Acta Bioethica 2025; 31(1): 25-34. DOI: http://dx.doi.org/10.4067/S1726-569X2025000100024

THE ETHICAL RISKS AND SOLUTIONS OF THE ALGORITHM BLACK BOX IN MEDICAL ARTIFICIAL INTELLIGENCE

Hu Qinggui[1], Tang Xiuqiong[2], Chen Hui[3], Wang Jinsong[4]

Abstract: The application of artificial intelligence (AI) in the medical industry is becoming increasingly widespread. Relying on its powerful machine-learning capabilities, it has gradually become an important auxiliary diagnostic device. At the same time, it gradually has a certain degree of autonomy. But this also leads to the problem of lack of transparency in algorithms. A critical ethical issue known as the "algorithmic black-box" problem has emerged. For the ethical challenges associated with the opacity of medical artificial intelligence algorithms, enabling artificial intelligence systems to have reasoning and moral judgment abilities, namely constructing medical artificial moral agents, has been proposed as a viable solution. This article analyzes the solutions of constructing medical artificial moral agents including "top-down" and "bottom-up" approaches. After that, the new hybrid ethical design approach is proposed, integrating the advantages of both top-down and bottom-up approaches.

Keywords: medical artificial intelligence, algorithmic black-box, top-down approach, bottom-up approach, hybrid approach

Los riesgos éticos y las soluciones de la caja negra algorítmica en la inteligencia artificial médica

Resumen: La aplicación de la inteligencia artificial (IA) en la industria médica se está extendiendo cada vez más. Basándose en sus potentes capacidades de aprendizaje automático, se ha convertido gradualmente en un importante dispositivo auxiliar de diagnóstico y, al mismo tiempo, de forma gradual, con cierto grado de autonomía. Pero esto también conduce al problema de la falta de transparencia en los algoritmos y, debido a esta situación, ha surgido un cuestionamiento ético crítico, conocido como el problema de la "caja negra algorítmica". Para los desafíos éticos asociados con la opacidad de los algoritmos de inteligencia artificial médica se ha propuesto como solución viable permitir que los sistemas de inteligencia artificial tengan capacidades de razonamiento y juicio moral, es decir, construir agentes morales artificiales médicos. Este artículo analiza las soluciones para construirlos, incluidos los enfoques "de arriba hacia abajo" y "de abajo hacia arriba". Después de eso se propone el nuevo enfoque de diseño ético híbrido, que integra las ventajas de ambos enfoques.

Palabras clave: inteligencia artificial médica, caja negra algorítmica, enfoque de arriba hacia abajo, enfoque de abajo hacia arriba, enfoque híbrido

Os riscos e soluções éticas da caixa preta algorítmica em inteligência artificial médica

Resumo: A aplicação de inteligência artificial (IA) na indústria médica está se tornando amplamente difundida. Contando com poderosos recursos de aprendizado de máquina, ela gradualmente ase tornou um importante dispositivo auxiliar de diagnóstico. Ao mesmo tempo, ela gradualmente tem um certo grau de autonomia. Mas isso também leva ao problema de uma falta de transparência em algoritmos. Um aspecto ético crítico conhecido como o problema da "caixa preta algorítmica" emergiu. Para os desafios éticos associados com a opacidade dos algoritmos de inteligência artificial médica, permitir que os sistemas de inteligência artificial tenham capacidade de raciocínio e julgamento moral, nomeadamente a construção de agentes morais artificiais médicos, foi proposto como uma solução viável. Esse artigo analisa as soluções de construção de agentes morais artificiais médicos, incluindo abordagens "de cima para baixo" e "de baixo para cima". Depois disso, a nova abordagem de planejamento ético híbrido é proposta, integrando as vantagens de ambas abordagens de cima para baixo e de baixo para cima.

Palavras chave: inteligência artificial médica, caixa preta algorítmica, abordagem de cima para baixo, abordagem de baixo para cima, abordagem híbrida

[1] Information Center, Sichuan Neijiang Normal University, Neijiang, Sichuan, 641100, China.
Corresponding author: hu646100178@126.com

[2] Nursing Department, Sichuan Neijiang Health Vocational and Technical College, Neijiang, Sichuan, 641100, China.

[3] School of Marxism Sichuan, Neijiang Normal University, Neijiang, Sichuan, 641100, China.

[4] Department of Anesthesiology, Neijiang Dongxing People's Hospital, Neijiang, Sichuan, 641100, China.

Introduction

The rapid development of technological progress has had a profound impact on society in many aspects, including health area*(1,2)*. Today, artificial intelligence (AI) technology is gradually becoming an important force in the medical development field*(3)*. At the same time, the ethical issues of medical artificial intelligence are becoming increasingly prominent. Although the prospects provided by artificial intelligence in the medical field are very promising, we need to control its use to avoid potential dangerous drift in this sensitive field*(4-6)*.

To find the key ethical concerns about medical artificial intelligence, the author searched for the topic of "medical artistic intelligence ethics" in the Web of Science database, and he obtained over 500 relevant literature (1975-2023). Then, a keyword co-occurrence map was drawn using VOS viewer 1.6.18 software (Figure 1). As shown in the figure, it indicated that issues such as privacy, autonomy, and responsibility have become key topics in medical artificial intelligence ethics research. The transparency issue of algorithms is an important node connecting thematic clusters such as "ethical principles" (yellow), "algorithm models" (red), and "medical practices" (blue). This indicates the algorithm transparency has become a key ethical question in the field of medical artificial intelligence

This article focuses on the algorithm transparency issue of medical artificial intelligence. The meanings and characteristics of the algorithm black box are introduced first. Later, both the ethical challenges and the solutions are analyzed. Then, combining with the current global ethical research trends of moral pluralism and monism, referring to specific cases in medical artificial intelligence research and application in recent years, it proposes a hybrid ethical design approach to address the algorithm black box problem.

The Algorithm Black Box and Ethical Challenges of Medical Artificial Intelligence

The application of artificial intelligence in the medical field is developing rapidly, AI technology relies on its powerful machine learning capabili-

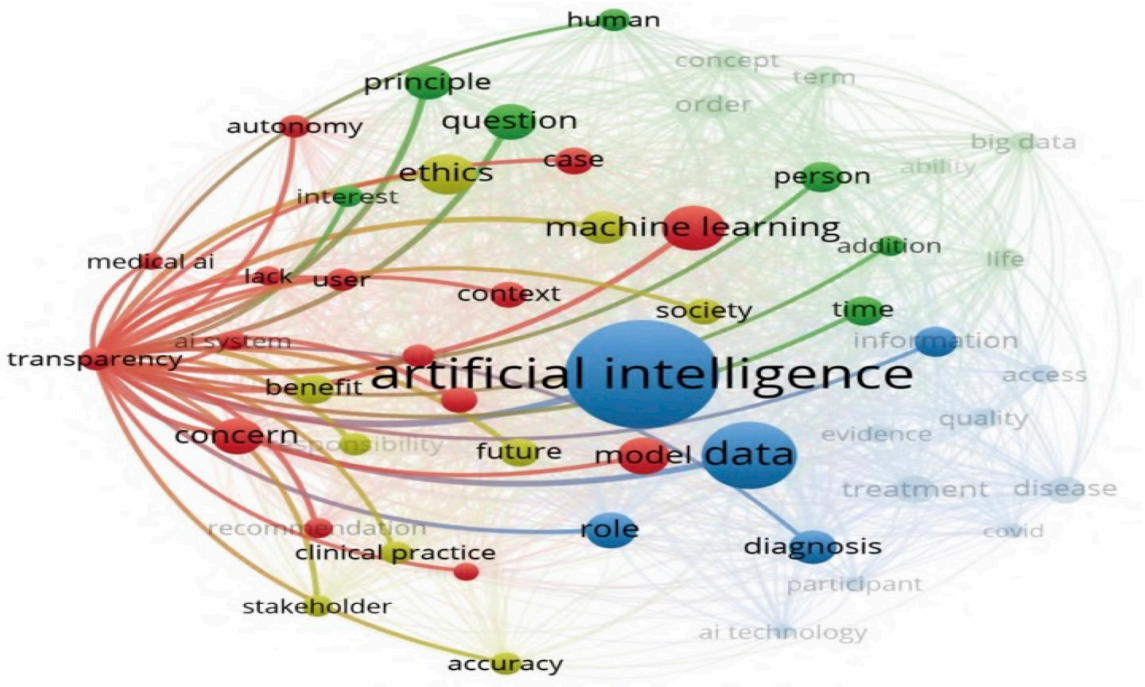

Figure 1: Literature keyword co-occurrence map on topic of "medical artificial intelligence ethics". Web of Science database/VOSviewer

ties to have a certain degree of autonomy, but this also leads to the problem of lack of transparency in algorithms. Specifically, after extensive training, the internal state of the model becomes quite complex, and the operations between input and output are automatic. This makes it difficult for people to accurately predict the behavior of the algorithm and understand the mechanism. Therefore, people refer to this phenomenon as the "algorithmic black box" problem*(7,8)*.

For ordinary human doctors, modern medicine emphasizes experience and evidence. The examination and diagnosis process in diagnosis and treatment activities are a set of empirical deductions based on the causal relationship between phenomena and results. The diagnostic and treatment measures taken are also based on long-term repeated clinical experience summaries.

In contrast, for medical Artificial Intelligence, the deep learning algorithms used in diagnostic and therapeutic are essentially a set of statistical mathematical models. The input and output layers are more based on a certain probability correlation rather than causal relationships. Algorithms can extract certain correlations by processing massive amounts of data to provide treatment plans.

This means that artificial intelligence can provide accurate diagnostic and therapeutic judgments, but cannot explain how these judgments are made. In other words, the decision-making process is difficult to understand. Medical personnel can only verify these judgments based on observational data rather than clinical trials. This opacity of interpretability is not intentionally created by humans, it is an inherent attribute of the algorithm's technical logic. The algorithm is a complex architecture when processing massive amounts of data.

The internal state of the algorithm becomes quite complex, and the operations between input and output are automatic. This also makes it difficult for people to understand the mechanism. This leads to black box issues.

The algorithm black box brings a lot of confusion, people always hope for the transparency of the algorithm.

At present, in the field of machine learning, to solve the algorithm black box problem, some researchers have developed several interpretable tools to improve the transparency and interpretability of algorithms. For example, local interpretable model agnostic interpretations (LIME) technology can help humans understand the classification criteria in image recognition models; Shapley value can be used to describe the contribution of each feature value to the model prediction results, and thus to improve the interpretability of the algorithm.

However, in the field of medical artificial intelligence, interpretable models are rarely used. There are many reasons for the problems. On the one hand, the medical profession itself is profound, and on the other hand, artificial intelligence algorithms are also profound. For Medical Artificial Intelligence, comprehensive knowledge and skills are required develop the local interpretable model agnostic interpretations (LIME) technology. This increases the difficulty to develop such technologies. From the actual effect perspective, a lot of such models are unable to provide truly satisfactory explanations in clinical practice*(9,10)*.

There are many "black box" issues in medical care. Medicine itself is full of unknowability, and we have accepted various black boxes in medicine. For example, electroconvulsive therapy is very effective for severe depression, but we do not know how it works; Many drugs seem to be very effective, but no one can provide an appropriate explanation*(11,12)*. Moreover, the judgments of human doctors are not always interpretable. In many cases, the diagnosis of human doctors is based on experience, intuition, and even speculation, rather than understanding the mechanisms of the disease. In other words, sometimes, human doctors themselves do not know the mechanism when diagnosing.

The opacity could bring about danger, for the risks brought about by the algorithmic black-box, we can divide them into the following categories, as shown in Tab.1.

Table 1. Ethical matrix analysis of algorithmic black-box of medical AI

moral bodies	**Data ethics**	**algorithm ethics**	**social ethics**
Internal hierarchy (algorithm expert)	data abuse	intellectual property	discrimination and bias
Internal and external interaction levels (medical experts)	data misuse issues	autonomy issues	responsibility attribution issues
External level (patient, public)	privacy violations	security issues	information cocoon issues

The moral responsibility persons include three groups: Internal level (algorithmic experts), Internal and external interaction level (medical experts), and External level (patient, public). In the following, we discuss their ethical risks from the perspectives of data, algorithms, and society.

In terms of data dimension, the performance of machine learning is highly dependent on the quantity and quality of the dataset. There are differences in the understanding and utilization of data among different groups. This leads to ethical issues, and affects the accuracy and safety of artificial intelligence's medical decision-making.

For example, the medical artificial intelligence companies use the full information patient data, it could cause data abuse. The patient´s privacy rights are violated.

On the issue of data sharing, many companies are unwilling to share their key data, which affects the accuracy and safety of artificial intelligence's medical decision-making. As of 2022, the US Food and Drug Administration has approved over 200 machine-learning algorithms for clinical practice. However, most of these algorithms lack sufficient data validation to evaluate model performance*(13,14)*. This may lead to data misuse issues characterized by "garbage in, garbage out".

In terms of algorithms, programs developed by experts should be protected by intellectual property rights, and some algorithms may even be classified as confidential for security reasons. This conflicts with transparency. In other words, it increases the opacity. When patients are unable to understand the raw information of the data and model, they will question the safety of the product naturally.

The algorithm black box can also threaten the autonomy of doctors in diagnosis and treatment, that is, the medical artificial intelligence, which should be used as an auxiliary means, may lead to doctors overly relying on it. When doctors do not understand the mechanism of artificial intelligence algorithms, it is difficult for them to make modifications and adjustments to AI medical decisions. Then, doctors can only rely on the diagnosis and treatment decisions provided by algorithms. In fact, many doctors report that the IBM Watson diagnosis and treatment system often provides confusing medication recommendations, and even some treatment plans are quite dangerous for specific patients. However, doctors cannot "inquire" the algorithm´s black box about why it makes such unreasonable decisions*(15)*, and some inexperienced doctors may blindly accept the advice given by medical artificial intelligence and cause misdiagnosis.

In terms of the social dimension, it is possible for algorithmic companies to use black boxes to make the public unconsciously accept algorithmic control. However, due to differences in technical literacy between algorithmic experts and the public, an information cocoon is formed, making it even difficult to achieve a transparent algorithm. Due to the opacity of the algorithm black box, it is difficult for people to detect and correct the biases and discrimination that may be included in the model,

which in turn can pose a threat to the health of patients*(16)*. In addition, the joint participation of doctors and medical artificial intelligence in the medical decision-making process can make the issue of responsibility attribution particularly complex: who should be responsible for errors in diagnosis and treatment? The algorithm developer or the doctor? For this question, the opinions of the public and doctors are different. a survey targeting the American public shows that when medical artificial intelligence causes medical accidents, the public (66.0%) is more inclined to believe that the doctor is the main responsible party, while doctors (43.8%) are more inclined to believe that the provider of artificial intelligence products should bear the main responsibility*(17,18)*. However, with the development of artificial intelligence technology, it is still necessary for us to consider how to make artificial intelligence systems have moral judgment ability to cope with the social risks and ethical challenges.

The ethical design approaches

The algorithm black box problem has become the current key ethical challenge in the field of medical artificial intelligence. The development of artificial intelligence with autonomous moral judgment ability, that is, the construction of artificial moral agents (AMAs), has been looked at as a feasible solution to resolve ethical issues. It means that AMAs have basic moral reasoning and moral judgment abilities.

This plan attempts to avoid the ethical risks that may be brought about by algorithmic black boxes by incorporating the paradigm of ethical into the research and development process. It enables artificial intelligence systems to have the ability to mitigate ethical risks.

How to develop artificial intelligence with autonomous moral reasoning capabilities?

Currently, many scholars have put forward some approaches to achieve AMAs*(19,20)*, those solutions can be divided into two categories: Top-down approach; Bottom-up approach.

However, according to the author's viewpoint, both top-down and bottom-up approaches have too limitations. In order to resolve the ethical challenges caused by the black box, the hybrid ethical design approach that combines "top-down and bottom-up" should be more reasonable.

1. Limitations of the top-down and bottom-up approaches

The top-down approach refers to the design of medical artificial intelligence ethical agents based on specific ethical principles to achieve a transparent and interpretable medical artificial intelligence system. Many scholars and institutions have proposed various norms and initiatives regarding ethical design for medical artificial intelligence. For example, in the "Ethics and Governance of Artificial Intelligence in the Health Sector" guidelines released by the World Health Organization in 2021, "ensuring transparency, interpretability, and comprehensibility" is one of the basic ethical principles that medical artificial intelligence should follow, and requirements are made for the transparency of relevant information such as technical limitations, operational records, data properties, and algorithm models*(21)*. However, the effectiveness of this top-down approach is quite limited, and the problem is that these preset ethical principles sometimes are hard to respond appropriately to complex ethical situations.

Firstly, as an emerging discipline, medical artificial intelligence inherently lacks a certain degree of ethical consensus among experts. There is a divergence between moral individualism and holism among data scientists. The former believes that facts and values are independent and they should adopt a technically neutral stance, while the latter believes that facts and values are inseparable and they believe that the ethical risks of artificial intelligence technology should be controlled strictly.

On the other hand, among ethicists, it is also difficult to form a consensus on what ethical principles should be adopted. In situations where there are conflicts between different principles, the rationality of using which set of ethical frameworks to guide the design of medical artificial intelligence may be questioned. For example, the principle of transparency requires medical artificial intelligence to be subject to scrutiny, but the principle of privacy protection requires it to keep data in-

formation confidential. Thus, the ethical dilemma caused by the top-down approach will bring more challenges to designers.

Secondly, there may be a contradiction between the ethical principles and practical goals of medical artificial intelligence, and achieving transparency often requires sacrificing a certain degree of accuracy. For example, during surgery, anesthesiologists need to monitor many physiological indicators to adjust the depth of anesthesia, and these physiological indicators often have linear relationships. Based on this fact, some scholars have developed a machine-learning algorithm that uses gradient descent to build a regression model to achieve automatic regulation of anesthesia. This is a highly understandable algorithm; However, in clinical practice, this algorithm cannot provide the best dose recommendation.

While, other algorithms based on neural networks perform better, although they have lower transparency. Do we need accuracy and security, or comprehensibility and transparency? The top-down approach does not provide a suitable answer.

Finally, when we face an environment with multiple moral standards, the ethical principles are difficult to provide us with specific guidance. When the designers design a logically monotonous moral reasoning algorithm for medical artificial intelligence based on the principle of transparency, the first problem is how this algorithm to matches the diverse persons, the diverse moral standards in different communities and cultures. For example, manufacturers, doctors, and ordinary users have different purposes and different cognitive abilities. On the other hand, the ethical principles are often too abstract and lack a certain degree of flexibility, they are hard to implement in the design of medical artificial intelligence.

The top-down approach make it difficult to fully consider the comprehensibility standards of differences, so this approach still cannot resolve the ethical risks of algorithmic black boxes.

To achieve AMAs, some scholars support other approaches, which is called the "bottom-up approach". This approach does not require engineers to follow a set of established ethical principles to design artificial intelligence, but rather allows the artificial intelligence to autonomously evolve a set of operating methods that conform to human moral standards in a series of reinforcement learning scenarios based on specific cases. In other words, it is to enable artificial intelligence to evolve a set of ethical systems that are in line with human standards through autonomous learning.

However, according to the author's viewpoint, relying solely on this approach cannot solve the ethical issues brought about by the algorithmic black box.

Firstly, the bottom-up approach, as a means of regulation and adjustment after the fact, needs to constantly learn and evolve. During its trial and error process, the negative moral consequences generated cannot be avoided.

Secondly, there are differences in moral standards, views, and behaviors among different individuals. The artificial intelligence machine's imitation of human moral behaviors is difficult to form a unified moral reasoning framework. For example, diagnostic and therapeutic robots may not be able to recognize the concealment and deception of the medical history of patients with mental disorders in their speech during training.

Thirdly, the bottom-up approach may allow medical artificial intelligence to learn some behavior patterns that violate moral standards due to the lack of guidance from ethical principles. Human beings do not know what behaviors artificial intelligence will evolve through learning, this actually increases the opacity of algorithms.

Therefore, the bottom-up approach cannot successfully build a medical artificial intelligence moral body that meets our requirements, and the ethical problems caused by the algorithm black box still cannot be resolved.

In summary, both top-down and bottom-up approaches have significant limitations.

2. Prospects of the hybrid ethical design approach

According to the author's viewpoint, a "hybrid approach" that combines both top-down and bottom-up can better cope with medical artificial

intelligence algorithms ethical challenges brought by black boxes.

The hybrid ethical design approach requires engineers to set a certain elastic ethical framework for medical artificial intelligence through a top-down approach, and embed more ethical moral requirement design into the entire process of medical artificial intelligence research and development*(22,23)*, the ethical framework only contains the most basic ethical principles. At the same time, it should consider the different moral environments in the process of learning evolution.

It also adopts a bottom-up approach, allowing the algorithm to learn human moral behavior patterns, giving full play to the advantages of medical artificial intelligence in processing multi-situational information, let it develop multiple moral reasoning models. What's more, the content of the ethical framework can be adjusted appropriately based on the learning and evolutionary process.

At the same time, we should note that in the ethical design of artificial intelligence, the following two mandatory requirements need to be added.

Firstly, some important algorithm source codes should be mandatory to open under certain conditions. We know, at present, the source code of the Windows system has always been confidential, and it is difficult for other experts to repair and remedy Windows system. Similarly, in the future, highly artificial intelligent systems will appear, if the algorithm source codes are in secret, it would be difficult for other experts to fix the vulnerabilities. And it may be difficult for other experts to control the system. If such a situation occurs, it could bring bad things for human beings. On the other hand, to protect intellectual property rights, a public interest organization would be established to receive and store those source codes, only under some specific circumstances, could the source codes be used.

Secondly, the most basic ethical principles should be forcibly embedded in artificial intelligent systems. For example, "No harm" is the most basic moral principle for doctors. It should be forcibly embedded in AI systems. Doctors have a moral obligation not to cause unnecessary harm. It means doctors have a moral obligation not to cause both the unnecessary physical bodily injury and economic injury. For the artificial intelligence systems, it is the same. From a social perspective, let us suppose, for example, that powerful artificial intelligence systems have the ability of intentionally harming to the general population, at the same time, they are controlled by only several persons, if those persons control the powerful AI systems to do wrong things. It would be of ethical failure.

Early researchers still adopted a computational stance, they believed that artificial intelligence only had the mechanical ability to perform moral computation, and could not form true moral reasoning abilities.

With the continuous advancement of ethical research in artificial intelligence, the ethical design approach has achieved theoretical breakthroughs. For example, some scholars have elaborated on the specific process of embedding human values into artificial intelligence systems*(24,25)*. At the practical level, some scholars have gradually shifted from limited machine training in early laboratory environments to handling moral dilemmas in the real world. That is to say, artificial intelligence has had moral judgment abilities, and such abilities are becoming stronger and stronger.

However, some scholars think that artificial intelligence does not need moral reasoning ability or moral judgment ability. Artificial intelligence technology is only neutral, it is a matter of human beings about how to utilize it. They think what we need is a safer and accurate artificial intelligence, rather than an ethical artificial intelligence.

However, according to the author's opinion, in today's information society, the various value systems and cultures are spreading globally, the conflict and integration cannot be avoided. The ethical challenges faced by artificial intelligence are becoming increasingly complex. It requires us to reflect on the ethical issues of artificial intelligence. According to our viewpoint, due to the increasing power of artificial intelligence, it is reasonable to equip it with basic moral judgment abilities. The hybrid approach could achieve this goal.

In the field of life medicine, people have basic

moral requirements for artificial intelligence. The hybrid approach could meet this need. Firstly, contrary to traditional moral monism, moral pluralism supports that moral decision-making in medicine is complex and diverse, and different subjects have different expectations for the moral behavior of artificial intelligence in different situations. A hybrid approach can respond to moral differences in the real world better. secondly, moral contextualism opposes the absolutist stance on morality within the framework of normative ethics, emphasizing that the criteria for comprehensibility and transparency are strongly related to the situation in which people are located, and there is no single criterion. The hybrid approach provides a value alignment path for resolving algorithmic black box ethical problems, which can align the moral behavior of medical artificial intelligence with the value judgments of stakeholders. finally, this approach follows the reflective equilibrium method in moral philosophy. Based on the practical goals of medical artificial intelligence, it identifies the ethical challenges and continuously adjusts according to the actual situation on the basis of formulated moral principles. This enables medical artificial intelligence to respond to complex moral issues and better meet people's needs for safety and accuracy.

From the internal perspective of medical artificial intelligence algorithms, hybrid ethical design requires algorithm engineers to consider the transparency of the algorithm at the beginning of the design. On the one hand, by developing interpretability tools that are separated from the underlying machine learning model, we can avoid the ethical difficulties that may arise from algorithmic black boxes. For example, when algorithm engineers build a machine learning model for neuroimaging, they add a tool that supports prediction, complementary procedures of verification and interpretation, evaluate the impact of interference in the model, and indicate possible discrimination and bias contained in the black box of the algorithm. This can improve the transparency.

For deep neural networks that analyze medical images, methods to improve the interpretability of the algorithm include concept learning models, counterfactual explanations, internal network representations, etc.

On the other hand, the hybrid approach requires engineers to build algorithms with self-explanatory capabilities. This can help to open the black box.

In addition, we can adopt many information technologies to visualize patient data, providing medical artificial intelligence with a more intuitive human-computer interaction interface and improving the comprehensibility of algorithms. Thus, the transparency of both the algorithms and the decision-making processes can be improved.

From the external perspective of the medical artificial intelligence algorithms, the hybrid approach requires all stakeholders to participate in the algorithm design process to resolve the problem of the algorithm black box. At present, some medical artificial intelligence developers have integrated doctors and patients into the algorithm design process.

A team developed an algorithm to evaluate treatment options, it included patients and doctors in the research of the algorithm design. It solicits opinions on transparency and understandability; this can ensure the autonomy of the patients. Medical experts can not only assist algorithm engineers in handling labels and supervised learning, but also play a key role in model validation. For example, some researchers have used a moral reasoning neural network called "Delphi" to process the opinions and consensus of medical experts. This algorithm can adjust its comprehensibility on time based on the clinical practice of doctors. The open participatory research in the algorithm design stage can improve the algorithm transparency. It can help build a "humanistic" medical artificial intelligence moral system, so that medical artificial intelligence can protect human dignity and subjectivity, at the same time, it can enable developers to assume corresponding moral responsibilities.

In summary, adopting a hybrid ethical design approach can make medical artificial intelligence learn moral reasoning in real-life situations and develop moral models embedded in its own algorithms. It allows medical artificial intelligence to effectively deal with conflicts between different ethical principles, and adapt to moral needs in

practical applications. It can avoid the difficulties caused by the rigidity of the top-down approach. In addition, the hybrid approach advocates for the widespread participation of multiple stakeholders, which can help medical artificial intelligence better handle complex moral scenarios, and align its behavior in different cultures.

It should be pointed out that in terms of the current development status of artificial intelligence, the hybrid approaches can only achieve relatively limited comprehensibility, and the algorithm black box cannot be completely "solved "in the short term. In addition, there are still many controversies about whether artificial intelligence systems can possess consciousness and free will. Those controversies make it difficult for AI to obtain a complete moral subject status. Based on this, the "algorithmic gray box" design with local interpretability can achieve a good balance between accurate "black boxes" and transparent "white boxes", which meets the moral judgment ability demands of medical artificial intelligence. For example, some studies have extracted classification information from brain cancer images using convolutional neural networks, and then extracted feature information such as the location and size of brain cancer from medical history, combining, them to improve the interpretability of brain cancer diagnostic models.

In the process of developing medical artificial intelligence moral system, through this hybrid approach, it is possible to combine the advantages of both top-down and bottom-up approaches. It can respond better to more complex ethical challenges in medicine.

Conclusion

Artificial intelligence technology is leading the transformation in the field of healthcare, and the algorithmic black box has brought significant ethical challenges to the development of medical artificial intelligence. Resolving the ethical challenges of algorithmic black boxes and building a medical artificial intelligence ethical system not only requires the participation of algorithm engineers, companies, governments, doctors, patients, ethicists, and other parties to provide ethical principles with a certain consensus for the design of medical artificial intelligence but also requires setting functional ethical status for medical artificial intelligence, to ensure its behavior and value framework are in line with human comprehensibility and autonomy. With the hybrid approach, moral design can better respond to the current societal demands for moral pluralism and contextualism can allow artificial intelligence to conduct moral reasoning and adopt appropriate actions based on diverse value systems and specific moral situations, thus promoting the innovation and development of medical artificial intelligence.

Acknowledgements

We acknowledge the financial support of the (1) National Natural Science Foundation project of China (Grant No. 6137073); (2)Sichuan Provincial Department of Science and Technology Provincial Science and Technology Plan Soft Science Project (Grant No.2019JDR0018).

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 (1)National Natural Science Foundation project of China (6137073); (2) Sichuan Provincial Department of Science and Technology Provincial Science and Technology Plan Soft Science Project (2019JDR0018)

Author contributions

Chen Hui put forward the idea; Hu Qinggu completed the first draft of paper and Kyle Michael James; Tang Xiuqiong and Wang Jinsong polished it and strengthened some parts of arguments.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

References

1. Umbrello S, van de Poel I. Mapping value sensitive design onto AI for social good principles. *AI Ethics.* 2021; 1(3): 283–96. https://doi.org/10.1007/s43681-021-00038-3
2. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. *Minimally Invasive Therapy & Allied Technologies.* 2019; 28(2): 73–81. https://doi.org/10.1080/13645706.2019.1575882
3. Ploug T, Holm S. Right to contest AI diagnostics. Artificial intelligence in medicine editors. *Cham: Springer International Publishing*. 2022: 227-238. https://doi.org/10.1007/978-3-030-64573-1_267
4. Ploug T, Holm S. The four dimensions of contestable AI diagnostics – a patient-centric approach to explainable AI. *Artif Intell Med*. 2020; 107: 101901. https://doi.org/10.1016/j.artmed.2020.101901
5. Smith H, Fotheringham K. Artificial intelligence in clinical decision-making: rethinking liability. *Med Law Int*. 2020; 20(2): 131–54. https://doi.org/10.1177/0968533220945766
6. von Eschenbach WJ. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philos Technol.* 2021; 34(4): 1607–22. https://doi.org:10.1007/s13347-021-00477-0
7. Schuilenburg M, Peeters R. *The Algorithmic Society*. New York: Routledge, 2021.
8. Durán J M, Jongsma K R. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics*. 2021, https://doi.org/10.1136/medethics-2020-106820
9. Antoniadi A M, Du Y, Guendouz Y, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Appl Sci.* 2021, 11: 5088.
10. Forsberg E M. The Ethical Matrix—A Tool for Ethical Assessments of Biotechnology. *Global Bioethics: Problemi di Bioetica* 2004;17(1): 167–172.
11. Lauritsen SM, Kristensen M, Olsen MV, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications* 2020; 11(1): 3852. https://doi.org/10.1038/s41467-020-17431-x
12. 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. https://doi.org/10.1016/s2589-7500(21)00208-9
13. Ebrahimian S, Kalra M K, Agarwal S, et al. FDA-regulated AI algorithms: Trends, strengths, and gaps of validation studies. *Acad Radiol*. 2022 Apr; 29(4): 559–566.
14. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy Versus Explainability. *Hastings Cent Rep*. 2019; 49(1): 15–21. https://doi.org/10.1002/hast.973
15. Smith H. Clinical AI: Opacity, accountability, responsibility and liability. *AI and Society*, 2021; 36: 535–545.
16. Khullar D, Casalino L P, Qian Y, et al. Public vs physician views of liability for artificial intelligence in health care. *Journal of the American Medical Informatics Association*. 2021 Jul; 28(7): 1574–1577.
17. Chen A T, Zhang X Q. Discussion on the medical ethical responsibility problems caused by artificial intelligence assisted diagnosis and treatment. *Chin Med Ethics*. 2020, 33: 803–808. (in Chinese)
18. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence in driven healthcare. *Artificial intelligence in healthcare*. 2020 Jun; 295-336. https://doi.org/10.1016/B978-0-12-818438-7.00012-5
19. Wallach W, Allen C. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press; 2008.
20. Heaven D. Why deep-learning AIs are so easy to fool. *Nature.* 2019 Oct; 574(7777): 163–6. https://doi.org/10.1038/d41586-019-03013-5
21. World Health Organization. *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization, 2021.
22. Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell*. 2019; 1: 501–507.
23. Li R C, Muthu N, Hernandez-Boussard T, et al. Explainability in medical AI. In: Cohen T A, Patel V L, Shortliffe E H, eds. Intelligent Systems in Medicine and Health. Cham: *Springer* 2022 Nov; 252–253.
24. Teng Y, Wang G Y, Wang Y C. Ethics and Governance of General Models: Challenges and Countermeasures (in Chinese). *Bull Chin Acad Sci*. 2022; 37(9): 1290-1299.
25. van de Poel I. Embedding values in artificial intelligence (AI) systems. *Minds Mach*. 2020; 30(3): 385–409.
26. Daniel W. Tigard; Maximilian Braun; Svenja Breuer, et al. Toward best practices in embedded ethics: Suggestions for interdisciplinary technology development. *Robotics and Autonomous Systems* 2023 May; 167(2): 104467.

Received: April 10, 2024
Accepted: May 9, 2024