medical ethics [53–55]. The second work [18]

systematically reviewed 45 academic documents and ethical guidelines related to AI in healthcare and found 12 common ethical issues: justice and fairness, freedom and autonomy, privacy, transparency, patient safety and cyber security, trust, beneficence, responsibility, solidarity, sustainability, dignity, and conflicts. The guidance provided by the WHO [19] outlined six consensus principles to make sure that AI works to the public benefit of all countries: protect autonomy, promote human well-being, human safety and the public interest, ensure transparency, explainability and intelligibility, foster responsibility and accountability and ensure inclusiveness and equity and promote artificial intelligence that is responsive and sustainable. To summarise the common ethical issues and principles found in the literature, these were matched, when possible, across the considered sources (Table 1).

## 3 Results: five-facts characterisation

Five human-centered facts to characterise the clinicians, patients, and developers' role that can guarantee an ethical AI in medicine are defined. The facts framing is motivated by the questions introduced in Sects. 2.2, and follow the extended collaborative model presented in Sect. 2.3. In particular, the facts aim to suggest an answer to the crucial question: what is the role of clinicians, patients, and developers that can guarantee an ethical AI in healthcare? Four pillar ideas that arouse from the prospects of the PCC and EBM medical perspectives, the Collaborative models, and modern healthcare needs form the fundamentals for the facts' definition. The fundamentals are as follows:

(i) Collaboration and shared responsibility.
(ii) Respect for clinicians' decisions.
(iii) Education in ethics and AI for all stakeholders.
(iv) Empowerment of citizens.

Most of the ethical issues and principles covered by the five facts matched those found at high level of common consensus amongst the considered ethical codes (Table 1). The four ethical pillars [52] found convergence through exact matching within Li et al. [18]. Seven out of twelve ethical issues in [18] found exact word matching either with the WHO guidance and/or the ethical pillars. In general, the matching was done using exact word matching [56]. There were exceptions with the word "equity" that was matched to "justice and fairness" and with "Non-maleficence" that was matched to "Patient Safety". Five out of twelve ethical issues in Li et al. [18] ("Privacy", "Trust", "Solidarity", "Dignity" and "Conflict") found no word matching across sources. However, we could argue that "Privacy" is associated with "human safety", "trust" with "transparency", "solidarity" with "patient protection" and "justice", "dignity" with "non-maleficence", and "conflicts" emerge with "responsibility".

The 1st fact applies to each agent—clinicians, patients and developers—and it works as an ethical grounding for facts 2 to 5. The 2nd and 3rd facts involve clinicians, the 4th fact involve patients, and the 5th fact involve developers. Throughout the facts' presentation, we have italicised the previously published ethical concerns and principles to ease the identification of the ethical prospects underlying each fact.

The five facts are as follows:

**Fact 1: The four classical ethical pillars of the medical profession are valid for assessing AI ethical risks in healthcare**

Four principles are considered by many as the standard theoretical framework from which to analyse ethical situations in medicine [28, 52, 57]. The principles apply as follows:

1. *Respect for autonomy*: Patient autonomy and freedom should be maximised in informed medical decisions. Patients are autonomous agents are entitled to hold their own viewpoints, are free to make choices, and act voluntarily according to their values, beliefs, and preferences.
2. *Beneficence*: Any human agent involved on patients' health care should act in a patient's best interests. Beneficence is an act of charity, mercy, and kindness with a strong connotation of doing good to others including moral obligation.
3. *Non-maleficence*: Patients should be treated as ends in themselves. The principle of non-maleficence holds that there is an obligation not to inflict harm on others. It is closely associated with the maxim "primum non nocere" (above all, do no harm) as stated in the Hippocratic Oath.
4. *Justice*: Medical benefits should be distributed fairly. A concept that emphasises fairness, equality, and equity amongst individuals.

This fact works as an "ethics umbrella" as it can be applied to assess any ethical situation in medicine, and in particular, when AI is in use. We argue that clinicians, but also patients and AI developers, should be aware of the four principles and facilitate that any medical decision is made accordingly to them. Clinicians are usually exposed to the principles, so this would be no new for the collective. Following the "patient-extended" collaborative model, we claim that also patients should be informed of the ethical principles. It would have an empowering effect on patients if they could know that their respect for autonomy should be respected, or that they deserve an equal amount of resources, as it will be discussed in Fact 4. Also, developers should be introduced to the four pillars. For example, the idea of justice and fairness strongly applies to the ethical role of

developers that are entitled to build AI tools that respect humans' equality (as it is discussed in Fact 5).

**Fact 2: AI technologies are a complement and not a replacement of clinician's knowledge**

The universe of clinician's knowledge should not be replaced in whole by an automatically deployed AI recommendation but complemented by it. The ethical principle says that *doctors should make use of all their available knowledge and skills to make a clinical decision* [51]. The knowledge can come in the form of (1) *Explicit Knowledge*, that knowledge that can be codified and written, expressed in mathematical and logical language and that can be transferred to others, or (2) informed medical intuition, a type of *Tacit Knowledge*, that knowledge that cannot be codified as language or mathematics and refers more to how we do things rather than to what we do [58–60]. Tacit Knowledge can lead to decisions not readily explainable by the physician [51]. The information provided by an AI system, if codifiable, becomes part of the Explicit Knowledge. Under an EBM and a PCC perspective, the best available science should be combined with the healthcare professional's clinical experience and the patient's values to arrive at the best medical decision shared with the patient. By best available external clinical evidence, Sackett [32] meant "clinically relevant research, often from the basic sciences of medicine, but especially from patient centered clinical research into the accuracy and precision of diagnostic tests (including the clinical examination), the power of prognostic markers, and the efficacy and safety of therapeutic, rehabilitative, and preventive regimens". AI models, if appropriately built, provide such kind of information. Clinicians should incorporate the information suggested by the AI algorithm at their own discretion, always in search of meeting the *Beneficence* principle of acting in patient's best interest. To facilitate this, clinicians should be *free* to develop their clinical judgement and their tacit knowledge. There is the risk of novice clinicians becoming too dependent on AI-based recommendations and not growing their own clinical judgement [61], particularly for those difficult cases that they might feel unconfident to solve [13]. This scenario might risk disempowerment of clinicians that should be avoided by promoting self-clinical judgement development [51].

**Fact 3: Clinicians are accountable for their clinical decisions and their decisions are to be respected, regardless the assistance of an AI system**

Clinicians as a human competent agent are *responsible* of their clinical decisions and their decisions must be respected [62]. It is vital that clinicians' judgement is respected even if this is contrary to a machine's suggestion (i.e., there is *conflict*) [63, 64]. Clinicians have the potential to know the facts, science, patient context, and their own clinical skill set better than the AI [46] and should never be forced to act against their own believes per the *principle of freedom of action* [65]. A consultation with other clinicians might be helpful to agree on a final decision in case of conflict. However, the physician has the moral obligation of caring for themselves and having a deepen knowledge of themselves to identify a need to acquire further knowledge, in particular, about an AI decision support system if this is to improve patient's health [66–68]. To establish *trust* towards machine-based recommendations, under the prospect of the collaborative model, clinicians should work together with developers to learn how to use the system, understanding how it works and learn how to interpret the outputs [15, 69]. If clinicians understand how AI algorithms deploy medical suggestions, they will be more able to assess the outputs, incorporate such information in their decisions or alerting of inaccurate or unfair predictions by making sure that they always behave according to the principles of *Beneficence* and *Justice*.

**Fact 4: The empowerment and education of patients is necessary for an ethical AI in healthcare**

Patients should be considered as active agents. A patient is not a merely passive agent waiting for a diagnosis or treatment. Patients make decisions that affects their health, as if having a treatment as prescribed or attending a visit, so they are an active playing part and should be considered also as a *responsible* agent about clinical decisions in benefit of the *Respect for Autonomy* principle that states that patients should be treated as autonomous agents.

Patients can be empowered in, at least, two ways. First, clinicians and developers should make patients aware on the relevance of including their subjective experience in medical decisions as this is essential to achieve a good treatment response [29, 51, 69–74]. The EBM approach states that the unique preferences, concerns, and expectations each patient brings to a clinical encounter must be integrated into clinical decisions if they are to serve the patient [32]. AI medical tools can hardly listen to humans or incorporate their subjective patients' experiences in their automatic decisions even if big efforts are being done in the field [75, 76]. Even if Chatbots or chatGPT can show an apparent conscious behaviour in a human conversational way, this is not spontaneous or intelligent behaviour, but a task learnt from existing patterns and performed unconsciously. Only humans have consciousness about the patient situation, can develop empathy with other human beings, and have a knowledge of the context environment; these are essential factors to meet the global moral imperative of the medical profession that "each patient must be treated as a person" to preserve human *dignity* [51, 77]. AI-based decision tools are fundamentally linked with the biomedical model of disease—imperative since mid-20th in clinical practice. The biomedical model focuses on understanding human bodies as physical bodies analysable into separate parts. This mechanistic view of biology that separates body from mind is deeply set in the Western culture, mostly because of the influential work of

René Descartes [78]. The biomedical model may risk objectification and mechanisation of humans, that are the main causes of *dehumanisation* in medicine [27, 28] in breach with the *Non-maleficence* ethical principle. The WHO recognises the biopsychosocial model of disease [79] as the model to adopt. Based on that model, the health organisation defines health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" [80]. This model, in contrast with the biomedical model, states that human health cannot be categorised into biological, psychological, or social factors alone, but in their interaction; thoughts and emotions such as fear or social situation like family circumstances are considered in interaction with the biological evidence. This model can reverse the dehumanisation of medicine and disempowerment of patients [51]. In this line, giving a patient the chance and time to elaborate about their suffering should be part of any doctor visit, regardless of the assistance of any AI tool. Following the PCC and EBM perspectives, and as included in the "patient-extended" collaborative model, patients should be engaged in the decision-making that should be shared between patients and clinicians in benefit of the *Respect for Autonomy* principle. Patients should be (1) informed about an AI system making decisions on their health and educated on how the system works, (2) informed and educated on the ethical concerns this raises, and (3) empowered to receive the information from clinicians and to take part on shared decisions with them. More than 100 countries have enacted data protection laws that recognise the right not to be subject to decisions guided solely by automated processes where the outcome produces significant effect on a patient. For example, under the European Union's General Data Protection Regulation (GDPR) Article 22 [81] the EU law states that "patients' perspective on data sharing, consent and data privacy should be taken into account in healthcare and research". AI should only be used in a health care system when an informed and free consent is given.

Patients should be involved actively with doctors and developers during the AI model generation stage in benefit of the *Respect for Autonomy principle*. Recent trends on implementing Patients and Public Involvement (PPI) activities in research facilitate to actively involve patients in research projects development, educate patients in the understanding of the technology, and include patients' point of views, experiences, and expectations in algorithms design. This approach lies more closely with the public deliberation model [15]. For example, in a demo session run by our team at the King's AI festival (London, 2023) [82], a group of participants (~20 people) were (1) introduced to how an AI tool for clinical decision works, (2) introduced to related ethical dilemmas, and (3) invited to express their concerns, fears, and desires. The attendants were very engaged in asking questions about the AI agent functioning and limitations, knowing more about their rights as patients, discussing what information patients would agree to include in a model, how the relevant information to patients' health should be translated, understanding clinicians' liability, and worrying if patients would be listened by practitioners assisted with AI systems. The researchers and developers that were organising the event listened carefully to the public claims and worries, and reflected on how their practice may incorporate such sensible information. The session met the aims of both empowering and educating citizens. Clinicians and developers should run these activities systematically and regularly, and should promote the involvement of patients in search of the *Respect for autonomy*. Patients are *responsible* in enrolling themselves in PPI activities to better understand how decisions on their health are made.

**Fact 5: Developers are accountable for the automated decisions provided by the tools they develop. Their awareness and education on the ethical concerns can ensure a better alignment between algorithms and values**

Even if a medical decision is totally made in line with an automatic recommendation, the system cannot be responsible of such decision as even if AI may surpass humans in some aspects, they do not possess free will and does not have moral subjectivity [24, 83, 84]. Moreover, so far, no AI algorithm could demonstrate consciousness [85]. AI tools' developers hold ethical *responsibility* on AI performance and final medical decisions. Developers should be aware of the relevance of their actions for the principle of *beneficence*. To this end, developers should be educated in the ethical aspects related with the development of AI systems to assist health decisions [15]. If educated in AI ethics, developers would be conscious of the risk and potential harm their models could produce on humans and this could contribute for them to be more proactive in seeking strategies for a better alignment between algorithms and ethical values [23]. University Departments, Health-tech companies, and any kind of institution developing AI models for medicine should promote the education on ethics amongst their workers and should implement and use protocols to guarantee transparent models' development that can produce fair and non-discriminant outputs in search of the *Beneficence* and *Justice* ethical principles. The idea of *transparency* [2, 86] is opposite to that of non-transparent or so-called "black-box" AI algorithms, in which the patterns the algorithm follows to derive an output for a given person are opaque to the person and even to the expert developer [87, 88]. Opacity may risk the *Respect for Autonomy* ethical conduct, as in many cases, it will be very challenging if not impossible for the affected person to understand how the system worked out an output for him/her. This risks disempowerment of both patients and clinicians. *Explainable* AI, a recently developed field that allows humans understand the reasoning behind decisions or predictions made by an AI system even if it

is a black-box algorithm, should be considered to ensure transparency, as it contributes to legitimacy [89]. Bias is another central concern in fair AI development [88, 90, 91] as it risks the development of unfair models that could be *discriminatory*. For example, Obermeyer et al. found a racial bias in one widely used algorithm [92]. Black patients were assigned the same level of risk by the algorithm even if Black people were sicker than White patients, so the allocation of resources was unfair. This is clearly in conflict with both the *Justice* and *Non-maleficence* endeavours and developers should work to prevent this. Bias in AI mainly arises when the dataset in which the model is trained is not diverse enough, i.e., the training dataset is not representative of the population or phenomenon of study. An AI model trained in such data might hurt groups that were underrepresented. Bias can be diminished by deep caring of data pre-process, training algorithms in big and divers samples that are representative of the population, thoroughly testing algorithms in independent data and real settings, and using human-in-the-loop strategies where humans step in and intervene to solve a problem, what is known as the "human warranty" mentioned above. Human warranty requires application of regulatory principles upstream and downstream of the algorithm by establishing points of human supervision [19]. Other forms of discrimination may arise in models that involve predictor variables like race, gender, origin, or language in search of optimal accuracy. To battle this is challenging as a loss in accuracy may be produced by the exclusion of a politically critical feature. Even if those potentially discriminatory predictors are left out of the model, surrogate variables correlated with the excluded set might still become relevant for prediction, this being in conflict with the Justice principle. Avoiding discrimination and ensuring *solidarity* [93] and model *fairness* is central for patient protection and *safety*. Belenguer [94] suggests a full pipeline to deal with discriminatory bias in Artificial Intelligence inspired on the clinical trials testing-phases methodology. Developers should consider the diversity around the world, for example in languages, to facilitate the use of the systems. It is ethical to implore that developers have a thorough comprehension and mastery of the computational and statistical methodology involving ML algorithms' development and that are in continuous education. They should promote *sustainability* and responsiveness by regularly update their tools and/or adjust them if they seem ineffective [95]. This would contribute to build trustworthy models development [46, 96]. On the other hand, users' identity, *data security* and *privacy*, should be assured by the institutions before any AI system is deployed. Methodological limitations, such as using a small sample size or publication bias, or failure to rigorously employ nested cross-validation or testing the predictions of an AI programme on a fully independent sample need to be mentioned [97, 98]. Those developing ML tools are prompted to follow the many guidelines available on good practices in ML models' development to avoid such methodological issues [99–102].

## 4 Discussion

In this paper, I argued about the crucial need to promote the human presence in a medicine assisted by artificial agents, and the relevance of ethics to delineate the role that humans have to incorporate AI in medicine whilst respecting human values. Five facts were proposed to frame and guide the human action that can contribute to enable an AI that ethically supports clinical decisions in healthcare. The facts aimed to facilitate the understanding of the ethical challenges and the related moral actions that could prevent ethical risks if adopted by practitioners, patients, and designers. Two important advancements in our facts definition were (1) the consideration of the PCC and EBM approaches of individualised healthcare as a cornerstone to integrate ethical values in the AI pipeline [33–36] and (2) the introduction of a novel "patient-extended" collaborative model as an extension of the collaborative model [15] that emphasises the need for mutual collaboration between patients, developers and medical practitioners to achieve an ethical AI in healthcare. For each factual argument, relevant underlying ethical principles like fairness, transparency, autonomy, or responsibility were highlighted. The ethical issues and principles involved in the facts definition were found to be common ethical dilemmas in relevant ethics literature [18, 19, 52]. We found convergence on most of the ethical issues across recent sources, including a WHO guidance elaborated under consensus of more than 100 experts in the field [19]. The facts were presented as human-centered aiming to invite human stakeholders to take an ethical action. Each fact relied on a human agent, this helping to clarify who may take action.

In choosing the "facts" terminology, we were following the work by Santamaría-Velasco and Ruiz-Martínez in which authors defined the role of factual assertions as "guiding principles for action" [103]. Their definition of action linked empirical facts with normative reasons to form an explanation of rational agency with predictive capabilities. The authors conceived facts as "empirical information that is cognitively apprehended" and "regarded as an input which is later contrasted to expected (liable) behavioural responses from the agent" [103]. The facts we presented have normative reasons and intend to serve as an input for expected ethical behaviour in patients, clinicians, and developers.

We integrated the PCC and EBM medical approaches, the "patient-extended" collaborative model, jointly with the recommendations by the WHO [19] to form four pillars that served as the fundamentals for our five-fact definition.

The first pillar "Collaboration and shared responsibility" focussed on the idea that responsibility on AI-assisted clinical decisions should be shared and distributed amongst numerous human agents [19]. This pillar connects with the theoretical basis of PCC, EBM, and the "patient-extended" collaborative model for which shared decisions, collaboration, and mutual engagement between human stakeholders is central to enable an ethical AI in health. The second pillar "Respect for clinicians' decisions" placed the clinician as the potentially skilled professional who has the capability to interpret and incorporate the AI information if this is to enrich the clinical decision [46]. Following the prospects of EBM, the clinician has the duty to integrate the available science, now possibly in the form of an automated decision, with her/his own experience and the individual patient's unique circumstances to delineate a final agreed decision with the patient [33]. Based on this pillar, our facts claimed to respect the clinicians' decision not as opposed to an AI automated decision, but as a final consensus that integrates the AI outcome with the rest of available knowledge to form the clinicians' judgement. Our third pillar "Education in ethics and AI for all stakeholders" defended the idea that clinicians educated on AI and ethics will be more able to develop a knowledge-based opinion, that will serve to make an informed decision on whether stablishing trust towards an AI clinical system/recommendation or not. Educated citizens will be more capable to make their own informed decisions and become empowered citizens, idea that strongly determined our four pillar proposal "Empowerment of citizens" [104–106]. Empowerment is key to enable a PCC and EBM where patients are empowered to be central in the active discussion of medical decisions affecting their health, and clinicians feel empowered to make such decisions freely collating the information at hand. Based on these two pillars, our facts strongly advocated education on ethics and AI for practitioners, patients and developers, and empowerment of clinicians and patients.

In this work, we discussed about the role of human agents in making of the AI an ethical tool for medicine. We focussed the discussion on the role of patients, developers, and clinicians to implement the ethical principles into practice. We stressed that clinicians can contribute to an ethical AI in medicine when collaborating with developers in designing and understanding AI systems and outputs, making their own decisions in terms of deciding whether or not incorporating the AI recommendations, battling to keep on developing their own self-judgement, making the AI information interpretable to the patients, elaborating and promoting PPI activities so that patients involve themselves on development stages for a better understanding of AI-based decisions, or by alerting of inaccurate/discriminatory predictions. Patients can also contribute to an ethical AI for healthcare when being proactive in taking part in PPI activities, making and understanding health decisions, and in claiming for their rights about AI outcomes. Developers contribute to an ethical AI by working to generate "good" models, where good means that algorithms are aligned with human values, and facilitate their understanding to non-expert human agents—i.e., clinicians and patients. However, it is crucial to stress that patients, developers, and clinicians should work together with the Ministries of Health and Ministries of Information Technology to integrate ethical norms at every stage of a technology's design, development, and deployment [20, 55, 107, 108].

The practical implementation of the five facts here presented would benefit the whole community. Using the four pillars of the medical profession, any ethical situation involving AI would be assessed with a robust and validated set of principles (Fact 1). By openly acknowledging the clinicians' opinion value and by promoting education on AI systems and related ethical issues clinicians would feel safer with the implementation of AI tools, would not fear about potential human replacement and disempowerment in medicine, could better welcome the integration of automatic systems, feel more competent, and ultimately better developing their job (Facts 2, 3 and 4). By educating patients in the AI but also in the related ethical concerns, this would contribute to patients becoming empowered people able to express their circumstances and desires, therefore enriching the medical conversation and increasing patients' satisfaction. If approached by empowered and confident patients, clinicians would be more prone to listen to their patients and incorporating their views (Fact 4). However, clinicians should facilitate that patients feel safe, welcome, and listened in the doctor visit, regardless the patients' level of confidence or empowerment, for the principle of justice. By promoting ethical awareness, fostering responsibility and mastery in AI methods, fairer and less discriminatory algorithms would be developed and offered to the community (Fact 5). All of these are important advancements that would be expected to have a direct positive impact on citizens' health.

An important challenge is on how to properly align AI algorithms with human values [23, 109]. This challenge has a double focus, a normative focus that wonders what principles should be encoded, and a technical focus on how the ethical principles can be actually coded in artificial agents, so that systems reliably do what they are intended to do. For the normative focus, we considered a common consensus approach between the existing ethical codes as a proposal of values [18, 19, 23, 52]. For the technical, we highlighted the education on ethics as crucial to motivate developers to search on and apply strategies to battle bias and ensure fairness, transparency, and explainability. However, achieving this is extremely challenging particularly for artificial agents with cognitive abilities potentially surpassing our own [23, 110–112].

Whilst the consideration of the PCC and EBM approaches that were cornerstones of our work may be considered empowering and beneficial for some patients, others might find the additional responsibility stressful. These approaches could also reduce an individual's access to formal health care services [19]. Also relevant, only institutions that had active, innovative improvement-oriented cultures in which accountability and staff engagement in problem solving is promoted were found to be able to provide medical care that is both evidence based and patient centered. Implementing both goals in institutions where there is a lack of accountability, blaming, and resistance to change could be challenging [39]. However, with the emergence of AI for medical applications those institutions that are resistant to change could soon find themselves in a challenging position. The AI revolution should be taken as an opportunity to bring profound changes to their care models and start working towards adopting a more individualised and patient-centered care approach.

## 5 Conclusion

In an ever and rapidly evolving world, the future of a medicine assisted by AI is unforeseeable, even for an AI predictive algorithm. However, there is consensus that ethics will play a dramatic role in enabling the future integration of AI in healthcare, and that patients must be considered at the center upon AI implementation. The collaborative models based on PCC and EBM care approaches which advocate for an active involvement of patients together with the rest of human stakeholders in the AI scene emerge as the optimal choice to ensure a patient centered approach that in turn enables an ethical AI deployment. By educating and empowering citizens, and promoting collaborative and human interaction between medical practitioners, patients, and developers, a patient-centered healthcare could flourish in a very challenging period where machines and humans seem to be placed on a twin-pan balance that measures who will stay and who should go. For such collaborative models to work, there is a need for frameworks to guide the human action that guarantees an ethical implementation of AI in healthcare, as the five facts presented in this article intend to be.

AI have an extreme big potential for medical applications, but in the AI era, we should not forget that a person is not only made of data. Even when we talk about personalised medicine, we should keep on asking ourselves "Where is the *person* in AI-based *personal*ised medicine?" Personhood is a deep notion associated with phenomenal consciousness, intention, and free will. If automatic AI-deployed clinical suggestions are integrated straightforward, this would prevent clinicians of developing their own clinical judgement and would risk disempowerment of clinicians. If AI programmes treat patients like systems made of interacting parts, there is a risk to increase patients' mechanisation and dehumanisation, where patients' unique circumstances would not be listened, and the holistic character of human beings would not be fully respected. We, humans that develop AI tools, should make sure that the AI preserves our health and well-being, and above all, our own dignity as persons.

**Acknowledgements** This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

**Data availability** No data was used for the research described in the article.

## Declarations

**Conflict of interest** The author has no competing interests to declare that are relevant to the content of this article.



## References

1. Trujillo, A.C., Gregory, I.M., Ackerman, K.A.: Evolving Relationship between Humans and Machines, in IFAC-PapersOnLine, Elsevier B.V., 2019, pp. 366–371. https://doi.org/10.1016/j.ifacol.2019.01.015.
2. Endsley, M.R.: Toward a Theory of Situation Awareness in Dynamic Systems. Hum. Factors **37**(1), 32–64 (1995). https://doi.org/10.1518/001872095779049543
3. Howard, K.L.U.S.G.A.O.: Technology Assessment: Artificial intelligence in health care: Benefits and challenges of technologies to augment patient care, 2020. [Online]. https://www.gao.gov/products/gao-21-7sp. Accessed 22 Sep 2023
4. Bajwa, J., Munir, U., Nori, A., Williams, B.: Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc. J. **8**(2), e188–e194 (2021). https://doi.org/10.7861/fhj.2021-0095
5. Woods, T., Ream, M. Demestihas, M.-A., Hertz, S., Steere, A., Roberts, S.: Artificial Intelligence: How to get it right Putting policy into practice for safe data-driven innovation in health and care, 2019. [Online]. https://transform.england.nhs.uk/media/documents/NHSX_AI_report.pdf. Accessed 22 Sep 2023

6. S. O'Meara, "China's data-driven dream to overhaul health care, Nature, **598** (2021). [Online]. https://www.nature.com/articles/d41586-021-02694-1. Accessed 1 Aug 2023
7. Whicher, D., Rapp, T.: The value of artificial intelligence for healthcare decision making—lessons learned. Value Health **25**(3), 328–330 (2022). https://doi.org/10.1016/j.jval.2021.12.009
8. Giordano, C., Brennan, M., Mohamed, B., Rashidi, P., Modave, F., Tighe, P.: Accessing artificial intelligence for clinical decision-making. Front Digital Health (2021). https://doi.org/10.3389/fdgth.2021.645232
9. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future Healthc. J. **6**(2), 94–98 (2019). https://doi.org/10.7861/futurehosp.6-2-94
10. Ali, O., Abdelbaki, W., Shrestha, A., Elbasi, E., Alryalat, M.A.A., Dwivedi, Y.K.: A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. J. Innov. Knowl. (2023). https://doi.org/10.1016/j.jik.2023.100333
11. Bohr, A., Memarzadeh, K.: "The rise of artificial intelligence in healthcare applications. In: Artificial intelligence in healthcare, pp. 25–60. Elsevier, Amsterdam (2020). https://doi.org/10.1016/B978-0-12-818438-7.00002-2
12. Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. **25**(1), 44–56 (2019). https://doi.org/10.1038/s41591-018-0300-7
13. Chekroud, A.M., et al.: The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry **20**(2), 154–170 (2021). https://doi.org/10.1002/wps.20882
14. Aristidou, A., Jena, R., Topol, E.J.: Bridging the chasm between AI and clinical implementation. Lancet **399**(10325), 620 (2022). https://doi.org/10.1016/S0140-6736(22)00235-5
15. Gundersen, T., Bærøe, K.: The future ethics of artificial intelligence in medicine: making sense of collaborative models. Sci. Eng. Ethics (2022). https://doi.org/10.1007/s11948-022-00369-2
16. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
17. Murphy, K., et al.: Artificial intelligence for good health: a scoping review of the ethics literature. BMC Med. Ethics (2021). https://doi.org/10.1186/s12910-021-00577-8
18. Li, F., Ruijs, N., Lu, Y.: "Ethics & AI: a systematic review on ethical concerns and related strategies for designing with AI in healthcare. AI **4**(1), 28–53 (2023). https://doi.org/10.3390/ai4010003
19. World Health Organisation: *Ethics and governance of Artificial Intelligence for health*. 2021. [Online]. http://apps.who.int/bookorders.
20. European Parlament: Artificial Intelligence Act. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html, 2023.
21. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4
22. Bleher, H., Braun, M.: Reflections on putting AI ethics into practice: how three AI ethics approaches conceptualize theory and practice. Sci. Eng. Ethics (2023). https://doi.org/10.1007/s11948-023-00443-3
23. Gabriel, I.: Artificial intelligence, values, and alignment. Minds Mach. (Dordr) **30**(3), 411–437 (2020). https://doi.org/10.1007/s11023-020-09539-2
24. Zhang, J., Ming Zhang, Z.: "Ethics and governance of trustworthy medical artificial intelligence. BMC Med. Inform. Decis. Mak. (2023). https://doi.org/10.1186/s12911-023-02103-9
25. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**(2), 185–193 (2019). https://doi.org/10.1007/s13347-019-00354-x
26. Haslam, N., Loughnan, S.: Dehumanization and infrahumanization. Ann. Rev. Psychol. **65**, 399–423 (2014). https://doi.org/10.1146/annurev-psych-010213-115045
27. Lekka, D., et al.: Dehumanization of hospitalized patients and self-dehumanization by health professionals and the general population in Greece. Cureus (2021). https://doi.org/10.7759/cureus.20182
28. Haque, O.S., Waytz, A.: Dehumanization in medicine: causes, solutions, and functions. Perspect. Psychol. Sci. **7**(2), 176–186 (2012). https://doi.org/10.1177/1745691611429706
29. Stewart, M.: Towards a global definition of patient centred care. Br. Med. J. **322**, 444–445 (2001). https://doi.org/10.1136/bmj.322.7284.444
30. World Health Organization: Framework on integrated, people-centered health services: report by the secretariat. Geneva, 2016. [Online]. https://iris.who.int/handle/10665/252698 Accessed 22 Sep 2023
31. Yehualashet, D.E., Seboka, T., Mamo, T.T., Yawo, M.N.: Evidence-Based Medicine-A New Approach for Medical Education and Practice. InTechOpen, New York (2022). https://doi.org/10.5772/intechopen.107298
32. Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., Richardson, W.S.: Evidence-based-medicine: what it is and what it isn't. Br. Med. J. **312**, 71–72 (1996)
33. Tenny, S., Varacallo, M.: Evidence Based Medicine. StatPearls Publishing LLC, New York (2023)
34. Straus, S., Haynes, B., Glasziou, P., Dickersin, K., Guyatt, G.: Misunderstandings, misperceptions, and mistakes. Evid. Based Med. **12**(1), 2–3 (2007). https://doi.org/10.1136/ebm.12.1.2-a
35. Reynolds, A.: Patient-centered care. Radiol. Technol. **81**(2), 133–147 (2009)
36. Gartner, J.-B., Abasse, K.S., Bergeron, F., Landa, P., Lemaire, C., Côté, A.: Definition and conceptualization of the patient-centered care pathway, a proposed integrative framework for consensus: a Concept analysis and systematic review. BMC Health Serv. Res. **22**(1), 558 (2022). https://doi.org/10.1186/s12913-022-07960-0
37. Stewart, M., Brown, J.B., Weston, W.W., McWhinney, I.R., McWilliam, C.L., Freeman, T.R.: Patient-centered medicine: Transforming the clinical method. Sage Publications Inc, Thousand Oaks (1995)
38. Little, P.: Preferences of patients for patient centred approach to consultation in primary care: observational study. BMJ **322**(7284), 468–468 (2001). https://doi.org/10.1136/bmj.322.7284.468
39. Engle, R.L., et al.: Evidence-based practice and patient-centered care: doing both well. Health Care Manage. Rev. **46**(3), 174–184 (2021). https://doi.org/10.1097/HMR.0000000000000254
40. Baker, A.: Book: crossing the quality chasm: a new health system for the 21st century. BMJ **323**(7322), 1192–1192 (2001). https://doi.org/10.1136/bmj.323.7322.1192
41. Institute of Medicine of America: Delivering High-Quality Cancer Care. National Academies Press, Washington, D.C. (2013)
42. Engineering, medicine National Academies of sciences: *Crossing the Global Quality Chasm*. Washington, D.C.: National Academies Press, 2018. doi: https://doi.org/10.17226/25152.
43. Topol, E.:"The Topol Review," London, 2019. [Online]. https://topol.hee.nhs.uk/. Accessed 5 Aug 2023.
44. Ng, A.Y., et al.: Artificial isntelligence as supporting reader in breast screening: a novel workflow to preserve quality and reduce workload. J Breast Imaging **5**(3), 267–276 (2023). https://doi.org/10.1093/jbi/wbad010

45. de Vries, C.F., et al.: AI in breast screening mammography: breast screening readers' perspectives. Insights Imaging (2022). https://doi.org/10.1186/s13244-022-01322-4
46. Jones, C., Thornton, J., Wyatt, J.C.: Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. Med. Law Rev. (2023). https://doi.org/10.1093/medlaw/fwad013
47. Liberati, E.G., et al.: What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implementation Sci. (2017). https://doi.org/10.1186/s13012-017-0644-2
48. Petkus, H., Hoogewerf, J., Wyatt, J.C.: "What do senior physicians think about AI and clinical decision support systems: Quantitative and qualitative analysis of data from specialty societies. Clin. Med. J. R. Coll. Physicians Lond. **20**(3), 324–328 (2020). https://doi.org/10.7861/clinmed.2019-0317
49. Laï, M.C., Brian, M., Mamzer, M.F.: Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. J. Transl. Med. (2020). https://doi.org/10.1186/s12967-019-02204-y
50. Rousseau, N.: Practice based, longitudinal, qualitative interview study of computerised evidence based guidelines in primary care. BMJ **326**(7384), 314–314 (2003). https://doi.org/10.1136/bmj.326.7384.314
51. Borrell-Carrió, F., Suchman, A.L., Epstein, R.M.: The biopsychosocial model 25 years later: Principles, practice, and scientific inquiry. Ann. Fam. Med. **2**(6), 576–582 (2004). https://doi.org/10.1370/afm.245
52. Beauchamp, T.L., Childress, J.F.: Principies of Biomedical Ethics. Oxford Univeristy Press, Oxford (1979)
53. Gillon, R.: Defending the four principles approach as a good basis for good medical practice and therefore for good medical ethics. J. Med. Ethics **41**(1), 111–116 (2015). https://doi.org/10.1136/medethics-2014-102282
54. Beil, M., Proft, I., van Heerden, D., Sviri, S., van Heerden, P.V.: Ethical considerations about artificial intelligence for prognostication in intensive care. Intensive Care Med. Exp. (2019). https://doi.org/10.1186/s40635-019-0286-6
55. European Commission: "Ethics By Design and Ethics of Use Approaches for Artificial Intelligence," 2021. [Online]. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf. Accessed 22 Sep 2023
56. Im, D., Pyo, J., Lee, H., Jung, H., Ock, M.: Qualitative research in healthcare: data analysis. J. Prev. Med. Public Health **56**(2), 100–110 (2023). https://doi.org/10.3961/jpmph.22.471
57. Jahn, W.T.: The 4 basic ethical principles that apply to forensic activities are respect for autonomy, beneficence, nonmaleficence, and justice. J. Chiropr. Med. **10**(3), 225–226 (2011). https://doi.org/10.1016/j.jcm.2011.08.004
58. Polanyi, M.: Personal Knowledge: Towards a Post-Critical Philosophy. 1958.
59. Thornton, T.: Tacit knowledge as the unifying factor in evidence based medicine and clinical judgement. Philos. Ethics Humanities Med. (2006). https://doi.org/10.1186/1747-5341-1-2
60. Kothari, A., Rudman, D., Dobbins, M., Rouse, M., Sibbald, S., Edwards, N.: The use of tacit and explicit knowledge in public health: a qualitative study. Implement. Sci. (2012). https://doi.org/10.1186/1748-5908-7-20
61. Brangier, É., Hammes-Adelé, S.: Beyond the technology acceptance model: elements to validate the human-technology symbiosis model. In: Robertson, M.M. (ed.) Ergonomics and Health Aspects of Work with Computers, pp. 13–21. Springer, Berlin (2011)
62. Van Cauwenberge, D., Van Biesen, W., Decruyenaere, J., Leune, T., Sterckx, S.: 'Many roads lead to Rome and the Artificial Intelligence only shows me one road': an interview study on physician attitudes regarding the implementation of computerised clinical decision support systems. BMC Med. Ethics (2022). https://doi.org/10.1186/s12910-022-00787-8
63. Starke, G., De Clercq, E., Borgwardt, S., Elger, B.S.: Computing schizophrenia: Ethical challenges for machine learning in psychiatry. Psychol. Med. **51**(15), 2515–2521 (2021). https://doi.org/10.1017/S0033291720001683
64. Faden, R.R., Kass, N.E., Goodman, S.N., Pronovost, P., Tunis, S., Beauchamp, T.L.: An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. Hastings Cent. Rep. **43**(s1), S16–S27 (2013). https://doi.org/10.1002/hast.134
65. Boff, K.R.: Revolutions and shifting paradigms in human factors & ergonomics. Appl. Ergon. **37**(4), 391–399 (2006). https://doi.org/10.1016/j.apergo.2006.04.003
66. Gauld, C., Micoulaud-Franchi, J.-A., Dumas, G.: Comment on Starke et al.: 'Computing schizophrenia: ethical challenges for machine learning in psychiatry': From machine learning to student learning: pedagogical challenges for psychiatry—Corrigendum. Psychol. Med. **51**(14), 2514–2514 (2021). https://doi.org/10.1017/S0033291721000684
67. Epstein, R.M.: Mindful practice. JAMA **282**(9), 833–839 (1999). https://doi.org/10.1001/jama.282.9.833
68. Novack, D.H.: Calibrating the physician. JAMA **278**(6), 502 (1997). https://doi.org/10.1001/jama.1997.03550060078040
69. Tizón, J.: Componentes psicológicos de la práctica médica. Una perspectiva desde APS Barcelona, 1988.
70. Entralgo, P.L.: Doctor and Patient. Weidenfeld & Nicolson, London (1969)
71. Epstein, R.M.: Just being. West J Med, vol. 174, 2001, [Online]. www.umassed.edu
72. Tresolini, C., Pew-Fetzer Task Force: Health Professions Education and Relationship-Centered Care. San Francisco, California, 1994. [Online]. https://healthforce.ucsf.edu/publications/health-professions-education-and-relationship-centered-care. Accessed 22 Sep 2023
73. Mead, N., Bower, P.: Measuring patient-centredness: a comparison of three observation-based instruments. Patient Educ. Couns. **39**(1), 71–80 (2000). https://doi.org/10.1016/S0738-3991(99)00092-0
74. Dordević, V., Braš, M., Brajković, L.: Person-centered medical interview. Croat. Med. J. **53**(4), 310–313 (2012). https://doi.org/10.3325/cmj.2012.53.310
75. Khanbhai, M., et al.: Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. Int. J. Med. Inform. (2022). https://doi.org/10.1016/j.ijmedinf.2021.104642
76. Sertolli, B., Ren, Z., Schuller, B.W., Cummins, N.: Representation transfer learning from deep end-to-end speech recognition networks for the classification of health states from speech. Comput. Speech Lang. (2021). https://doi.org/10.1016/j.csl.2021.101204
77. Luxton, D.D.: Recommendations for the ethical use and design of artificial intelligent care providers. Artif. Intell. Med. **62**(1), 1–10 (2014). https://doi.org/10.1016/j.artmed.2014.06.004
78. Rocca, E., Anjum, R.L.: Complexity, reductionism and the biomedical model. In: Anjum, R.L., Rocca, E., Copeland, S. (eds.) Rethinking Causality, Complexity and Evidence for the Unique Patient, pp. 75–94. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-41239-5
79. Engel, G.L.: The need for a new medical model: a challenge for biomedicine. Science (1979) **196**(4286), 129–136 (1977). https://doi.org/10.1126/science.847460.

80. World Health Organisation: Constitution of the world health organization. 1948. [Online]. https://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1. Accessed 4 Aug 2023
81. European Parliament and Council of the European Union: GDPR: General Data Protection Regulation (L119), 2016. [Online]. https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed 22 Sep 2023
82. Iniesta, R.: How an AI-based clinical decision tool works. https://www.kcl.ac.uk/events/ethics-ai-based-medical-tools, 2023.
83. Gan, S.P.: How can robots have rights. Ethics Res. **3**, 126–130 (2017)
84. Sparrow, R.: Killer Robots. J. Appl. Philos. **24**(1), 62–77 (2007). https://doi.org/10.1111/j.1468-5930.2007.00346.x
85. Coeckelbergh, M.: Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci. Eng. Ethics **26**(4), 2051–2068 (2020). https://doi.org/10.1007/s11948-019-00146-8
86. Gerber, A., Derckx, P., Döppner, D. A., Schoder, D.: Conceptualization of the Human-Machine Symbiosis A Literature Review. 2020. [Online]. https://hdl.handle.net/10125/63775 978-0-9981331-3-3. Accessed 4 Aug 2023
87. Burrell, J.: How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. **3**(1), 205395171562251 (2016). https://doi.org/10.1177/2053951715622512
88. Müller, V. C.: Ethics of Artificial Intelligence and Robotics. The Stanford Encyclopedia of Philosophy. 2021. [Online]. https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/. Accessed 4 Aug 2023
89. Vilone, G., Longo, L.: Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion **76**, 89–106 (2021). https://doi.org/10.1016/j.inffus.2021.05.009
90. Floridi, L., Taddeo, M.: What is data ethics? Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **374**(2083), 20160360 (2016). https://doi.org/10.1098/rsta.2016.0360
91. Mittelstadt, B.D., Floridi, L.: The ethics of big data: current and foreseeable issues in biomedical contexts. Sci. Eng. Ethics **22**(2), 303–341 (2016). https://doi.org/10.1007/s11948-015-9652-2
92. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science (1979) **366**(6464), 447–453 (2019). https://doi.org/10.1126/science.aax2342
93. Comite consultatif national d'éthique pour les sciences de la vie et de la sante: Digital technology and healthcare: which ethical issues for which regulations? Paris, 2018.
94. Belenguer, L.: AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. AI Ethics **2**(4), 771–787 (2022). https://doi.org/10.1007/s43681-022-00138-8
95. Van Houten, H.: Five guiding principles for responsible use of AI in healthcare and healthy living, 2020. https://www.philips.com/a-w/about/news/archive/blogs/innovation-matters/20200121-five-guiding-principles-for-responsible-use-of-ai-in-healthcare-and-healthy-living.html. Accessed 4 Aug 2023.
96. Bærøe, K., Miyata-Sturm, A., Henden, E.: How to achieve trustworthy artificial intelligence for health. Bull. World Health Organ. **98**(4), 257–262 (2020). https://doi.org/10.2471/BLT.19.237289
97. Mechelli, A., Vieira, S.: From models to tools: clinical translation of machine learning studies in psychosis. NPJ Schizophr. **6**(1), 4 (2020). https://doi.org/10.1038/s41537-020-0094-8
98. Stahl, D., Pickles, A.: Fact or fiction: reducing the proportion and impact of false positives. Psychol. Med. **48**(7), 1084–1091 (2018). https://doi.org/10.1017/S003329171700294X
99. Wolff, R.F., et al.: PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann. Intern. Med. **170**(1), 51 (2019). https://doi.org/10.7326/M18-1376
100. Steyerberg, E.W., Vergouwe, Y.: Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur. Heart J. **35**(29), 1925–1931 (2014). https://doi.org/10.1093/eurheartj/ehu207
101. Steyerberg, E.W., et al.: Assessing the performance of prediction models. Epidemiology **21**(1), 128–138 (2010). https://doi.org/10.1097/EDE.0b013e3181c30fb2
102. European Information Technologies Certification Academy: European AI certificate, 2023. https://eitca.org/eitca-ai-artificial-intelligence-academy/. Accessed 4 Aug 2023.
103. Santamaría-Velasco, F., Ruiz-Martínez, S.: Redefining action: facts and beliefs in the social world. Cinta de moebio **73**, 24–35 (2022). https://doi.org/10.4067/s0717-554x2022000100024
104. Feste, C., Anderson, R.M.: Empowerment: from philosophy to practice. Patient Educ. Couns. **26**(1–3), 139–144 (1995). https://doi.org/10.1016/0738-3991(95)00730-N
105. NHS England: Empowering people in their care, 2019. https://www.england.nhs.uk/blog/empowering-people-in-their-care/. Accessed 5 Aug 2023.
106. Lawson, T.: Empowerment in education: liberation, governance or a distraction? A review. Power Educ. **3**(2), 89–103 (2011). https://doi.org/10.2304/power.2011.3.2.89
107. European Commission: Ethics guidelines for trustworthy AI, 2019. [Online]. https://ec.europa.eu/digital-
108. T. A. and A. R. to C. R. Science: Artificial Intelligence in Health Care: Benefits and Challenges of Technologies to Augment Patient Care With content from the National Academy of Medicine, 2020. Accessed: Sep. 22, 2023. [Online]. https://www.gao.gov/products/gao-21-7sp
109. Russell, S.J.: Human Compatible: AI and the Problem of Control. Allen Lane/Penguin Books, c2009, London (2019)
110. Irving, G., Christiano, P., Amodei, D.: AI safety via debate, 2018, https://doi.org/10.48550/arXiv.1805.00899.
111. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., Legg, S.: Scalable agent alignment via reward modeling: a research direction. 2018. [Online]. https://users.cs.utah.edu/~dsbrown/readings/scalable_alignment_direction.pdf. Accessed 22 Sep 2023
112. Christiano, P.: Prosaic AI alignment. AI Alignment, 2016. [Online]. https://ai-alignment.com/prosaic-ai-control-b959644d79c2. Accessed 4 Aug 2023