AI Ethics: Should you trust AI with your medical diagnosis?

Weerawut Rainmanee✉
Chulalongkorn University

(Received: 1st December 2024; Revised: 23rd February 2025; Accepted: 25th February 2025)

Abstract

The use of AI in the medical field leads to one main ethical question: Should you trust AI with your medical diagnosis? To answer this question, one must understand how AI decides, which faces two main problems: The black box and validation data. The first is a problem of transparency in which we have no idea how AI makes decisions. The second is a problem of how we can set the validation data for training AI. This paper aims to analyze both problems and clarify how AI decides in medical diagnoses. Then I will show that AI use for medical image processing is one of the models that doesn't face these problems since it can provide the evidence of diagnosis. Next, I will address the question of whether you should trust AI in medical diagnoses, and I will show that the answer depends on comparing the functions between humans and AI.

Keywords: AI ethics, AI in medical diagnosis, medical image processing, consequentialism

✉ weerawut420@gmail.com

Introduction

The use of AI is widespread in many fields: art, writing, mathematics, history, and medical diagnosis. Students use AI to answer homework assignments. AI can help you design logos for brands faster than human artists can. My focus here is on using AI in the medical field as it will directly affect our lives and well-being if its use is permitted. This paper includes two parts: First, I will clarify the problems of using AI: the black box and the validity of datasets. I will show that neither of these problems constitutes sufficient reason to reject the use of AI models in medical diagnosis. Second, I will attempt to answer the question: 'Should you trust AI with your medical diagnosis’?

There are two models of medical devices: the rules-based model and the AI-based model. The first model is based on rules or conditions that a clinician inputs into the machine, and it processes output accordingly. This model lacks deep learning capabilities. Examples of this type of system include the blood pressure monitoring system, which assists the clinician in measuring blood pressure (Anwar, 2023). The clinician trusts the results of this machine. When you go to the hospital for a health check-up, a nurse may ask you to use the blood pressure machine. Simply place your arm in the machine, wait a moment, and it will print out a small paper with your blood pressure results. I would say that most patients have no idea how the blood pressure machine works, but they still believe in the results. However, the nurse knows that certain conditions, such as fatigue, can affect blood pressure readings.[1] If she sees you looking tired, she may suggest that you sit down and wait until your blood pressure stabilizes before taking a reading. Therefore, the rule-based model still requires experts to use it properly and be aware of any potential errors. We trust the machine because we assume it has undergone rigorous testing and validation by experts or public health standards before being approved for hospital use. Additionally, we see that doctors often rely on its readings for diagnoses.

The rules-based model is predictable since the outcome comes from the conditions we set. If it is programmed to choose only black or white, it will not choose purple, blue, or any other color that it is not programmed to choose. I think there are two main problems with using a rules-based model: First is the problem of bias, which can be solved by reading and following the instructions (e.g., do not use the machine after you exercise). Second is the problem of error machines, which can be easily solved by calling an engineer to fix it. I think we have no problem using this model to support clinicians in diagnoses since it has been doing a great job until now, and there are not many ethical concerns about this model. The next one is the AI-based model, which has deep learning abilities and can make decisions based on what it learns by itself. There are many examples of these models, such as Coupled Plasmonic infrared sensors, which can detect neurodegenerative diseases such as Parkinson's and Alzheimer's. CT panda is a model that detects pancreatic cancer (Perrault & Clark, 2024). AI may be used as a planning and simulation tool to assist neurosurgeons in advanced planning, show optimal access paths, and mark high-risk structures for the removal of tumor tissue (Palm, 2023). These are intelligent models and beneficial for human well-being. However, using these models for medical diagnoses raises an ethical question: Should you trust AI with your medical diagnosis? If AI diagnoses you with cancer or a tumor in your lung, brain, or other organs, would you believe it? What is the difference between a diagnosis from a human doctor and an AI doctor? We know that a doctor must have evidence or reasons to support her diagnosis. For example, a doctor might say, ‘I found this black spot on her lung x-ray film, and she has lost a lot of

[1] I express my gratitude to Sipang Sreprasert, Registered Nurse at Si Prachan Hospital, for invaluable assistance in developing my understanding of medical tools.

weight and other related symptoms'. In contrast with the case of AI diagnosis, AI can provide diagnoses but it's not transparent how it decides. We call this a black box problem. Another problem is determining the validity of the dataset used to train AI. Both problems are the main reasons to reject using AI in medical diagnoses.

The problem of AI in the medical field

In the previous section I show that there are two main problems of using AI in medical diagnoses: The black box problem and the validation of data. Both problems are mainly used for rejecting AI in the medical field. In this section I will justify the claim that some models don't face these two problems.

The black box problem refers to the fact that AI users don't understand how AI algorithms work (Barkal et al., 2023). The problem is similar to how we don't fully understand the complex decision-making processes of animals, such as why a dog might bark at one person but not another. I would say that the animal decision-making process is also a black box from the human perspective. Similarly for AI, even if we know what knowledge or data we used to train it, we don't know how it makes a decision. We know the input and output, but we don't know how it processes this output. This can be a serious problem of using AI in medical diagnosis since, in general, to diagnose one must have some reasons or evidence to support the diagnosis statement; for example, "You have cancer" or "You have a tumor in your lung/brain" as shown in your MRI scan with your other related symptoms. The support diagnosis statement is necessary since it is a decision of how one should be cured and how one should live. A precise diagnosis is good for patients since they will change their lifestyle for well-being. On the other hand, wrong diagnoses might even cause death, and there are many cases in Thailand in which patients sue doctors for misdiagnosis.

One might think that if AI cannot explain its decision-making process, how can we trust its diagnoses? Is there any supporting diagnosis statement? I think that AI Image segmentation for medical diagnosis model doesn't face this problem. This model aids doctors in the analysis of medical images. The task includes liver and liver-tumor, brain and brain-tumor, lung and other organ segmentation (Wang, 2022). This model is quite successful, as it can provide the data used for decision-making. If it diagnoses that you have lung cancer, it also shows or locates the lung cancer in the film (Kavasidis et al., 2023). The epistemic question remains: Is only locating a tumor in the films enough to use it for diagnosis? Bertrand Russell said, *"Facts have to be discovered by observation, not by reasoning; when we successfully infer the future, we do so by means of principles which are not logically necessary, but are suggested by empirical data"* (Russell, 1945). I think this is similar to diagnosing some diseases like tumors and cancers. If a doctor sees a tumor via MRI scan, then it is evidence to support the statement 'you have a tumor.' Similarly, if a camera captures A killing B with a gun, then that is evidence to support the statement 'A killed B with a gun.' I don't deny that a diagnosis requires detailed examination of other symptoms and organs. My point here is not to justify the result of AI diagnosis but to justify that there is no black box problem with the AI image segmentation model. If it finds a tumor in MRI scans, then there is no need to find other reasons to prove how it finds it. Some research shows that this model can perform fast and more accurately in diagnosis (Wang, 2022). Again, I don't mean that we should prefer using AI in diagnosis independently from clinicians. I mean that locating the spot in the image is enough to support the doctor's diagnosis. The doctor might agree or disagree with AI, but I think the black box in this case is no longer a problem that might lead us to reject this AI model.

Another problem of using AI in medical diagnosis is the problem of valid datasets. How can we know that the data used to train AI is correct? I think this is a kind of technical problem that can be solved by each discipline expert. Engineers develop AI models while clinicians label the datasets. If each helps to develop the model, and its result is accurate to 99% and ready to use in hospitals all around the world, then it's not necessary to know how they train AI. If it works perfectly, then there is no problem in using it. However, two problems might arise: First, since we use our current human knowledge to train it, and our knowledge of curing or diagnosing some diseases might be incorrect at the present time, we might train AI with incorrect data. I think in this case, it's not an AI problem itself, but rather a problem of developing human knowledge. Now AI even helps develop knowledge; for example, it helps to understand complicated structures like proteins (Perrault & Clark, 2024) and some complicated disease structures like Parkinson's disease (Khachnaoui et al., 2020). Second, the medical data used to train AI might contain bias. Suppose that model X is trained on medical data exclusively from European patients. The question arises: is this model suitable for use with Asian patients as well? Generally, the doctor will always inquire about medication allergies and existing medical conditions because certain treatments or medications may pose risks to specific individuals. Similar to using AI in diagnosis, the experts who label the data and the clinicians who will utilize the machine must possess knowledge of specific conditions to use it appropriately with patients. The experts must determine whether the model is suitable for use with Europeans, Asians, or patients with specific medical conditions. This concern parallels the use of a blood pressure machine. The doctor must be aware of and understand the patient's medical condition before using it to prevent misdiagnosis.

In this section, I show that there are two problems of using AI in the medical field: the black box and the validity of datasets. I argue that the AI image model doesn’t face the first problem since locating a tumor in the film is enough to be the evidence for diagnosis. For the second one, I show that it’s only a technical problem which can be solved by experts in each discipline. I concluded that neither problem constitutes sufficient reason to reject using AI in the medical field. However, if we accept using AI in medical diagnosis, one problem might still arise: should you trust AI in medical diagnoses? For the next section I will try to answer this question.

Should you trust AI in medical diagnosis?

Some problems of using AI in the medical field are still future problems, such as: Can AI be used independently from doctors? Will AI replace doctors? Should we trust AI diagnoses and what to do if AI and doctors have different diagnoses? These are all future problems if humans succeed in developing strong AI. In this section, I will try to clarify one problem in using AI in medical diagnosis: 'Should you trust AI in medical diagnosis?' I believe this question is significant, as it is similar to the situation with COVID-19 vaccines. People around the world debated whether or not to take the COVID-19 vaccine, as they were unsure about vaccine information, such as which vaccine is better, whether vaccines can cause death, and how to know if they are not allergic to the vaccine. During that time, the Thai government forced foreigners and Thai people to get vaccinated before entering the country, requiring at least two doses. Even after receiving two doses, there was no guarantee of protection from the virus. We are still figuring out if there are any long-term side effects, facing this challenge together in human history, as no one can travel into the future to warn us about whether or not to get vaccinated. The problem in this situation is 'What should we do about getting vaccinated or not, and who should we trust: specialists, company researchers, or clinicians?' I think this kind of problem is most likely to occur in the future when using AI in medical diagnosis, as people might debate and be confused about using it.

We are surprised that AI can beat humans at chess. It's reasonable to choose AI to be your assistant if you play the game since it has a recorded winning history. In the case of contrast diagnoses between clinicians and AI, will you choose AI too? If we develop a test for reading films to diagnose brain disease, we can use it to evaluate both AI systems and human experts. There are two logically possible scenarios that will happen: First, AI demonstrates significantly lower diagnostic capacity for brain disease compared to human experts. In this case, rejecting its use is reasonable as it poses an unacceptable risk. This scenario does not present a serious concern. Second, AI exhibits higher diagnostic capacity than human experts. This indicates that the AI system achieved a higher score on the test. Let's consider a scenario where this AI system is used to assist a clinician. This clinician, who has consistently scored lower than the AI system in multiple tests, is now using the AI system in a real-life clinical situation. A patient visits the hospital for a medical check-up. The clinician is curious about whether her diagnosis would differ from the AI's or not. So, she reads the film and compares the results to the AI’s diagnosis. The AI diagnoses the patient as being in the early stages of Parkinson's disease, while the clinician disagrees.[2] In this case, it is reasonable to give more weight to the AI's diagnosis since it has a higher capacity than experts. However, it might be unreasonable to solely rely on the AI's diagnosis. This is because human experts lack the ability to verify the AI's results. This is a paradoxical situation: AI can potentially provide accurate diagnoses, but there's no reliable way for human experts to confirm or refute these diagnoses. Consequently, the results will be considered valid only if humans can prove them.

Suppose this scenario is true: AI can provide a valid diagnosis, but the expert cannot independently verify the validity of that result. What should the expert do? There are clearly two answers: First, the expert could choose to trust the AI's diagnosis and proceed with medical treatment because she believes that the AI has achieved a higher score than her in multiple tests, and thus that its diagnosis may be more reliable than hers. This reasoning appears to be an appeal to authority fallacy. An AI machine, considered an expert in Parkinson's disease (PD), indicates that the patient has PD. Therefore, I believe that the patient has PD. This constitutes a logical fallacy because the expert lacks the ground or evidence to support the AI's diagnosis. The only way to solve this problem is to find evidence or grounds to support the AI. As I stated earlier, the expert must understand how the diagnosis is validated, but this may be impossible since there's no reliable way for human experts to confirm or refute these diagnoses. If the clinician blindly trusts the AI and proceeds with treatment without a thorough understanding of the diagnosis, the patient may be harmed by the treatment. My point here is that even if AI can perform diagnoses better than experts, we will still face the challenge of proving its accuracy even if its diagnoses are already correct. Next, the second answer is that the expert does not believe in the AI's diagnosis and continues to believe that the patient does not have Parkinson's disease. It's logical to think that this might be harmful to the patient, as they may waste time and money on unnecessary treatment. One might argue that one way to solve this problem is to send the patient data to other experts for diagnosis. However, the problem lies in determining how many experts would be needed to reconfirm the diagnosis before it can be considered justified. In my view, answering the question of whether we should trust AI in medical diagnosis requires comparing the functions of humans and AI. If AI has lower capacity than humans, then we should not use or trust it. However, if AI has a higher capacity than humans, we should not trust it until we can find the grounds that support its

[2] Parkinson’s disease is one of the diseases that is hard to diagnose, and doctors sometimes misdiagnose it.

results, even if the AI can provide a valid answer initially. It is a better choice for clinicians to understand how the AI arrives at its diagnosis before providing medical treatment to the patient.

The future problem of deciding who to believe between humans and AI still requires new solutions. The serious problem now is that there is a shortage of clinicians in most countries, including Thailand. At some hospitals in Bangkok I have experienced that they schedule an appointment for me almost seven months later for only a liver X-ray. You can pay extra to get an appointment sooner, but then you must come for the X-ray after 5:00 PM. This must be the worst situation for both doctors and patients, as doctors have to work overtime and patients have to wait for a long time to get treatment for their diseases, which may reach a point where they cannot be cured. If AI can actually help doctors read MRI scans, CT scans, X-ray films, and other tasks faster, then there are many good reasons to continue developing the model to improve clinician well-being and save more human lives.

Conclusion

I began my paper by trying to clarify two problems: the black box problem and the validation of data problem. I claim that AI image segmentation models do not face either problem, because reading or detecting tumors in X-ray films, CT scans, or MRI scans does not require a reason to support the diagnosis. It needs only evidence of whether or not someone has a tumor, and clinicians may agree or disagree with the evidence that AI provides, but the transparency of how AI decides is not significant. The black box problem is no longer a reason to reject using this model. For the validation of data problem, I claim that it's only a technical problem which can be solved by experts. In the last section, I answer the question "Should we trust AI in medical diagnosis?" The answer depends on the capacities of humans and AI. If AI demonstrates low diagnostic capacity compared to human experts, it should not be used or trusted. Conversely, if AI demonstrates higher capacity than experts, then we should trust it if we have the grounds or evidence to support its results.

REFERENCES

Anwar, S. M. (2023). Expert Systems for Interpretable Decisions in the Clinical Domain. In *AI in Clinical Medicine* (pp. 66-72). https://doi.org/https://doi.org/10.1002/9781119790686.ch7

Barkal, J. L., Stockert, J. W., Ehrenfeld, J. M., Aunger, C. E., & Cohen, L. K. (2023). General Framework for Using AI in Clinical Practice. In *AI in Clinical Medicine* (pp. 13-26). https://doi.org/https://doi.org/10.1002/9781119790686.ch2

Kavasidis, I., Salanitri, F. P., Palazzo, S., & Spampinato, C. (2023). History of AI in Clinical Medicine. In *AI in Clinical Medicine* (pp. 39-48). https://doi.org/https://doi.org/10.1002/9781119790686.ch4

Khachnaoui, H., Mabrouk, R., & Khlifa, N. (2020). Machine learning and deep learning for clinical data and PET/SPECT imaging in Parkinson's disease: a review. *IET Image Processing*, *14*(16), 4013-4026. https://doi.org/https://doi.org/10.1049/iet-ipr.2020.1048

Palm, C. (2023). History, Core Concepts, and Role of AI in Clinical Medicine. In *AI in Clinical Medicine* (pp. 49-55). https://doi.org/https://doi.org/10.1002/9781119790686.ch5

Perrault, R., & Clark, J. (2024). Artificial Intelligence Index Report 2024.

Russell, B. (1945). A history of western philosophy. Simon and Schuster.

Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., & Nandi, A. K. (2022). Medical image segmentation using deep learning: A survey. *IET Image Processing*, *16*(5), 1243-1267. https://doi.org/https://doi.org/10.1049/ipr2.12419