Intelligent Medicine 4 (2024) 52–57

Contents lists available at ScienceDirect

Intelligent Medicine

journal homepage: www.elsevier.com/locate/imed

Review

Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm"

Hanhui Xu [1,*], Kyle Michael James Shuttleworth [2]

[1] School of Medicine, Nankai University, Tianjin, 300071, China
[2] Department of Global Connectivity, Akita International University, Okutsubakidai-193-2 Yuwatsubakigawa, Akita, 010-1211, Japan

ARTICLE INFO

Keywords:
Medical artificial intelligence
Black box problem
Do no harm
Paternalism

ABSTRACT

One concern about the application of medical artificial intelligence (AI) regards the "black box" feature which can only be viewed in terms of its inputs and outputs, with no way to understand the AI's algorithm. This is problematic because patients, physicians, and even designers, do not understand why or how a treatment recommendation is produced by AI technologies. One view claims that the worry about black-box medicine is unreasonable because AI systems outperform human doctors in identifying the disease. Furthermore, under the medical AI-physician-patient model, the physician can undertake the responsibility of interpreting the medical AI's diagnosis. In this study, we focus on the potential harm caused by the unexplainability feature of medical AI and try to show that such possible harm is underestimated. We will seek to contribute to the literature from three aspects. First, we appealed to a thought experiment to show that although the medical AI systems perform better on accuracy, the harm caused by medical AI's misdiagnoses may be more serious than that caused by human doctors' misdiagnoses in some cases. Second, in patient-centered medicine, physicians were obligated to provide adequate information to their patients in medical decision-making. However, the unexplainability feature of medical AI systems would limit the patient's autonomy. Last, we tried to illustrate the psychological and financial burdens that may be caused by the unexplainablity feature of medical AI systems, which seems to be ignored by the previous ethical discussions.

1. Introduction

The global artificial intelligence market in healthcare was estimated to be worth $15.1 billion in 2022 [1]. The size is expected to close at $12.63 billion as of 2023 and is projected to flourish to $187.76 billion by 2031 [2]. In terms of North America, the healthcare artificial intelligence market was estimated to be worth $6.8 billion, and 521 medical artificial intelligence (AI) devices have been approved by US Food and Drug Administration (FDA) by 2022 [3]. The healthcare artificial intelligence industry is growing so rapidly that we need to be cautious about the problems that may arise.

AI, as defined by Umbrello and van de Poel [4] "is understood as a class of technologies that are autonomous, interactive, adaptive, and capable of carrying out human-like tasks". In this article, our focus will be on AI technologies based on deep learning (DL), which is understood as "a subset of machine learning which is structured similar to human brain processing, taking into account multiple data sets at the same time, which are evaluated and reprocessed for second and third different evaluations and so on, until reaching an output" [5]. These programs are commonly referred to as "black boxes" due to the fact that their internal workings are undisclosed and only the inputs and outputs are known.

Such a feature has caused tension between the accuracy and explainability of these systems, as opaque models seem to be more accurate, yet difficult to interpret. From an ethical perspective, the existing literature shows different attitudes to this issue. Some scholars emphasize that although some explainable medical AI has been created it has a long way to go before medical AI systems become explainable enough [6–11]. Accordingly, medical AI systems are required explainable enough to make it possible for a physician to identify the incorrect results and for an individual to contest the decision of the system. However, the current medical AI systems cannot seem to achieve it. Lauritsen et al. recognize transparency and explainability are a necessity for introducing AI into clinical practice [12]. As they note, "Clinicians must be able to understand the underlying reasoning of AI models so they can trust the predictions and be able to identify individual cases in which an AI model potentially gives incorrect predictions" [12]. In an attempt to address this issue, they present an explainable AI early warning score (xAI-EWS) system for the early detection of acute critical illness. Whilst Lauritsen et al's xAI-EWS system is shown to have high predictive performance while enabling the possibility to explain predictions, this is only applicable to the early detection of acute critical illness.

* Corresponding author: Hanhui Xu, School of Medicine, Nankai University, Tianjin 300071, China (Email: 018205@nankai.edu.cn).

https://doi.org/10.1016/j.imed.2023.08.001
Received 20 May 2023; Received in revised form 29 July 2023; Accepted 1 August 2023

Ghassemi et al. [13] argue against the position that trust can be engendered by explainable AI, and claim such a position represents a false hope. Their position is based on the insufficiency of current approaches to explainability. As they emphasize, "we suggest that end users of explainable AI, including clinicians, lawmakers, and regulators, be aware of the limitations of explainable AI as it currently exists" [13]. On account of current explainable AI, Ghassemi et al. [13] go on to pessimistically suggest that "explainability for patient-level decision making is unlikely to advance these goals in meaningful ways". However, whilst we recognize the limitations of current explainable AI, our approach is to request more developed means of explainability are augment in medical AI in the future.

By contrast, some scholars deny the necessity of any such concern [14–16]. According to them, although medical AI systems do not provide any rationale for their diagnoses or treatment suggestions, their decisions deserve our trust. The reason is that such machines are consistently more accurate than human doctors in terms of disease diagnosis, and as such, it can be inferred that they will perform better, and be safer than human doctors. Studies have indicated that medical AI can perform equally well to, or even outperform expert radiologists and pathologists in terms of accuracy when detecting, classifying, and segmenting tumors in ultrasonography, X-ray imaging, MRI scans, and digitalized microscopy slides [17–19]. Thus, if the primary concern regarding the black box in medical AI is that such systems are unsafe, then this concern cannot be reasonably justified since safety is not grounded on explainability but on accuracy. In addition, human doctors' judgements are not always explainable. In many cases, a human doctor's diagnosis is based on experience, intuitions and even guesswork, rather than an understanding of the mechanism of the disease. Hence, if we can trust human doctors, it seems reasonable to trust the results produced by medical AI.

In this article, we focus on the possible harm caused by the unexplainability feature of medical AI and try to show that such possible harm is underestimated. We will seek to contribute to the literature from three aspects. Firstly, we will appeal to a thought experiment to show that although the medical AI systems perform better on accuracy the harm caused by medical AI's misdiagnoses may be more serious than that caused by human doctors' misdiagnoses in some cases. Due to the black box feature, medical AI systems might make incomprehensible mistakes that are difficult to detect. Failure to detect such mistakes may cause harm to patients. And the harm caused by such misdiagnoses may be more serious than that caused by human doctors' misdiagnoses in some cases. If it is logically sound, then the view that accuracy grounds safety will be challenged. Secondly, in patient-centered medicine, physicians are obligated to provide adequate information to their patients in medical decision-making. However, the opacity of medical AI systems seems to make it impossible and then returns patients to a new paternalistic model of the physician-patient relationship, in which the patient's best interests may not be appropriately protected. Namely, the treatments suggested by medical AI systems may not be in the patient's best interests, by which we mean a decision made on behalf of the patient which is considered to be the most beneficial course of action for that patient. The reason is that the treatment suggestions provided by medical AI are usually based on a professional perspective which ignores the patient's values, which we understand to entail the system of beliefs which shapes a patient's outlook on life and orders their actions and goals in terms of importance. Thirdly, the uncertainty caused by the unexplainability of medical AI systems may cause emotional harm. In addition, the financial burden brought by medical AI should also be considered.

2. The black box problem

As explained in the introduction, our concern with AI is specifically with DL systems. Such technology is based on the "artificial neural network (ANN)", which is inspired by the human brain. The basic unit responsible for activity in the human brain is the "neuron". There are tens of billions of neurons in the human brain, which are interconnected into a larger structure called a "neural network." In DL, scientists and engineers try to imitate the "neural network" of the human brain to build a similar learning strategy, also named "neural network" [20]. Medical AI systems have three features. Firstly, they can develop the capacity for self-learning. Traditional computational programs are rule-based, that is, the mechanism within these programs is set by a collection of rules. By contrast, DL systems can deal with a large amount of data and develop the capacity for self-learning. Such algorithms enable them to produce the desired output based on the input data [21-22].

Secondly, medical AI systems usually have high predictive accuracy in diagnosis. As mentioned above, medical AI systems require large amounts of data to develop their capacity for self-learning. Such data is separated into training and testing. Designers use the training data to fit the model and testing data to test it. After being trained, medical AI systems show extremely high accuracy in testing. They even surpassed human experts in identifying particular diseases. For instance, the deep patient system was able to accurately predict "hidden" diseases in new cases of more than 700,000 electronic health records through DL [21]. Furthermore, it appeared to be unbelievably accurate in predicting mental illnesses such as schizophrenia, which impressed human doctors due to the fact that psychiatric disorders have always been extraordinarily difficult to predict. In another case, the AI system made by Google's DeepMind surpassed human experts in breast cancer prediction [22]. After learning X-ray images from nearly 29,000 female patients, the system was able to predict breast cancer in new mammograms, outperforming six human radiologists.

Thirdly, the diagnoses and treatment suggestions provided by medical AI systems are unexplainable. As mentioned above, medical AI systems discussed in this article are based on artificial neural networks, which are imitations of the neural networks of the human brain. Just as human beings have not yet figured out how the human brain works, medical AI systems are "black boxes" for patients, doctors, and even designers. The so-called "black box" indicates that people do not understand the internal working mechanism of such systems. medical AI systems can (for instance) predict tumor response to a particular drug based on allelic patterns among thousands of genes or predict lung cancer prognosis by analyzing microscopic images—all without understanding or identifying why or how those patterns matter. In the case of Deep Patient, researchers only knew that the system can accurately predict diseases including psychiatric disorders like schizophrenia through DL. However, the questions that have puzzled researchers include how such accurate predictions can be achieved and how the system can reach its diagnostic conclusions. As Watson makes explicit, "(the primary concern regarding the application of medical AI systems is) …… a lack of understanding among patients and doctors about how predictions are made. This is especially true of some top performing algorithms, like the deep neural networks used in image recognition software. These models may reliably discriminate between malignant and benign tumors, but they offer no explanation for their judgments" [23].

As mentioned above, it seems to be a long way to go before medical AI systems become explainable enough to make it possible for a physician to identify the incorrect results and for an individual to contest the decision of the system. Thus, the potential harm caused by medical AI systems should be seriously taken into account.

3. Medical malpractice and medical AI

The maxim "do no harm" is derived from *The Epidemics*, a collection of clinical syndromes recorded by Greek physicians, and which is associated with Hippocrates [16]. In one of the treatises, ethical codes are defined as follows: "As to diseases, make a habit of two things—to help, or at least to do no harm." [24]. We can also see a similar expression in the Hippocratic Oath, which is usually regarded as the oldest ethical document in medicine: "I will use treatment to help the sick according

Table 1 Cases of mega harms caused by AI in medicine

Time	Cases	Problems
2016	Tay, the Twitter bot [26]	Within a day of its release, the bot began to spout racist and sexist comments, having been taught by some users to repeat offensive language.
2016	CAMPAS, Correctional Offender Management Profiling for Alternative Sanctions tool [27]	The system uses historical crime data to predict where crimes are likely to occur and who is likely to commit them. Since the data used to train these systems often reflects existing biases in law enforcement, the predictions are questioned to reinforce and amplify existing biases.
2018	Uber self-driving car [28]	The car struck and killed a pedestrian in Arizona.
2020	Facebook AI facial recognition [29]	The system was found to be racially biased, putting 'primates' label on video of black men.
2020	Dutch "System Risk Indication" (SyRI) program [30]	SyRI was designed to identify potential welfare fraud by analyzing data from various government agencies, including tax authorities, social security agencies, and housing agencies. The system used machine learning algorithms to identify patterns and anomalies in the data that could indicate potential fraud. However, the SyRI program was controversial, with critics arguing that it violated privacy rights and targeted vulnerable populations.

to my ability and judgment, but never with a view to injury and wrongdoing" [24].

Since then, "do no harm" has become the most fundamental ethical rule for physicians. Here, a few points should be clarified. The proper interpretation of "do no harm" is that physicians have a moral obligation not to cause unnecessary harm. Unnecessary harm, here, can be understood as harm which is neither a necessary process to avoid more serious harm nor an option reasonable enough for bringing given benefits. In terms of medical practices, unnecessary harm caused by physicians can be divided into two types. The first type of unnecessary harm is usually caused by physicians' misdiagnoses and erroneous treatments. The second type of unnecessary harm involves what is termed "medical paternalism". That is, simply speaking, physicians make medical decisions on behalf of their patients on the grounds that the physician knows what is best for their patients. Common examples of medical paternalism are those in which physicians induce patients to do what they believe to fit the patients' best interests by withholding information or misinforming them. Allen Buchanan makes this position explicit when he writes:

"……paternalism is interference with a person's freedom of action or freedom of information, or the deliberate dissemination of misinformation, where the alleged justification of interfering or misinforming is that it is for the good of the person who is interfered with or misinformed" [25].

It can be inferred that medical paternalism, in which physicians know their patients' interests better than the patients themselves, and the physicians subsequently have the authority to choose for their patients, even without any explanation of the patient's informed consent, is morally unacceptable. It is morally unacceptable because it ignores patients' values and preferences, which we define as the decision between options and the choice of a particular treatment based on the patient's physical and psychological needs. Doing so causes serious harm.

While AI has made significant advancements in recent years, there have been several instances where AI systems have failed to perform as expected or have produced unintended consequences. Here are some examples (Table 1).

Some mistakes also occur in the clinical practice. Watson was deployed at UB Songdo Hospital in Mongolia. The system was reported to have inappropriately recommended the drug taxane for a patient whose history would contraindicate the use of that drug [10]. Fortunately, this error was noted by an oncology specialist. In some cases, the results provided by medical AI systems even contradict our common sense. Rich Caruana and colleagues report that although one medical AI system was more accurate than other methods in diagnosing the probability of death from pneumonia, it ranked asthmatic patients lower than the general population. This finding is counterintuitive because patients with a history of asthma are usually admitted directly to the intensive care unit (ICU) because patients with a history of asthma are usually admitted directly to ICU to receive aggressive medical care. Medical care; it is this extra care that gives them a lower probability of dying. Without this active care, asthmatics patients have a higher probability of dying from pneumonia. Their score in the system is considered misleading because it does not reflect the patient's underlying medical needs. The unexplainability feature of medical AI systems makes it much more difficult to identify and detect medical error. Considering it will put the patient at potential risk of harm, it seems reasonable to require medical AI systems to be explainable enough to make it possible for a physician to identify the incorrect results.

However, as mentioned above, some scholars deny the necessity of any such concern. One reason is that although medical AI systems are not 100 percent accurate in predicting diseases, they perform better than human doctors on accuracy.

As Alex John London wrote: "… the ability to explain how results are produced can be less important than the ability to produce such results and empirically verify their accuracy…. a blanket requirement that machine-learning systems in medicine be explainable or interpretable is unfounded" [31].

Furthermore, in terms of unexplainability, human doctors cannot always explain their diagnoses and treatment recommendations clearly. Actually, they usually make judgments based on their experience, intuitions, and even guesswork. This phenomenon was not only prevalent in the pre-scientific period, but also present medical practice. If both medical AI systems and human doctors have the unexplainability feature, then fears about the implications of medical AI systems seem overcautious and misplaced.

As Vijay Pande said: "Human intelligence itself is – and always has been – a black box.… Sure, the doctor would probably give a few indicators about what pointed her in a certain direction – but there would also be an element of guessing, of following hunches" [32].

In our opinion, the misdiagnosis by medical AI systems differs from the misdiagnosis by human doctors not in accuracy but in nature. As mentioned above, due to the unknown of the internal mechanism of medical AI systems, it is difficult to realize what factors have been taken into account and how these factors are used. Hence, some irrelevant factors may be taken into account when medical AI systems deal with massive amounts of data. The involvement of these irrelevant factors may produce counterintuitive results similar to those mentioned above. If an AI system can make errors that a human doctor would never make, then even if the AI is more accurate than a human doctor, the harm caused to the patient by an AI's misdiagnosis could be extremely serious, so serious that it could exceed the harm caused by the average human doctor's misdiagnosis. This is overlooked in previous discussions in terms of potential medical malpractice made by medical AI systems.

Such low-level errors that do not normally occur for a human being made by AI have been reported. In one test, an AI-driven vehicle misread a "stop" sign as "speed limit 45" causing it to speed up into the busy road rather than slow down [33]. It made this mistake because there were four small rectangular stickers stuck to the face of the sign. Such a mistake seems unlikely for human drivers. And even if a human driver were to make the same mistake, the reason would be different. A more serious example can be seen in the case of Elaine Herzberg, who was the first pedestrian to be killed by a driverless car. As Elaine was pushing a bicycle across the road, it is likely that the emergency

brake system was not applied because the car's sensors were unable to distinguish between Elaine and her metal bicycle [34]. Thus, whilst AI systems perform better than humans in particular aspects, they can also make some incomprehensible low-level mistakes.

In medical practices, this has the potential to put patients at risk of serious harm. Here, we try to show an extreme case with a thought experiment. Suppose there are four diseases: A, B, C, and D. Disease A is triggered by short-term overwork and the symptoms will disappear after a few days of rest. Disease B is followed by severe muscle pain which requires taking painkillers for three days and regular sleep during these days. The patient who is diagnosed with disease C needs to take antibiotics three times a day for one week. Compared to these diseases, disease D is much more severe, and requires the patient to have an emergency amputation. One day, Robert experiences acute discomfort and decides to visit his doctor. From Robert's symptoms, all four diseases are possibilities on the doctor's list. After the check-up, the doctor firstly excludes disease C and D from his list because he knows that there is no causal relationship between these and Robert's symptoms. However, the doctor is unsure whether it is disease A or disease B. Based on their experience and intuition, the doctor provides the final diagnosis and suggests Robert has a couple of days of rest. Although the doctor cannot clearly explain why Robert has disease A rather than B, based on their experience, the human doctor is able to make the correct judgement. They have 90 percent accuracy in disease A's diagnosis. Suppose Kevin has the same symptoms and visits his "AI doctor", a medical AI system. After checking Kevin's symptoms, the AI doctor also flags the same four diseases (disease A, B, C, and D). Then, combing Kevin's health records, the AI doctor firstly removes disease B and disease C from the list. Lastly, it makes the final choice between the two remaining options, and determines that Kevin has disease A. The AI doctor has higher predictive accuracy than the human doctor in disease A, let's say, 99 percent.

This thought experiment fits the claims that medical AI systems outperform human doctors in disease detection and the diagnoses produced by human doctors are not always explainable. If the thought experiment can be imagined, then the harm caused by medical AI's misdiagnoses might seem to be more serious than that caused by human doctors' misdiagnoses. In the thought experiment, if the human doctor makes the mistake to misdiagnose Robert with disease B, Robert has to take painkillers for three days and have regular sleep. By contrast, if the AI doctor makes a misdiagnosis between disease A and disease D, then Kevin would have an amputation. The harm caused by an unnecessary amputation would be much more serious than that caused by the misuse of prescription painkillers. Although this is a thought experiment and not a real-world occurrence, it illustrates the potential harm which may be caused by black boxes being applied in medicine before they are explainable enough. It is the opaque working process within the medical AI systems that contributes to the serious harm caused by misdiagnoses.

4. New medical paternalism

The potential harm to patients from medical AI systems may also stem from the fact that the patient's right to be fully informed would not be effectively guaranteed due to its unexplainability. Medical AI systems produce diagnoses and treatment suggestions without adequate explanation that may aid the patient in making medical decisions. Such a model returns patients to a paternalistic model of the physician-patient relationship, in which physicians make medical decisions for their patients without disclosing information which is relevant for decision making. The difference is that in the era of artificial intelligence, AI systems critically contribute to paternalistic decisions.

Several regulatory approaches are currently being used for AI clinical decision support (CDS) systems in the US and European countries, including:

(1) Pre-Market Review. FDA requires pre-market review of certain AI CDS systems that meet the definition of a medical device. This review process evaluates the safety and effectiveness of the system, and can result in a clearance or approval for marketing [35]. Similarly, in the EU, AI CDS systems are subject to a conformity assessment, which is a process used to verify that the system meets the necessary safety and performance requirements [36]. The process includes evaluating the design and manufacturing processes, as well as the documentation and labeling of the system. However, the process can't guarantee 100 percent safety and may not always keep up with the rapidly evolving technology landscape, leading to delays in getting innovative products to market.

(2) Post-Market Surveillance. The FDA also requires post-market surveillance of medical devices, including AI CDS systems, to monitor their safety and effectiveness in real-world use [37]. This can include monitoring adverse events and taking action if safety issues arise. In the EU, in addition to be monitored, AI CDS systems are also subject to clinical evaluation to assess their safety and effectiveness in the intended use [38]. This includes evaluating the clinical data supporting the use of the device, as well as any potential risks associated with its use. However, the current post-market surveillance can be resource-intensive and may not always catch all safety issues, particularly if they are rare or difficult to detect. One strategy for improving post-marketing surveillance is to enhance reporting requirements for adverse events associated with medical devices. However, the risk of misdiagnosis still exists due to the black box feature of medical AI.

Generally, the application of medical AI will create two diagnostic models: the AI-patient model and the AI-physician-patient model. In the former, medical AI systems replace human doctors to directly provide diagnoses and treatment recommendations to patients. This requires AI systems to independently complete relevant tasks, which seem to raise more technical and ethical challenges than the AI-physician-patient model. In terms of unexplainability, it is more severe than the AI-physician-patient model, so the application of this model seems impossible until the black-box problem in the AI-physician-patient model is well resolved. Although some smart wearable devices have been applied to monitor users' physiological indices, and will alert the wearer once certain indices are no longer within the normal range, these devices often only collect and monitor users' data, and do not have the authority or ability to make the independent diagnosis and provide treatment plans for users. In fact, in order to avoid getting into legal or other disputes, the designers of these devices often remind users that the test reports are for reference only and to consult a professional doctor if required.

In the AI-human physician-patient model, physicians and AI systems work together to provide diagnosis and treatment recommendations to patients. In such a model, medical AI plays an assistant role to contribute to physicians' diagnoses and treatment suggestions. For example, the Sepsis Watch, a medical AI system, was designed to rapidly identify patients who need treatment for sepsis [39]. Sepsis has no standardized diagnosis criteria, which makes it easy to misdiagnose or miss, so that doctors have to spend a lot of time and effort to monitor patients for sepsis risk. The Sepsis Watch is a good solution to the problem of monitoring sepsis, saving doctors' time and effort. Patients monitored by the Sepsis Watch have their risk index evaluated every hour, and when the risk value exceeds 60%, the system alerts the doctor. After receiving the alarm, the doctor comes to diagnose whether the patient needs immediate medical treatment [39]. The model works in this case because, as it makes explicit, the Sepsis Watch is not a diagnostic device. The function of the Sepsis Watch is simply to "identify patients for further evaluation so that the attending physician caring for an individual patient can make the final diagnostic determination to start treatment for sepsis" [39]. Such cooperation between medical AI systems and physicians really increases efficiency and reduces the burden on doctors.

However, this model does not seem to bring out the full advantages of medical AI. As mentioned above, medical AI can gain the ability to diagnose diseases and then provide relevant treatment suggestions through self-learning of massive amounts of data. Furthermore, they show better than human doctors in accuracy. Thus, replacing or partially replacing human doctors with medical AI systems to make diag-

noses for patients, which may improve diagnosis accuracy seeming to be what we were expecting more. However, the unexplainability feature of medical AI limits such a model in which medical AI systems replace or partially replace human doctors to undertake diagnostic work. In this model, physicians may also be required to play the role to check the accuracy and safety of the results provided by medical AI systems, and communicates this to the patients. However, to some extent, medical AI systems are "super experts" in comparison to human doctors. Then, how can doctors check and confirm these super experts' conclusions? One response would be that the high accuracy of medical AI systems is enough to guarantee safety, and that human doctors do not need to check it again. Whether the safety can be guaranteed, as it suggests, has been discussed in the previous section.

Here we will focus on how to communicate with patients based on the results produced by medical AI systems. We try to use a hypothetical case to show that medical AI systems should be more explainable to make it possible for an individual to contest the decision of the system. Suppose Julia, who is ten weeks pregnant, is diagnosed with breast cancer by a Medical AI system. Although Julia's attending doctor does not draw the same conclusion after reading her mammogram, he accepts the diagnosis and then informs Julia. In terms of treatment options, suppose the AI system lists three plans: A, B and C, and ranks them in terms of success rates from A to C. The doctor then highly recommends plan A and explains the possible risks involved in as much detail as possible. If Julia decides to receive the treatment, she will have to terminate the pregnancy because each of these treatment plans will injure the fetus. Julia wants to delay the treatment until after the birth, if the risk is under control. And in terms of treatment, she prefers plan C. Hence, Julia may want to know the risk of treatment delay and the difference between plans A and C in terms of potential risk and benefit. However, the doctor cannot seem to meet such a requirement since the information that Julia wants is determined by factors like the number and size of tumors, the tumor grade, which indicates how quickly cancer can be expected to grow and spread, and the specific type of breast cancer. All of these are unknown to the doctor. As we can imagine, all the doctor could do is to emphasize the accuracy of the result and then persuade Julia to accept Plan A.

Julia's case does not deny the high accuracy of the diagnosis produced by medical AI systems. Rather, the case indicates that the unexplainable results would limit the patient's autonomy, which creates a new type of medical paternalism, as mentioned above. Regardless of the diagnosis or the treatment suggestion, what is produced by medical AI systems is data-based rather than individual patient-centered. It is the medical AI systems that determine what is best for the patient.

In addition, the unexplainability feature of medical AI increases the patient's mistrust and suspicion [11]. From the patient's perspective, without access to adequate information, they are not sure whether they should trust the diagnosis and treatment recommendations given by medical AI systems; nor do they know who will be responsible for these diagnoses [40]. These can add to the patient's psychological burden, causing anxiety and confusion. Further, as an emerging technology, the application of medical AI devices has raised the cost of healthcare. In some developing countries, these costs are not covered by health insurance, ultimately adding financial burdens to patients. For instance, in China, the cost of da Vinci surgical robot is expensive and not reimbursed by medical insurance in some cities [41]. Many hospitals are keen to recommend the da Vinci surgical robot to patients for profit and promotional purposes, which increases the financial burden on the patient.

In this article, based on the principle of "do no harm", we have argued that the unexplainability feature of medical AI systems put the patient at potential risk of harm from three aspects. Compared with existing literature, in this article, we used a thought experiment to show that the potential harm caused by medical AI's malpractice might be underestimated. Then, from an anti-medical paternalism perspective, we argued that the current medical AI systems need to be further explainable to ensure that patients' right to informed consent will be effectively protected. In addition, we showed the psychological and financial burdens that may be caused by the unexplainability feature of medical AI systems, which seems to be ignored by the previous ethical discussions.

Lastly, this article does not hold a pessimistic perspective of the application of medical AI, but rather a reminder that the public needs to be aware of the possible risks caused by the black box feature. Actually, many things can be done by medical societies. For example, training programs can be offered to help doctors and medical students to identify areas where AI may be less effective. Medical societies can foster collaboration between doctors, researchers, and AI experts to ensure that AI tools are developed with a deep understanding of the clinical context and the needs of patients. In addition, some guidelines and standards for the use of medical AI tools should be worked out. These guidelines should address issues such as data privacy, transparency in algorithms, and quality control measures to ensure that AI tools are used ethically and effectively. All these are particularly important to address the black box problem of medical AI.

Conflicts of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This study was supported by the Young Scholars Program of the National Social Science Fund of China (Grant No. 22CZX019).

Author contributions

Hanhui Xu completed the first draft of paper and Kyle Michael James Shuttleworth polished it and strengthened some parts of arguments.

References

[1] Artificial Intelligence (AI) in Healthcare Market. Available from https://www.precedenceresearch.com/artificial-intelligence-in-healthcare-market/2023 (Accessed on 24 July 2023).
[2] Artificial Intelligence (AI) in Healthcare Market. Available from https://www.transparencymarketresearch.com/artificial-intelligence-in-healthcare-market.html/2022 (Accessed on 24 July 2023).
[3] Food & Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. Available from https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices/2022 (Accessed on 24 July 2023).
[4] Umbrello S, van de Poel I. Mapping value sensitive design onto AI for social good principles. AI Ethics 2021;1(3):283–96. doi:10.1007/s43681-021-00038-3.
[5] Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol 2019;28(2):73–81. doi:10.1080/13645706.2019.1575882.
[6] Knight W. The dark secret at the heart of AI. Available from https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/(Accessed on 20 January 2021).
[7] Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: the system causability scale (SCS): comparing human and machine explanations. Kunstliche Intell 2020;34(2):193–8. doi:10.1007/s13218-020-00636-z.
[8] Ploug T, Holm S. Right to contest AI diagnostics. Artificial intelligence in medicine editors. Cham: Springer International Publishing; 2022. doi:101007/978-3-030-64573-1_267.
[9] Ploug T, Holm S. The four dimensions of contestable AI diagnostics - a patient-centric approach to explainable AI. Artif Intell Med 2020;107:101901. doi:10.1016/j.artmed.2020.101901.
[10] Smith H, Fotheringham K. Artificial intelligence in clinical decision-making: rethinking liability. Med Law Int 2020;20(2):131–54. doi:10.1177/0968533220945766.
[11] von Eschenbach WJ. Transparency and the black box problem: why we do not trust AI. Philos Technol 2021;34(4):1607–22. doi:10.1007/s13347-021-00477-0.
[12] Lauritsen SM, Kristensen M, Olsen MV, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun 2020;11(1):3852. doi:10.1038/s41467-020-17431-x.
[13] Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3(11):e745-e50. doi:10.1016/s2589-7500(21)00208-9.
[14] McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics 2019;45(3):156–60. doi:10.1136/medethics-2018-105118.

[15] Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. Philos Technol 2021;34(2):349–71. doi:10.1007/s13347-019-00391-6.
[16] Martin PM, Martin-Granel E. 2,500-year evolution of the term epidemic. Emerg Infect Dis 2006;12(6):976–80. doi:10.3201/eid1206.051263.
[17] Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol 2020;21(2):233–41. doi:10.1016/s1470-2045(19)30739-9.
[18] Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. N Engl J Med 2017;376(26):2507–9. doi:10.1056/NEJMp1702071.
[19] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318(22):2199–210. doi:10.1001/jama.2017.14585.
[20] Castelvecchi D. Can we open the black box of AI? Nature 2016;538(7623):20–3. doi:10.1038/538020a.
[21] Miotto R, Li L, Dudley JT. Deep learning to predict patient future diseases from the electronic health records. In: Proceedings of 38th European Conference on Information Retrieval Research, Padua; Italy. ECIR; 2016. doi:10.1007/978-3-319-30671-1_66.
[22] McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6.
[23] Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886. doi:10.1136/bmj.l886.
[24] Allbutt C. Hippocrates - Hippocrates. With English Translation by W. H. S. Jones; 1923. doi:101017/S0009840X00040361.
[25] Buchanan A. Medical paternalism. Philos Public Aff 1978;7(4):370–90.
[26] Elle Hunt. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. Available from https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter/(Accessed on 24 July 2023).
[27] Noiret S, Lumetzberger J, Kampel M. Bias and fairness in computer vision applications of the criminal justice system. In: Proceedings of 2021 IEEE Symposium Series on Computational Intelligence (SSCI); 2021. doi:10.1109/SSCI50451.2021.9660177.
[28] Phil McCausland. Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk. Available from https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281/ (Accessed on 24 July 2023).
[29] Facebook Apologizes After A.I. Puts 'primates' label on video of black men. Available from https://www.nytimes.com/2021/09/03/technology/facebook-ai-raceprimates.html#:~:textApologizes20After20A.I.-.Puts20Primates20Label20on20Video20of20Black20Men.other20issues20related20to20race/ (Accessed on 24 July 2023).
[30] Melissa Heikkilä. Dutch scandal serves as a warning for Europe over risks of using algorithms. Available from https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/(Accessed 24 July 2023).
[31] London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 2019;49(1):15–21. doi:10.1002/hast.973.
[32] Pande V. Artificial intelligence's 'black box' is nothing to fear. Available from https://www.nytimes.com/2018/01/25/opinion/artificial-intelligence-black-box.html/2018 (Accessed on 20 January 2021).
[33] Heaven D. Why deep-learning AIs are so easy to fool. Nature 2019;574(7777):163–6. doi:10.1038/d41586-019-03013-5.
[34] Efrati A. Uber finds deadly accident likely caused by software set to ignore objects on road. Available from https://www.theinformation.com/articles/uber-finds-deadly-accident-likely-caused-by-software-set-to-ignore-objects-on-road/2018 (Accessed on 17 January 2022).
[35] Clark P, Kim J, Aphinyanaphongs Y. Marketing and US Food and Drug Administration clearance of artificial intelligence and machine learning enabled software in and as medical devices: a systematic review. JAMA Netw Open 2023;6(7):e2321792. doi:10.1001/jamanetworkopen.2023.21792.
[36] Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artificial intelligence in healthcare. Academic Press; Elsevier; 2020. doi:101016/B978-0-12-818438-700012-5.
[37] Harvey HB, Gowda V. How the FDA regulates AI. Acad Radiol 2020;27(1):58–61. doi:10.1016/j.acra.2019.09.017.
[38] Niemiec E. Will the EU medical device regulation help to improve the safety and performance of medical AI devices? Digit Health 2022;8:20552076221089079. doi:10.1177/20552076221089079.
[39] Sendak M, Elish M, Gao M, et al. The human body is a black box": supporting clinical decision-making with deep learning. In: Proceedings of the 2020 conference on fairness, accountability, and transparency; 2020. doi:10.48550/arXiv.1911.08089.
[40] Price WN. Medical malpractice and black-box medicine. Cambridge University Press; 2017. doi:101017/9781108147972027.
[41] Liu Z, Wang J, Wang X. Ethical issues and countermeasures of Da Vinci robot surgery in urology. Med Philos 2021;42(13):31 24-7.