Philosophy & Technology (2026) 39:1
https://doi.org/10.1007/s13347-025-01009-w

RESEARCH ARTICLE

Critical Engagement: The Value of Transparency of AI in Healthcare

James Edgar Lim[1] · **Owen Schaefer**[1] · **Julian Savulescu**[1,2]

Received: 26 November 2024 / Accepted: 28 November 2025

Abstract

Why is transparency important for the use of AI in healthcare? Responses to this question typically claim that transparency is something owed to the patient – because it is a condition for informed consent, legitimacy, accountability to the patient, etc. In this paper, we draw attention to why transparency can be valuable for medical practitioners. We claim that transparent AI models facilitate *critical engagement* by medical practitioners with AI models that they use. That is, they enable practitioners to assess why AI models make the recommendations they do, think about how those reasons affect their own beliefs and judgments, and make reasoned decisions about whether to maintain or change their own judgments. Via this process, AI models can help medical practitioners to improve their practice in a distinctly valuable way. In turn, this benefits both medical practitioners and their patients. This conclusion has important implications for AI design in healthcare: if AI models are to be used in healthcare, they should be designed in ways which allow medical practitioners to understand how the models arrive at their recommendations, and engage with them critically.

Keywords AI · AI ethics · Healthcare · Transparency · Interpretability · Explainability · XAI · Black-box · Opaque · Critical engagement

✉ Julian Savulescu
julian.savulesu@uehiro.ox.ac.uk

[1] Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[2] Uehiro Oxford Institute, University of Oxford, Oxford, United Kingdom of Great Britain and Northern Ireland

Published online: 16 December 2025

1 Introduction

Imagine you are at the hospital, seeking treatment for a serious illness. You consult the doctor, and she keys in your symptoms and your information into her computer. A few seconds later, she reveals that she's consulted an AI model which has recommended exploratory surgery. Curious (and perhaps a little concerned), you ask how the AI arrived at its conclusion. She replies that she has no idea how the AI works – in fact, the AI is a "Black-Box" model, and no one – including its designers – can explain fully how it works in the sense of how exactly inputs are causally related to the recommendations (Vainio-Pekka et al., 2023). She does know, however, that the model has a very high degree of accuracy, comparable to or better than the best doctors around. This unsettles you.

Many ethicists have claimed that something has gone wrong here. Specifically, they claimed that AI models must be transparent (or interpretable/explainable) for medical practitioners (among other people) to ethically use them (Cortese et al., 2023; Günther & Kasirzadeh, 2022; Rasheed et al., 2022). That is, the models must be designed such that someone can provide a causal explanation for how they translate inputs to outputs. Transparent AI models can be contrasted with opaque or black-box models, which cannot be understood similarly. There are a number of reasons ethicists have provided in favor of transparent AI. Among these reasons are that they minimize risk (Chesterman, 2021; Cortese et al., 2023), engender trust (Mitchell, 2025), and ensure fairness (Balasubramaniam et al., 2022; Chesterman, 2021). While these concerns are certainly serious, they can in principle be assuaged. While we might not know how black-box models work, we can test the reliability and fairness of a model simply by comparing a sufficiently large set of its recommendations (or outputs) against existing data.[1] Because we can prove an AI model to be accurate and fair without knowing how it works, transparent models are not necessarily a prerequisite for AI to be considered low-risk, trustworthy (or at least reliable), or fair. That said, there are other concerns about opaque AI which are in principle harder to overcome. These include concerns that transparency is necessary for accountability (Kemper & Kolkman, 2019), legitimacy (Chesterman, 2021; Selbst & Barocas, 2018; Vainio-Pekka et al., 2023), or patient autonomy/informed consent (Amann et al., 2020; Bjerring & Busch, 2021; Vaassen, 2022). These concerns are about the very nature of black-box AI and cannot be dealt with just by proving that some AI model is reliable and accurate. Nevertheless, these concerns have been disputed as well (Kawamleh, 2023; Muralidharan et al., 2024; Prince & Savulescu, 2025; Prince & Lim, 2025).

In this paper, we set aside these concerns (and whether they have been overcome), and draw attention to a different reason for why transparency matters for the use of AI in healthcare. Typical arguments in favor of AI transparency in medicine focus on the patient, and why they are owed explanations for what is done to them. But it is worth also paying attention to medical practitioners, and whether transparency can help them improve their medical practice. We claim that transparency can facilitate

[1] For example, see (Johnson et al., 2023).

critical engagement on the part of doctors and other medical professionals. Critical engagement refers to the activity whereby practitioners engage with AI models as they might human interlocutors, by critically assessing the reasons given for judgments and modifying their own judgments and beliefs according. This activity allows practitioners to improve their judgments and gain knowledge, which in turn results in more accurate and appropriate medical decisions. As a result, practitioners carry out their duties better and more confidently, which in turn can result in better outcomes for patients and their wellbeing. To our knowledge, this paper is novel in putting forward and examining this argument in detail. This paper shall be organized as follows. In section II, we develop and explain the account of transparency, and how it facilitates critical engagement. In section III we address potential criticisms of the view. Finally in section IV we consider some implications of the claims developed here, particularly for the use of AI models in emergency medicine.

To be clear, the claim of this paper is not that the arguments here show that transparency, explainability, or interpretability is necessary for the justified use of AI in medicine. Nor do we claim that transparent AI ought to be preferred in medicine, all things considered. Thus, we do not provide a comprehensive overview of all the reasons for and against transparency. Rather, it is just that there is something distinctly (and instrumentally) valuable about having transparent AI, because it facilitates critical engagement and consequently, better patient outcomes. This value is not necessarily decisive. It provides healthcare institutions with a *pro tanto* reason to prefer the use of transparent AI models over opaque ones, but must be weighed against other considerations in favor of black-box models, such as potential gains in accuracy and efficiency. A second qualification worth making is the scope of this paper is restricted to the use of AI models in medicine. The claims made in this paper *may* be applicable to other domains, but examining the use of AI models in those other domains is outside the scope of this paper.

Our arguments in this paper assume that all else being equal, it is good to increase patient outcomes and wellbeing – which as we shall argue are effects of critical engagement in medicine. Relatedly, we also assume that all else being equal, we have *pro tanto* reasons to take actions to increase someone's happiness or wellbeing. These assumptions are compatible with most plausible theories of ethics, which agree that wellbeing is valuable.[2]

[2] See (Crisp, 2001), who writes that "A theory which said that [wellbeing] just does not matter would be given no credence at all". While consequentialist theories most obviously instruct agents to increase happiness or wellbeing, wellbeing also provides agents reasons for action in other theories. For example, Kant (1998, p. 4: 430) wrote that "the natural end that all human beings have is their own happiness.. there is still only a negative and not a positive agreement with humanity as an end in itself unless everyone also tries, as far as he can, to further the ends of others". In contractualism, the fact that some act reduces someone's wellbeing or fails to contribute to it can constitute a reason to reject the action (Ashford & Mulgan, 2018). With respect to virtue ethics, there are multiple virtues which could imply that doctors have reasons to improve patient outcomes, including generosity (Hursthouse & Pettigrove, 2023), compassion (Gardiner, 2003), trustworthiness (Gardiner, 2003), or the Scholastic virtues of prudence, justice, fortitude, and temperance (Gay, 2019).

2 Transparency and Critical Engagement

In this section we begin with a brief conceptual analysis of transparency and closely related terms. Then we shall define and explain the concept of critical engagement, before discussing why critical engagement is valuable for doctors and patients in healthcare.

2.1 Transparency

We define "transparency" the following way:

Process P (a medical AI's decision-making) is transparent to an agent S (any medical practitioner) iff it is practical, in virtue of feature F of the process P, for S (given their agential resources) to know the epistemically relevant elements of the process.[3]

Less formally, a transparent AI system as one where it is *practical* for medical practitioners (reasonable to access, given their likely resources and training) to know the epistemically relevant elements of the process.

Our definition of transparency is motivated by definitions of "opacity" in the philosophical literature. A system can be "opaque" in a number of senses. It can be opaque because (i) its operators intentionally conceal its operations, (ii) because understanding it requires specialized skills, or (iii) because of a deeper mismatch between "mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation" (Burrell, 2016). Opacity of the first two kinds is an interesting and important issue because it raises issues regarding how governments and corporations can gain disproportionate power over people (Pasquale, 2015). But our concern in this paper is the third sort of opacity. Most salient to us is Alvarez's (2021) definition of "agent-neutral opacity":

[P]rocess P is opaque in an agent-independent manner iff it is impossible, in virtue of a feature F (and/or set of features) of the process P, which is irresponsive to (not enhanced or minimized by) agential resources, to know the epistemically relevant elements of the process.

Our definition of transparency is a partial opposite of Alvarez's, which captures a distinct concern about modern AI models which use machine-learning techniques. They are opaque not just because AI companies conceal their inner workings or because they are hard for laypeople to understand (although these are often also the case). They are opaque because of a more fundamental feature of the AI models themselves; the procedures used by machine learning algorithms simply do not match the way humans understand and interpret data (Burrell, 2016). No human being, no matter how technically literate, can fully understand how a machine learning algorithm translates specific inputs into outputs.

Kästner and Crook (2024) point out that the phrase "epistemically relevant" is deliberately non-specific, and includes "any robust patterns which underlie or maintain a system's behaviour and are relevant to the epistemic goals of an agent". As a result, opacity is relative "to an agent's (e.g., a company, AI user, or developer) inter-

[3] We thank an anonymous reviewer for suggesting this formulation.

ests.. at a given time" (Kästner & Crook, 2024). In this paper, we are primarily interested in whether and how transparent AI systems can help improve medical practice. Therefore, epistemically relevant elements can refer to any information that could make a difference to a medical practitioner's goals (like meeting the health needs of her patient). As we shall argue later, such information can sometimes include explanations for how an AI system arrived at an output.

Within the category of transparent AI, there is "interpretable AI", which refers to AI models which are inherently transparent. We can determine how the model makes decisions simply by studying the algorithm's structure itself (Muralidharan et al., 2024). For example, an algorithm which uses a simple linear model for all operations is likely to be interpretable. Meanwhile, "explainable AI" refers to AI models for which we have secondary post-hoc models which tell us how the primary model works (Rudin, 2019). This secondary model can be, for example, another AI model which generates "a local linear equation that emulates the behaviour of the [primary] black-box in a range of nearby cases" (Muralidharan et al., 2024). Secondary models can vary in their level of *fidelity* to the primary models, in that they can be more or less representative of how the primary model actually arrives at its results (Rudin, 2019). A high-fidelity model is a secondary model which very closely approximates how the primary model works. For our purposes, only high-fidelity models count as transparent, because they provide a reasonably accurate picture of how a primary model works. They can therefore provide agents with knowledge about epistemically relevant elements about the model. In contrast, low-fidelity models are less accurate or precise. While they may still provide users with some useful information about how the primary model works, the arguments we present in this paper may not apply as well to them. For our purposes "transparent AI" refers to any interpretable AI model, or any primary model paired with a high-fidelity secondary model.

It is possible that no explainable models can be sufficiently high-fidelity to be counted as "transparent". It may be the case that for any black-box model worth using (because of their level of accuracy and sophistication), no sufficiently representative secondary model can be developed. For example, Babic and Cohen (2023) suggest that all explainable models will fail to be action-guiding, trustworthy, or suitable for accountability practices (like blame), because they do not reflect the real reasons for why a decision was made. Another possibility is that interpretable/explainable high-fidelity models will simply be so complicated that any explanation will be beyond human comprehension. As a result, it may be that transparency is impossible for black-box models which are worth using. In such a case, a trade-off must be made between the value of transparency and the other goals held by medical institutions and professionals. All things considered, there may be reasons to prefer opaque or lower-fidelity models in some cases. Nevertheless, given the constant rate of technical progress and change, the feasibility of explainable and high-fidelity models may well change. We will remain neutral on this issue for the purposes of this paper.

2.2 Critical Engagement

To grasp what we mean by *critical engagement*, consider what it means to critically engage with a friend, interlocutor, or peer. When we critically engage with each other

on a topic, we may begin by stating our conclusions (or beliefs), and then supplying premises in favor of those conclusions. Subsequently, we may analyze the relationship between the premises and the conclusions, determining if the premises successfully support the conclusions, if arguments need to be modified, or if our beliefs are strengthened or undermined by those arguments. In most cases, human beings can also engage in continued back-and-forth dialogue. Such dialogue can be very helpful, although not necessary for critical engagement (consider someone who writes a single letter to a friend to provide them with advice on a decision). Critical engagement, as we have described, is familiar to anyone who has enjoyed collegial and productive arguments about movies, politics, or food. Critical engagement is an important and vital way by which we revise and improve our beliefs. In engaging with one another, we learn about new facts and perspectives, improve our own arguments, and revise our flawed beliefs. All this in turn helps us to interact with the world in more productive ways.

Human agents can critically engage with transparent AI models in a way that's *partially* analogous to the way we critically engage with each other. A transparent AI model can give us information about how or why the model arrived at a decision. This information can give us reasons – which we might not otherwise have – to increase or decrease our confidence in our beliefs. Information about how a model arrived at its decision can therefore count as an epistemically relevant element. To be clear, such a model alone is not perfectly analogous to a human interlocutor, as it cannot engage in further discussion and dialogue (although it may be possible to supplement the model with an LLM which then engages in dialogue). Such a model will also be unable to revise its own opinions in light of some discussion. Nevertheless, further dialogue is not strictly necessary for critical engagement to occur, and information about how the model arrived at its decision may be sufficient to help a practitioner arrive at better conclusions.

Similar points have been previously observed by Kempt and Nagel (2022) and Kempt et al. (2023). Kempt and Nagel (2022) point out that AI Decision Making Systems (DSSs) can partially play the role of someone providing a second opinion for practitioners, such that if a practitioner is challenged by an AI-DSS, the disagreement must be resolved by another practitioner. They further point out an interpretable AI system "allows for the interpretation of its inner workings *as reasons* for the physician-in-charge"(Kempt & Nagel, 2022, p. 225). Thus, interpretable models can give their users evidence about the correctness of their own opinions or reasoning. In contrast, non-explainable systems do not similarly provide users the tools to reason about the recommendations provided by the system. As a result, "a disagreement is not resolved but merely decided in favour of one side" (Kempt & Nagel, 2022). In this section, we will build on these ideas by explaining in more detail how transparent AI systems can give practitioners reasons to make decisions. And in the following section, we argue that transparent AI – not just any AI capable of providing reasons – provides distinctly valuable information for medical practitioners.

Consider the following example:

1: A doctor, upon assessing a patient, believes that she ought to provide a certain course of treatment – treatment X. After forming this opinion, she consults a transparent AI model, which recommends another course of treatment – treatment Y – for

the patient. She examines how the model arrived at its recommendation and learns that the model made its recommendation partly because of the patient's ethnicity. The doctor investigates further and learns that the patient's ethnicity is associated with some genetic dispositions for a certain disease, which she previously overlooked. In light of the new information, she revises her opinion.

In the case, the doctor is able to engage critically with the AI model. She is able to see the reasons for why the model made a recommendation, assess those reasons, and determine how those reasons can affect her own decisions. She can thus revise her decisions or not, based on informed reasoning. If she revises as she does in case 1, it might be because the model has revealed some information that she might have previously overlooked or not known about. If she rejects the recommendation of an AI model, it may be because she has decided that the model (even one that is accurate 95% of the time) has made a mistake and that her knowledge of the situation is superior. For example, we can imagine a case where a very powerful model makes a rare mistake of associating a condition with a patient's ethnicity, which the doctor spots. In case 1, the eventual outcome is materially improved by the use of the transparent AI model – if not for the model, the doctor would have made an inferior decision which would have negatively impacted the patient.

Of course, the outcome would have been improved even if the AI model was opaque and if the doctor decided to defer to the model anyway. But the difference here is that because the AI model in case 1 is transparent, it is practical for the doctor to gain new epistemically relevant information about the patient and the condition. This process of gaining new knowledge has been pointed out by Okada et al. (2023), who write that "if an explanation from an AI model shows an unexpected contribution of a certain risk factor to the prediction of outcomes, a novel hypothesis might be developed regarding this factor and its association with outcomes". Indeed, they note a study where "clinical subgroups of cardiac arrest patients treated effectively with extracorporeal cardiopulmonary resuscitation (ECPR), creatinine value was associated with outcome" by an AI designed to predict outcomes, leading to "the development of a novel score for indications of ECPR that included creatinine" (Okada et al., 2023). Of course, further work should be done to determine why creatinine value is associated with certain outcomes. Another study found that AI models learned to detect a patient's race from medical images like x-rays, even having accounted for "trivial proxies" like "body habitus, age, or tissue density", and even when trained professionals could not detect race (Gichoya et al., 2022). This suggests some correlation between race and previously unknown observable factors in medical images. Similar observations have been made in other fields like genomics (Novakovsky et al., 2023), ecosystem management (Chakraborty et al., 2021), and in theoretical physics (Krenn et al., 2022). When researchers and medical practitioners engage critically with transparent AI, they can look at why an AI model made particular recommendations, study the relationships drawn between different pieces of data, and discover new hypotheses for how previously overlooked factors can influence medical outcomes. This translates to new and potentially valuable knowledge, which might have otherwise been hard to discover (particularly if relationships between data are very obscure and if the algorithm used by the model is very complicated).

In contrast, the lack of transparency, and hence the lack of critical engagement, can be bad for the practice of medicine because it can place practitioners in very difficult dilemmas. Consider the following case:

2: A doctor, upon assessing a patient, believes that she ought to provide a certain course of treatment – treatment X. However, after forming this opinion, she consults a *black-box* AI model, which recommends another course of treatment – treatment Y – for the patient. Because the model is a black box, she has no idea why the recommendation was made.

Things don't look as good here. On one hand, the doctor can stick to her original judgment of going with treatment X rather than treatment Y. But if the black-box model has been proven to be extremely accurate and reliable – more so than human practitioners[4] – sticking to her original judgment involves knowingly acting in a way is worse in expectation. Refusing to defer to the AI model may even be hubristic and irresponsible, given the circumstances. On the other hand, she can defer to the recommendation made by the AI model, knowing that it is more likely to have made an accurate decision. But because she doesn't know why the model arrived at the decision, this course of action involves rejecting her own independent judgment without knowing why her judgment is incorrect. In acting as such, she may sacrifice her own integrity – her loyalty to her own convictions, judgments, or conclusions (Scherkoske, 2010). This is not a trivial matter. As Bernard Williams (1973) points out (in response to impartial moral principles like utilitarianism), requiring an agent to act against their own convictions is to "alienate him in a real sense from his actions and the source of his action in his own convictions". A doctor who defers to an opaque AI system may be similarly alienated when she has to set aside her own conclusions and the reasons for those conclusions in favor of a conclusion she does not understand. As a result, the doctor is left with the unenviable dilemma where she must either act in a way which is in expectation worse, or sacrifice her integrity by acting against her own judgments (a third, even less desirable, possibility is that the doctor takes it upon herself to develop her own explanation and "filling in the gaps" for the AI recommendation. This may result in the doctor deceiving herself and her patient).

Compounding the doctor's dilemma is the issue of responsibility or accountability (Kempt & Nagel, 2022). Part of accounting for a decision is providing a justification or explanation for it. If a doctor defers to the recommendation of a black-box AI model, there is a sense in which they cannot provide a full account of their final decisions. The statement, "I made this decision because the accurate AI model said I should" clearly misses something epistemically relevant to the aim of *understanding* why a decision is better than the alternatives. As a result, having access to a transparent AI may allow a doctor to account for their decisions better or more fully. To be clear, it is possible that referring to the recommendation of a black-box AI –

[4] To date, several studies have indicated that some AI models perform comparably or favorably with human practitioners in assessing triage cases and in diagnosis. For examples, see (Baker et al., 2020; Gräf et al., 2022; Razzaki et al., 2018; Shen et al., 2019).

"the AI model said so" – may sometimes constitute an *adequate* account of a decision. Whether fuller and more satisfying explanations are needed for accountability is beyond the scope of this paper.

Transparent AI is not only useful in cases where it contradicts the opinions of medical practitioners. It can also be useful in cases where it confirms their opinions, especially when practitioners make decisions based on incomplete information and intuition. Medicine has been described by some people as an art (Detsky, 2022; Moseley, 1993), reflecting the idea that gut feeling and intuition often play important roles in medicine (Adler, 2022; Hall, 2002; Kozlowski et al., 2017; McCutcheon & Pincombe, 2001; Smith et al., 2021; Stolper et al., 2009; Trafton, 2018; Van Den Bruel et al., 2012; Woolley & Kostopoulou, 2013). An experienced and proficient doctor may be described as someone who possesses *practical wisdom*, a virtue consisting of a set of "motivations, habits, dispositions, beliefs, knowledge, or abilities" (Swartwood & Tiberius, 2019). Some philosophers have observed that wisdom is *uncodifiable*, in that we cannot gain a wise person's understanding simply by learning a list of principles or rules (Swartwood & Tiberius, 2019). This claim is traceable to Aristotle, who noted that practical wisdom involves both general and abstract knowledge as well as skill in reasoning about particulars (Aristotle, 1999). Hence, the development of practical wisdom requires experience and cannot be gained merely by studying general principles (Aristotle, 1999). As a result, an experienced and proficient doctor may arrive at a conclusion without being able to articulate general reasons for how they arrived at it.[5] They may sense, based on their years of experience that some conclusion is more likely to be correct, while younger and less experienced doctors rely on guiding rules or "informal yardsticks" (Nilsson & Pilhammar, 2009; Rikers et al., 2004). When doctors make important decisions based on intuition, transparent AI can help them articulate reasons for their judgments. Practitioners can in turn use this information to improve their own skills, and to communicate their decisions more clearly to their patients. For example:

3: A patient turns up at a general practice with some difficulty breathing. The attending nurse has a gut feeling that the patient might be going through cardiac arrest, rather than something else (like an asthma attack), but they can't articulate why. Fortunately, the clinic has an AI model which confirms what they believe – based on the specific breathing patterns, it confirms that the patient is likely going through cardiac arrest (Chan et al., 2019).

It's probably valuable for medical practitioners to have AI models which confirm their suspicions regardless of whether those AI models are transparent or not. These models can help practitioners execute their decisions with greater confidence, and may reduce incidences of wrongful risk-averse decisions. But it's probably even better for AI models in such contexts to be transparent, when they can help practitioners articulate reasons that they may not have access to. Imagine further that in Case 3,

[5] There are a number of different theories about how expert intuitions, like the ones held by experienced medical professionals, work. One possibility is that expert intuitions are a form of pattern recognition. At the same time, some scholars are skeptical about how reliable expert intuitions are. For a discussion on these two points, see (Kahneman & Klein, 2009).

the AI model explains its predictions; because the patient has a particular breathing pattern associated with cardiac arrest, it is likely that the patient is suffering from cardiac arrest. The attending nurse in this case could learn something they may not have otherwise known, and may pay special attention to the breathing patterns of future patients.

A final potential benefit of critical engagement is in counteracting some worries about automation bias – a tendency to over-rely on automation (Goddard et al., 2012) – and deskilling (Duran, 2021). These two worries are interlinked; when users of AI models over-rely on automation, they will be less likely to rely on their own skills, and may therefore lose those skills. And when practitioners are less skilled, they may be less able to determine when a model should be relied on, leading to either over or under-reliance. When an AI tool is opaque, practitioners may end up adopting different attitudes towards the model. One possibility is that they become overly skeptical of the model – as Dzindolet et al. (2003) found in three studies of people using automated aids "experience with the aid can quickly lead to distrust (and disuse) if the aid malfunctions in a way that the operator is unable to explain". This is a bad outcome when using an AI tool can improve outcomes. The other possibility is that practitioners become uncritically reliant on the AI model; when a practitioner knows that a model is very accurate but doesn't know why, they may decide to simply defer to the AI model all the time. This may be a form of over-reliance, if the AI model turns out to be less than perfect.

This is where transparent AI comes in. Transparent models facilitate critical engagement, which facilitates the development and maintenance of skill. In turn, skillful practitioners may be more likely to develop an appropriate level of reliance on AI models. This claim has been corroborated by some empirical research, where studies found that explanations decreased over-reliance on automated systems (although these findings were qualified based on the types of explanations and the domain) (Vasconcelos et al., 2023; Vered et al., 2023). When a user is presented with the reasons for a decision (presumably, in intelligible and user-friendly ways), they are encouraged to engage critically with the reasons. In doing so, they become less likely to accept the decision uncritically, and thus be less likely to over-rely on automation.

Thus, transparent AI is valuable because doctors, other healthcare practitioners, and researchers using it may gain skill, knowledge, and confidence. This outcome may be valuable in and of itself, but more importantly, this should improve patient outcomes. Doctors using transparent AI may deliver service more confidently and more appropriate to the patient's situation. This value gives us an instrumental reason to prefer the use of transparent AI over opaque AI. Of course, this reason is not by itself a decisive reason to use transparent AI instead of opaque AI. It may supplement other reasons in favor of transparent AI, including "in principle" considerations (e.g. that opaque AI violates patient autonomy) (Amann et al., 2020; Bjerring & Busch, 2021; Vaassen, 2022). It must also be weighed against the benefits of using opaque AI in medicine, such as the possibility of increased accuracy (London, 2019). Nevertheless, when we assess the balance of reasons, it is important to consider the benefits we have discussed.

3 Objections

Let us now consider some potential objections. Transparent AI models may not always provide any information that is useful for deliberation. Consider the following example:

[S]uppose that a patient, Hal, was looking at a LIME system AI and can see why given the information provided, palliative care instead of chemotherapy was recommended by the AI. However, this explanation invokes considerations that plausibly have little relation to whether he should go for chemotherapy. For instance, suppose that the primary algorithm heavily weighted the fact that Hal likes to play badminton in how it actually came to the recommendation.. However, it would not be clear whether playing badminton is a good proxy for some ground truth about factors that affect QALY or whether it reflects some sort of bias in the training data (Muralidharan et al., 2024).

According to Muralidharan et al. (2024), the AI system hasn't improved Hal's epistemic position, because the explanation provided for its decision – that Hal likes to play badminton – seems to have no relevant relation to whether he should go for chemotherapy. Such information seems to be a non sequitur, and is thus hard to engage with.

Nevertheless, the relevant lesson from this example and similar cases is not to jettison transparency. Rather, it is that transparent AI models should be designed in ways which make it clear how morally or epistemically relevant reasons factor into the recommendations made by the AI, and that doctors should be trained to engage with AI explanations in constructive ways. To illustrate, Jabbour et al. found that explanations do not improve outcomes when clinicians do not really understand the implications of explanations, and have insufficient "AI literacy" (Jabbour et al., 2023). But notably, they did not conclude that explainable AI is pointless. Rather, they emphasized good AI design and training:

.. Second, researchers developing explanation tools should involve clinicians to better understand their specific needs. Third, standardizing clinician-facing AI model information in simple language and empirically testing such standards may help clinicians understand appropriate model use and limitations. (Jabbour et al., 2023).

Obviously, more can and should be said about specific design principles for transparent AI. But the point here is that transparent AI models *can* be useful for deliberation and justification. Indeed, as we noted earlier, several studies have already demonstrated that transparent AI models can reveal new epistemically relevant information.

One way of pressing this objection further is to assert that AI algorithms will inevitably be so complicated that transparent AI will be completely unintelligible for users. For example, imagine a black-box model which, when "opened" reveals that it recommends palliative care based on the patient's yoghurt consumption, the number of plants in the hospital, and the climate conditions of the first ten years of the patient's life. Regardless of how much training a practitioner has, they may struggle to make sense of such a recommendation. Yet, significant research in computing is being carried out on making AI explanations intelligible or graspable (Coste-Marquis & Marquis, 2020; Ribera & Lapedriza, 2019; Zhang & Lim, 2022; Zuccarelli, 2024).

Furthermore, there are existing AI models, like the "Score for Emergency Risk Prediction" (SERP) model which produce very good results and are interpretable and intelligible (Look et al., 2024). Therefore, the jury is still out on whether we can indeed design transparent AI systems which are both intelligible and effective. It is premature to declare that transparent AI is doomed to unintelligibility, and we should not yet give up on trying to develop transparent AI.

Muralidharan et al. (2024) provide a helpful distinction between two chains of reasoning: *explanations* – how a decision was in fact made – and normative *justifications* – why a decision is correct or permissible according to some "normative or evaluative criteria". Decisions are normatively justified (from here on, simply "justified") by normative reasons – considerations which weigh in favor of a decision or course of action (Alvarez & Way, 2024; Scanlon, 1998). While explanations for our decisions often involve normative reasons, explanations and justifications can come apart. For example, suppose you notice a colleague, who you greatly dislike, stealing office supplies. You report his behavior, not because you care about the stealing, but because you dislike him and want him to suffer. The explanation of your behavior is that you want him to suffer, but this is not a normative reason. Rather, the justification for your actions is that stealing office supplies is wrong and should be stopped.[6] Muralidharan et al. (2024) go on to imagine a "justifiable AI", which is "a secondary AI, based on an LLM, which churns out one or more plausible justifications" for a recommendation made by a primary AI. This secondary AI is essentially a creative machine, inventing reasons which constitute post-hoc justifications for a decision. These reasons may or may not be faithful to how the decision was in fact made. This example allows us to see how a justifiable AI differs from a transparent AI. The former tells us some story about why a decision can be justified, while the latter tells us (or at least tries to approximate) why the decision was in fact made.

From here, we can see a second possible objection to our account in this paper: transparent AI models aren't strictly necessary for critical engagement. You could argue that ultimately what matters is that some other plausible normative reasons are offered for the recommendations offered by AI models. In particular, we should focus on justifications rather than explanations. These justifications may or may not be faithful to how a decision was made, and may be a post-hoc model supplying reasons which explain how a decision is compatible with a patient's goals and values. Such a justification is sufficient, and perhaps more useful, for the purposes of critical engagement.

In principle, we are sympathetic to this line of thought. In certain circumstances, justifications that do not provide an explanation of the basis of an algorithm's output could indeed promote critical engagement. But in practice, explanations may be more supportive of critical engagement in many cases. Consider that, for any given decision, there are multiple chains of reasoning (or arguments) which can be given in an attempt to justify the decision. Many of those chains may be flawed – they may contain irrelevant reasons, logical fallacies, non sequiturs, make bad use of data, and so on. Sometimes, it will be easy to identify flawed chains of reasoning where the reasons offered do not in fact support the eventual decision. But other times the flaws

[6] Assuming you don't work for an evil corporation.

may be subtle, and will be difficult for people untrained in critical reasoning to pick them out. As we know from arguments with friends, public discourse, and politics, it's not hard for a skilled orator to supply an apparently convincing story for almost anything! Furthermore, even skilled and conscientious deliberators are vulnerable to the effect of biases. We often make mistakes in assessing reasons for conclusions, especially when we agree with those conclusions. Even under good conditions, decent reasoners may make mistakes when assessing reasons. Given that patients and practitioners must often deliberate under emotionally trying circumstances, they cannot be expected to tell good chains of reasoning apart from (subtly) bad ones reliably. As a result, they may mistake some reasons as objective justifications for certain decisions even though those decisions are not in fact justified.

This brings us to a first potential problem with focusing on justification at the expense of transparency. Because people may not always have the capability or capacity to carefully differentiate between good and bad reasons for decisions, they may engage in bad reasoning themselves. If doctors focus on supplying their own post-hoc justifications for the decisions supplied by black-box AI models, those "justifications" could instead end up being faulty post-hoc rationalizations driven by motivated reasoning. Doctors can make these mistakes even when they are intelligent, conscientious, well-intentioned, and reason in good faith. If doctors use secondary justification AI models, like the kind Muralidharan et al. describe, those AI models can also make mistakes and effectively deceive their users. Contemporary LLMs like GPT-4 are trained on data from writing all over the internet, and are prone to replicating the kinds of errors in reasoning that human beings make, as well as making simple mistakes that humans don't make.[7] Users who put too much trust in AI models may be fooled into believing these mistakes.

What if the reasoning skills of LLMs improve? In such a case, we may end up in an even worse situation. As an analogy, imagine that your doctor hires a rhetorician to justify all of her decisions. This rhetorician is famously skillful and can invent a convincing story for almost any conclusion – even conclusions we know to be false. You probably wouldn't take the rhetorician's arguments seriously, nor would their arguments help you trust your doctor more (if anything, you may trust your doctor less). If we design an LLM to justify the decisions of another primary model, the LLM becomes like the very skilled but untrustworthy rhetorician. As Uwe Peters points out in a discussion of AI systems which provide reason-giving explanations (i.e. justifications),

AI explanation systems would produce explanations that, during their development, people have learned to accept as justifications for decisions.. The proposal to design ADM systems that produce reason-giving explanations may thus lead to the development of AI systems that create a false impression of transparency and trustworthiness by co-opting an often warranted pre-existing human tendency to assume reason-givers are self-regulators (Peters, 2023).

[7] A study on the logical reasoning skills of GPT-4 concluded that while the model outperformed a number of benchmarks, the model faced "challenges in handling new and out-of-distribution data", and that it does not perform well "on the natural language inference task requiring logical reasoning" (Liu et al., 2023). Similar findings about LLMs have been made by (Azaria et al., 2024; Dougrez-Lewis et al., 2024; Groza, 2023; Valmeekam et al., 2022; Zhang et al., 2024).

Just like the rhetorician, this sort of LLM will end up either deceiving its users, or be perceived as untrustworthy. Even when you can't pinpoint any obvious flaws in the "justification" provided, you still know that it's designed to provide "justifications" for all decisions, even the bad decisions. As a result, unless you have the capacity and skill to carefully analyze the arguments made by the LLM, you may be better off ignoring what it says.

But let's suppose that we can design a post-hoc justification AI model that can and does reason well, and does not provide "justifications" for objectively unjustified decisions. Presumably, such a model will give practitioners and patients useful information about how a decision is compatible with the patient's values and goals. It provides this information using chains of reasoning which meets some standards of logical coherence and relevance. It supplies reasons which are supported by plausible and relevant scientific and ethical theories. And instead of attempting to justify *any* recommendation made by the primary model, it points out when the primary model may have made a mistake. This sort of AI model could certainly be of great value. Nevertheless, there is still a limitation. Because the information presented here is a *post-hoc* justification for a given conclusion, we do not have information about whether this chain of reasoning would be *accurate* at making new predictions. Even when a chain of reasoning justifies a conclusion in a way that is consistent with a patient's values and goals, and abides by all standards of logical inference, it may not be great at making new predictions. Indeed, if the model was indeed highly accurate and also great at providing justification, we might wonder why we don't use that secondary model as the primary model instead, giving us a primary model which both explains and justifies its decisions simultaneously (in reality, secondary models of this sort may not be designed to make accurate predictions anyway).

This is where transparency comes in. AI models have shown potential to be highly accurate at making new predictions in healthcare, matching or improving on the assessments made by practitioners (Alowais et al., 2023; Jiang et al., 2017; Khalifa & Albadawy, 2024; Nadella et al., 2023). They are iteratively trained on large bodies of data to get the most accurate results, given certain parameters (like the type of data the model was trained on). Thus, we can expect that a proven AI model will use a process which is best or near-best in terms of accuracy at making predictions. So a transparent AI model, which reveals its processes to users, offers something uniquely valuable: a chain of reasoning which reaches the right conclusions with very high accuracy (in cases of explainable AI, which involve post-hoc explanations, those AI models will only provide this valuable chain of reasoning if they are high-fidelity models). Even though this chain of reasoning might not seem logically coherent at first, information about this chain of reasoning could be useful for human decision-makers who may discover relationships (between seemingly disconnected points of data) that they may never have thought of themselves. For example, in the example described by Muralidharan et al., an AI model recommends palliative care to Hal based on the fact that he likes to play badminton. While seemingly irrelevant at first glance, users could infer that the AI model is weighing "plays badminton" as a proxy for living an active lifestyle. A practitioner like Hal's doctor could then use this information as a prompt to discuss Hal's lifestyle with him, potentially revealing some preferences and values neither Hal nor his doctor may have been conscious of. They

can then use this information to help their deliberation about subsequent decisions, like the nature of palliative care. This point is not entirely theoretical. As we pointed out earlier, several studies have connected the use of transparent AI to new discoveries in medicine, genomics, and other sciences. In this way, transparent AI has the potential to provide practitioners and patients with unique insights and new knowledge about how certain facts can be related to certain conclusions about a patient's condition or the treatment available to them.

4 Implications

As we have argued, transparent AI models can help doctors and other healthcare practitioners improve their own practices by facilitating understanding of how decisions are made. Conversely, opaque AI can create undesirable situations where doctors may have to choose between acting on decisions they don't understand or holding fast to their own (probably worse) decisions. It is thus quite clear that transparent AI models can bring about material benefits. More capable and knowledgeable doctors should result in better patient outcomes. Transparent AI is also likely to engender more trust from patients, which "At least in the medium term.. will be strongly mediated by their trust in physicians" (Nickel, 2022). Doctors who are able to understand and explain the recommendations of AI models are more likely to gain the trust of their patients, which in turn will help their patients trust the recommendations.

One clear implication of the claims in this paper is that developers of AI models in healthcare should aim for transparency, either by making those models interpretable or explainable. But this implication needs to be qualified. As we have noted, the arguments here do not show that transparency is necessary for the use of AI models to be justified all-things-considered. Rather, the arguments only show that there's something distinctly valuable about transparent AI. If there are trade-offs involved in making AI transparent, then this value must be weighed against those trade-offs.

At the same time, one implication of this paper is also stronger than many other arguments in favor of transparent AI. As we noted earlier, for medical practitioners and their patients to critically engage with AI models, those AI models need to be sensible to them. This means that it's not enough for AI models to be transparent just in the sense that some people, like AI experts, can understand them. Those without any formal training in computing or AI should be able to understand (to some extent) how AI models make their decisions, assess them, and factor the AI chain of reasoning into their own reasoning. In other words, transparent AI should provide explanations of its decisions in plain language. This may be achieved with a secondary LLM which constructs a post-hoc model of how the primary AI works, and then explains it in everyday language. In addition, medical practitioners who use AI models in their work should be trained in basic "AI literacy". They should learn some basic information about how neural networks and machine learning work, the strengths and weaknesses of AI models, the risks of bias and error in AI models, and so on. Such training may give practitioners the skill to productively engage with AI models. The training is particularly useful when dealing with AI models which make use of connections and information which initially don't seem relevant, like a patient's ethnic-

ity or badminton hobby. AI literacy training can also help medical practitioners gain confidence in working with AI models, so that they can avoid dilemmas like those described in section II (with respect to case 2).

Another implication of the claims here is that transparent AI may be of different value in different medical contexts. Consider the difference between non-emergency healthcare and emergency healthcare. While stakes can be very high in both non-emergency and emergency healthcare, emergency healthcare is characterized by high urgency. The value of critical engagement may thus be different in the two sets of contexts. We want doctors in non-emergency cases to take their time and consider all the variables and factors in order to make the best decisions. But critical engagement takes time and careful and methodical consideration. These are luxuries in the world of emergency medicine, where taking the time to critically evaluate information may come with the cost of losing precious time.[8] Such costs may well outweigh the value of critical engagement.

This is not to say that transparency is of no value in emergency healthcare. Rather, depending on the nature of specific tasks, the value of transparency can differ. Indeed, in a qualitative study, Townsend et al. (2023) found that some emergency health practitioners see transparency as important for the use of AI in healthcare, with one practitioner noting "I would like to know why a diagnosis was made" and "I would want to know this [so as to know whether] to do further investigations to confirm or deny [the diagnosis]". This quote shows a desire on the part of the practitioner to engage critically with the recommendations made by AI models, even in emergency departments. Another qualitative study of emergency radiologists found that participants also valued transparency (Agrawal et al., 2023). However, a study by Stewart et al. (2024) on emergency clinicians raised fewer concerns with transparency and explainability while another study found that pathologists view transparency as less important (Drogt et al., 2022). For a conclusive picture on how emergency practitioners view AI transparency, more studies will have to be conducted, as well as on why they care about transparency.

The desire of emergency practitioners to understand why AI decisions are made reveals what is likely a shared desire among practitioners. Medical practitioners (at least sometimes) want to be actively and critically engaged in the decisions they ultimately have to make. They want to know why the decisions they make are the right ones, and to do that they need to be able to assess the recommendations made by AI models. Transparent AI models facilitate such critical engagement.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13347-025-01009-w.

Acknowledgements We would like to thank Anantharaman Muralidharan, Sinead Prince, and other members of the National University of Singapore's Centre for Biomedical Ethics for providing comments on versions of this paper.

[8] For example, Banerjee et al. found that in paediatric cardiac arrest cases, "Each additional minute for the EMS to arrive resulted in 5% decreased odds of ROSC [return of spontaneous circulation] and hospital admission, and 12% decreased odds of surviving to hospital discharge" (Banerjee et al., 2021).

Author Contributions The first draft of the manuscript was written by James Edgar Lim, and all authors made contributions to and commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Statement The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the NRF/AISG Governance [grant number AISG3-GV-2023-012]; and the Wellcome Trust [grant number 226801/Z/22/Z].

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.
Declaration of competing interests.
JS is a Bioethics Committee consultant for Bayer. JS is a Bioethics Advisor to the Hevolution Foundation.

Declarations

Ethical Approval No ethical approval was required for this study.

Consent for Publication Not applicable.

Consent To Participate Not applicable.

Competing interests JS is a Bioethics Committee consultant for Bayer. JS is a Bioethics Advisor to the Hevolution Foundation.

References

Adler, I. (2022). The medical gap: Intuition in medicine. *Medicine, Health Care and Philosophy, 25*(3), 361–369. https://doi.org/10.1007/s11019-022-10081-4

Agrawal, A., Khatri, G. D., Khurana, B., Sodickson, A. D., Liang, Y., & Dreizin, D. (2023). A survey of ASER members on artificial intelligence in emergency radiology: Trends, perceptions, and expectations. *Emergency Radiology*, *30*(3), 267–277. https://doi.org/10.1007/s10140-023-02121-0

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Saleh, B., Badreldin, K., Yami, H. A. A., Harbi, M. S. A., S., & Albekairy, A. M. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(1), 689. https://doi.org/10.1186/s12909-023-04698-z

Alvarez, M., & Way, J. (2024). Reasons for Action: Justification, Motivation, Explanation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2024/entries/reasons-just-vs-expl/

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

Aristotle (1999). *Nicomachean Ethics* (W. D. Ross, Trans.). Batoche Books.

Ashford, E., & Mulgan, T. (2018). Contractualism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/contractualism/

Azaria, A., Azoulay, R., & Reches, S. (2024). ChatGPT is a remarkable tool—For experts. *Data Intelligence, 6*(1), 240–296. https://doi.org/10.1162/dint_a_00235

Babic, B., & Cohen, I. G. (2023). *The Algorithmic Explainability Bait and Switch* (SSRN Scholarly Paper No. 4541487). https://papers.ssrn.com/abstract=4541487

Baker, A., Perov, Y., Middleton, K., Baxter, J., Mullarkey, D., Sangar, D., Butt, M., DoRosario, A., & Johri, S. (2020). A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2020.543405

Balasubramaniam, N., Kauppinen, M., Hiekkanen, K., & Kujala, S. (2022). Transparency and Explainability of AI Systems: Ethical Guidelines in Practice. In V. Gervasi & A. Vogelsang (Eds.), *Requirements Engineering: Foundation for Software Quality* (pp. 3–18). Springer International Publishing. https://doi.org/10.1007/978-3-030-98464-9_1

Banerjee, P., Ganti, L., Stead, T. G., Vera, A. E., Vittone, R., & Pepe, P. E. (2021). Every one-minute delay in EMS on-scene resuscitation after out-of-hospital pediatric cardiac arrest lowers ROSC by 5%. *Resuscitation Plus*, *5*, 100062. https://doi.org/10.1016/j.resplu.2020.100062

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology, 34*(2), 349–371. https://doi.org/10.1007/s13347-019-00391-6

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Chakraborty, D., Başağaoğlu, H., Gutierrez, L., & Mirchi, A. (2021). Explainable AI reveals new hydroclimatic insights for ecosystem-centric groundwater management. *Environmental Research Letters*, *16*(11), 114024. https://doi.org/10.1088/1748-9326/ac2fde

Chan, J., Rea, T., Gollakota, S., & Sunshine, J. E. (2019). Contactless cardiac arrest detection using smart devices. *Npj Digital Medicine*, *2*(1), 52. https://doi.org/10.1038/s41746-019-0128-7

Chesterman, S. (2021). Through a glass, darkly: Artificial intelligence and the problem of opacity. *American Journal of Comparative Law, 69*(2), 271–294. https://doi.org/10.1093/ajcl/avab012

Cortese, J. F. N. B., Cozman, F. G., Lucca-Silveira, M. P., & Bechara, A. F. (2023). Should explainability be a fifth ethical principle in AI ethics? *AI and Ethics*, *3*(1), 123–134. https://doi.org/10.1007/s43681-022-00152-w

Coste-Marquis, S., & Marquis, P. (2020). *From Explanations to Intelligible Explanations*.

Crisp, R. (2001). *Well-Being*. https://plato.stanford.edu/entries/well-being/?ref=pasteurscube.com

Detsky, A. S. (2022). Learning the art and science of diagnosis. *JAMA, 327*(18), 1759–1760. https://doi.org/10.1001/jama.2022.4650

Dougrez-Lewis, J., Akhter, M. E., He, Y., & Liakata, M. (2024). Assessing the reasoning abilities of ChatGPT in the context of claim verification. *arXiv*. https://doi.org/10.48550/arXiv.2402.10735

Drogt, J., Milota, M., Vos, S., Bredenoord, A., & Jongsma, K. (2022). Integrating artificial intelligence in pathology: A qualitative interview study of users' experiences and expectations. *Modern Pathology*, *35*(11), 1540–1550. https://doi.org/10.1038/s41379-022-01123-6

Duran, L. D. D. (2021). *Deskilling of medical professionals: An unintended consequence of AI implementation?*.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Gardiner, P. (2003). A virtue ethics approach to moral dilemmas in medicine. *Journal of Medical Ethics*, *29*(5), 297–302. https://doi.org/10.1136/jme.29.5.297

Gay, R. (2019). Virtue Ethics and Medical Law. In A. M. Phillips, T. C. de Campos, & J. Herring (Eds.), *Philosophical Foundations of Medical Law* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780198796558.003.0002

Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L. C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S. C., Kuo, P. C., Lungren, M. P., Palmer, L. J., Price, B. J., Purkayastha, S., Pyrros, A. T., Oakden-Rayner, L., Okechukwu, C., Seyyed-Kalantari, L., & Zhang, H. (2022). AI recognition of patient race in medical imaging: A modelling study. *The Lancet Digital Health*, *4*(6), e406–e414. https://doi.org/10.1016/S2589-7500(22)00063-2

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Gräf, M., Knitza, J., Leipe, J., Krusche, M., Welcker, M., Kuhn, S., Mucke, J., Hueber, A. J., Hornig, J., Klemm, P., Kleinert, S., Aries, P., Vuillerme, N., Simon, D., Kleyer, A., Schett, G., & Callhoff, J. (2022). Comparison of physician and artificial intelligence-based symptom checker diagnostic accuracy. *Rheumatology International*, *42*(12), 2167–2176. https://doi.org/10.1007/s00296-022-05202-4

Groza, A. (2023). Measuring reasoning capabilities of ChatGPT. *arXiv*. https://doi.org/10.48550/arXiv.2310.05993

Günther, M., & Kasirzadeh, A. (2022). Algorithmic and human decision making: For a double standard of transparency. *AI and Society, 37*(1), 375–381. https://doi.org/10.1007/s00146-021-01200-5

Hall, K. H. (2002). Reviewing intuitive decision-making and uncertainty: The implications for medical education. *Medical Education, 36*(3), 216–224. https://doi.org/10.1046/j.1365-2923.2002.01140.x

Hursthouse, R., & Pettigrove, G. (2023). Virtue Ethics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/

Jabbour, S., Fouhey, D., Shepard, S., Valley, T. S., Kazerooni, E. A., Banovic, N., Wiens, J., & Sjoding, M. W. (2023). Measuring the impact of AI in the diagnosis of hospitalized patients: A randomized clinical vignette survey study. *Journal of the American Medical Association*, *330*(23), 2275–2284. https://doi.org/10.1001/jama.2023.22295

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*. https://doi.org/10.1136/svn-2017-000101

Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., Scoville, E., Reese, T., Friedman, D., Bastarache, J., van der Heijden, Y., Wright, J., Carter, N., Alexander, M., Choe, J., & Wheless, L. (2023). Assessing the accuracy and reliability of AI-Generated medical responses: An evaluation of the Chat-GPT model. *Research Square*, *rs.3.rs-2566942*. https://doi.org/10.21203/rs.3.rs-2566942/v1

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kant, I. (1998). *Groundwork of the metaphysics of morals* (M. J. Gregor, Trans.). Cambridge Univ. Press.

Kästner, L., & Crook, B. (2024). Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science*, *14*(4), 52. https://doi.org/10.1007/s13194-024-00614-4

Kawamleh, S. (2023). Against explainability requirements for ethical artificial intelligence in health care. *AI and Ethics*, *3*(3), 901–916. https://doi.org/10.1007/s43681-022-00212-1

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information Communication & Society*, *22*(14), 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967

Kempt, H., & Nagel, S. K. (2022). *Responsibility, second opinions and peer-disagreement: Ethical and epistemological challenges of using AI in clinical diagnostic contexts* |. *Journal of Medical Ethics*. https://jme.bmj.com/content/48/4/222

Kempt, H., Heilinger, J. C., & Nagel, S. K. (2023). I'm afraid i can't let you do that, doctor: Meaningful disagreements with AI in medical contexts. *AI & SOCIETY, 38*(4), 1407–1414. https://doi.org/10.1007/s00146-022-01418-x

Khalifa, M., & Albadawy, M. (2024). AI in diagnostic imaging: Revolutionising accuracy and efficiency. *Computer Methods and Programs in Biomedicine Update*, *5*, 100146. https://doi.org/10.1016/j.cmpbup.2024.100146

Kozlowski, D., Hutchinson, M., Hurley, J., Rowley, J., & Sutherland, J. (2017). The role of emotion in clinical decision making: An integrative literature review. *BMC Medical Education*, *17*(1), 255. https://doi.org/10.1186/s12909-017-1089-7

Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A., Yao, Z., & Aspuru-Guzik, A. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics, 4*(12), 761–769. https://doi.org/10.1038/s42254-022-00518-3

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of ChatGPT and GPT-4. *arXiv*

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report, 49*(1), 15–21. https://doi.org/10.1002/hast.973

Look, C. S., Teixayavong, S., Djärv, T., Ho, A. F., Tan, K. B., & Ong, M. E. (2024). Improved interpretable machine learning emergency department triage tool addressing class imbalance. *Digital Health*, *10*, 20552076241240910. https://doi.org/10.1177/20552076241240910

McCutcheon, H. H. I., & Pincombe, J. (2001). Intuition: An important tool in the practice of nursing. *Journal of Advanced Nursing*, *35*(3), 342–348. https://doi.org/10.1046/j.1365-2648.2001.01882.x

Mitchell, T. (2025). Trust and transparency in artificial intelligence. *Philosophy & Technology*, *38*(3), 87. https://doi.org/10.1007/s13347-025-00916-2

Moseley, R. (1993). Intuition in the Art and Science of Medicine. In C. Delkeskamp-Hayes & M. A. G. Cutter (Eds.), *Science, Technology, and the Art of Medicine* (Vol. 44, pp. 211–218). Springer Netherlands. https://doi.org/10.1007/978-94-017-2960-4_13

Muralidharan, A., Savulescu, J., & Schaefer, G. O. (2024). AI and the need for justification (to the patient). *Ethics and Information Technology*, *26*(1), 16. https://doi.org/10.1007/s10676-024-09754-w

Nadella, G. S., Satish, S., Meduri, K., & Meduri, S. S. (2023). A systematic literature review of advancements, challenges and future directions of AI and ML in healthcare. *International Journal of Machine Learning for Sustainable Development, 5*(3), 115–130.

Nickel, P. J. (2022). Trust in medical artificial intelligence: A discretionary account. *Ethics and Information Technology*, *24*(1), 7. https://doi.org/10.1007/s10676-022-09630-5

Nilsson, M. S., & Pilhammar, E. (2009). Professional approaches in clinical judgements among senior and junior doctors: Implications for medical education. *BMC Medical Education*, *9*(1), 25. https://doi.org/10.1186/1472-6920-9-25

Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., & Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, *24*(2), 125–137. https://doi.org/10.1038/s41576-022-00532-2

Okada, Y., Ning, Y., & Ong, M. E. H. (2023). Explainable artificial intelligence in emergency medicine: An overview. *Clinical and Experimental Emergency Medicine*, *10*(4), 354–362. https://doi.org/10.15441/ceem.23.145

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Peters, U. (2023). Explainable AI lacks regulative reasons: Why AI and human decision-making are not equally opaque. *AI and Ethics*, *3*(3), 963–974. https://doi.org/10.1007/s43681-022-00217-w

Prince, S., & Savulescu, J. (2025).) When is black-box AI justifiable to use in healthcare? *Big Data and Society, 12*(4). https://doi.org/10.1177/20539517251386037

Prince, S., & Lim, J. E. (2025). Black-box AI and patient autonomy. *Minds and Machines, 35*(2), 1–19. https://doi.org/10.1007/s11023-025-09729-w

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, *149*, 106043. https://doi.org/10.1016/j.compbiomed.2022.106043

Razzaki, S., Baker, A., Perov, Y., Middleton, K., Baxter, J., Mullarkey, D., Sangar, D., Taliercio, M., Butt, M., Majeed, A., DoRosario, A., Mahoney, M., & Johri, S. (2018). A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *arXiv*. https://doi.org/10.48550/arXiv.1806.10698

Ribera, M., & Lapedriza, A. (2019). *Can we do better explanations? A proposal of user-centered explainable AI*. https://openaccess.uoc.edu/handle/10609/99643

Rikers, R. M. J. P., Loyens, S. M. M., & Schmidt, H. G. (2004). The role of encapsulated knowledge in clinical case representations of medical students and family doctors. *Medical Education, 38*(10), 1035–1043. https://doi.org/10.1111/j.1365-2929.2004.01955.x

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press. https://doi.org/10.2307/j.ctv134vmrn

Scherkoske, G. (2010). Integrity and moral danger. *Canadian Journal of Philosophy*, *40*(3), 335–358. https://doi.org/10.1080/00455091.2010.10716726

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review, 87*(3), 1085–1140.

Shen, J., Zhang, C. J. P., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S. Y., Fang, P. H., & Ming, W. K. (2019). Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Medical Informatics*, *7*(3), e10010. https://doi.org/10.2196/10010

Smith, C. F., Kristensen, B. M., Andersen, R. S., Hobbs, F. R., Ziebland, S., & Nicholson, B. D. (2021). GPs' use of gut feelings when assessing cancer risk: A qualitative study in UK primary care. *British Journal of General Practice*, *71*(706), e356–e363. https://doi.org/10.3399/bjgp21X714269

Stewart, J., Freeman, S., Eroglu, E., Dumitrascu, N., Lu, J., Goudie, A., Sprivulis, P., Akhlaghi, H., Tran, V., Sanfilippo, F., Celenza, A., Than, M., Fatovich, D., Walker, K., & Dwivedi, G. (2024). Attitudes towards artificial intelligence in emergency medicine. *Emergency Medicine Australasia*, *36*(2), 252–265. https://doi.org/10.1111/1742-6723.14345

Stolper, E., Van Royen, P., Van de Wiel, M., Van Bokhoven, M., Houben, P., Van der Weijden, T., & Dinant, J., G (2009). Consensus on gut feelings in general practice. *BMC Family Practice*, *10*, 66. https://doi.org/10.1186/1471-2296-10-66

Swartwood, J., & Tiberius, V. (2019). Philosophical foundations of wisdom. In R. J. Sternberg, & J. Glück (Eds.), *The Cambridge handbook of wisdom* (1st ed., pp. 10–39). Cambridge University Press. https://doi.org/10.1017/9781108568272.003

Townsend, B. A., Plant, K. L., Hodge, V. J., Ashaolu, O., & Calinescu, R. (2023). Medical practitioner perspectives on AI in emergency triage. *Frontiers in Digital Health*. https://doi.org/10.3389/fdgth.2023.1297073

Trafton, A. (2018, July 20). *Doctors rely on more than just data for medical decision making*. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2018/doctors-rely-gut-feelings-decision-making-0720

Vaassen, B. (2022). AI, opacity, and personal autonomy. *Philosophy & Technology, 35*(4), 88. https://doi.org/10.1007/s13347-022-00577-5

Vainio-Pekka, H., Agbese, M.O.-O., Jantunen, M., Vakkuri, V., Mikkonen, T., Rousi, R., & Abrahamsson, P. (2023). The role of explainable AI in the research field of AI ethics. *ACM Transactions on Interactive Intelligent Systems, 13*(4), Article 26:1–26:39. https://doi.org/10.1145/3599974

Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022, November 18). *Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change)*. NeurIPS 2022 Foundation Models for Decision Making Workshop. https://openreview.net/forum?id=wUU-7XTL5XO

Van Den Bruel, A., Thompson, M., Buntinx, F., & Mant, D. (2012). Clinicians' gut feeling about serious infections in children: Observational study. *BMJ (Clinical Research Ed.), 345*, Article e6144–e6144. https://doi.org/10.1136/bmj.e6144

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *arXiv*. https://doi.org/10.48550/arXiv.2212.06823

Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence*, *322*, 103952. https://doi.org/10.1016/j.artint.2023.103952

Woolley, A., & Kostopoulou, O. (2013). Clinical intuition in family medicine: More than first impressions. *Annals of Family Medicine, 11*(1), 60–66. https://doi.org/10.1370/afm.1433

Zhang, W., & Lim, B. Y. (2022). Towards Relatable Explainable AI with the Perceptual Process. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–24. https://doi.org/10.1145/3491102.3501826

Zhang, Y., Wang, H., Feng, S., Tan, Z., Han, X., He, T., & Tsvetkov, Y. (2024). Can LLM graph reasoning generalize beyond pattern memorization? *arXiv*. https://doi.org/10.48550/arXiv.2406.15992

Zuccarelli, E. (2024, April 2). *Trusting AI requires we move beyond black-box algorithms*. World Economic Forum. https://www.weforum.org/agenda/2024/04/building-trust-in-ai-means-moving-beyond-black-box-algorithms-heres-why/