AI and Ethics (2023) 3:901–916
https://doi.org/10.1007/s43681-022-00212-1

ORIGINAL RESEARCH

Against explainability requirements for ethical artificial intelligence in health care

Suzanne Kawamleh[1]

Received: 13 June 2022 / Accepted: 12 August 2022 / Published online: 29 August 2022

Abstract
It is widely accepted that explainability is a requirement for the ethical use of artificial intelligence (AI) in health care. I challenge this Explainability Imperative (EI) by considering the following question: does the use of epistemically opaque medical AI systems violate existing legal standards for informed consent? If yes, and if the failure to meet such standards can be attributed to epistemic opacity, then explainability *is* a requirement for AI in healthcare. If not, then based on at least one metric of ethical medical practice (informed consent), explainability is not required for the ethical use of AI in healthcare. First, I show that the use of epistemically opaque AI applications is compatible with meeting accepted legal criteria for informed consent. Second, I argue that human experts are also black boxes with respect to the criteria by which they arrive at a diagnosis. Human experts can nonetheless meet established requirements for informed consent. I conclude that the use of black-box AI systems does not violate patients' rights to informed consent, and thus, with respect to informed consent, explainability is not required for medical AI.

Keywords Explainability · Informed consent · Medical records · Artificial intelligence

1 Introduction

There is growing excitement about the potential of artificial intelligence (AI) and deep learning systems (DLS) to revolutionize healthcare.[1] For example, fully automated and autonomous medical AI systems have been recently approved by the FDA for diabetic retinopathy screening. However, the shift to automated medical decision-making has raised epistemic, legal, and ethical challenges.

1.1 The problem

AI models, specifically deep learning systems,[2] are "black boxes" and "epistemically opaque". This means the inner workings of AI models—how or why the model reaches a certain decision—are opaque or a "black box" to experts. This is problematic because critical healthcare decisions are being made and automated by machines that experts can neither understand nor control.

This has led scholars to conclude that "Relying on devices whose logic is opaque violates principles of medical ethics" (Kundu [16], 1328). In particular, the use of epistemically opaque medical AI systems is taken to be in direct conflict with patients' right to informed consent[3]:

> "The opacity problem of artificial intelligence seems to be in direct conflict with the right of patients to be provided with meaningful information about the logic involved as well as about the significance and the envisaged consequences of diagnostic or treatment procedures, or other interventions into health (Bachle 2019)" (Astromske et al. [1], 516).

✉ Suzanne Kawamleh
skawamle@iu.edu

[1] Department of Philosophy, Indiana University, Bloomington, IN, USA

[1] Deep learning is a specific type of artificial intelligence which refers to complex forms of machine learning, like neural networks with several layers. Epistemic opacity and explainability imperatives largely concern deep learning systems.

[2] Throughout this manuscript, when I refer to AI systems, I mean deep learning systems specifically.

[3] I accept that informed consent preserves and upholds certain bioethical values like personal autonomy and non-domination. I also assume that existing legal requirements and guidelines adequately secure informed consent. I set aside broader questions and legitimate concerns about the fundamental ethical value of informed consent or the means by which it is granted.

Supposedly, the use of black box AI systems in clinical healthcare both violates patients' rights to informed consent and precludes the ability of experts to secure informed consent from patients. Therefore, the use of epistemically opaque AI in healthcare appears to be unethical.

1.2 The proposed solution

Such ethical concerns have led to growing legal and ethical demands for the explainability[4] of ML algorithms used in decision-making, what I call the Explainability Imperative (EI):

> **Explainability Imperative:** Explainability is a necessary requirement for the ethical deployment of AI in safety-critical domains like healthcare and criminal justice.

EI reflects a popular belief that "In order to know if the model is [fair, safe, reliable, robust, sensible], we need to know how the model makes its decisions. Or perhaps we need to know why the model makes its decisions" (Lipton [18], 1). In other words, ethical AI requires explainable AI. With respect to AI in healthcare specifically, Char et al. state, "ML [machine learning] systems in medicine must have an explainable architecture, designed to align with human cognitive decision-making processes familiar to physicians, and directly tied to clinical evidence" (Char et al. [3], 6). An article published in Nature Medicine echoes this sentiment stating, "AI algorithms used for diagnosis and prognosis must be explainable and must not rely on a black box" [16]. In short, despite various challenges to the primacy of explainability,[5] it remains widely accepted that the ethical use of AI in healthcare requires explainable models. Conversely, the use of epistemically opaque or "black box" AI systems in healthcare is taken to be unethical because it violates foundational principles of medical ethics, such as transparency, and the legal requirements for informed consent.

1.3 Thesis

Although it is widely accepted that informed consent requires the explainability of medical AI systems, I argue that this view relies on a misunderstanding of what information is required for informed consent. A deeper look at the Information Requirement in ordinary clinical settings reveals that, to the contrary, explainability is not necessary. My argument against the Explainability Imperative for AI systems is based on the U.S. legal requirements for informed consent for human agents. This is because informed consent is the primary legal mechanism for upholding and preserving foundational principles of medical ethics like patient autonomy. While it is possible for ethical action to exceed legal requirements and conditions, meeting accepted legal criteria is usually prior to and necessary for ethical action. It strikes me as implausible that a medical procedure is ethical to perform, despite not meeting the most basic legal requirements of informed consent like being informed of relevant and substantial risks associated with the medical procedure.

Some may very well believe that our ethical standards should surpass legal requirements when it comes to informed consent. For example, Ursin et al. [30] expressly frame their proposal as describing ethical requirements for informed consent "that go beyond legal requirements" (Ursin et al. [30], 2). In other words, they endorse a maximal interpretation of the Explainability Imperative and use that interpretation to define the ethical rights and responsibilities of patients and health care providers, respectively.

The risk with using maximal interpretations of explainability to define ethical standards for informed consent are threefold. First, such requirements would effectively render most practitioners "unethical" for failing to disclose, for example, "the risk of cyber-attack" when using AI to diagnose diabetic retinopathy (Ursin et al. [30], 3)—information that most health care providers do not and would not normally disclose or be required to disclose. The reliance on maximal, and perhaps idealistic, standards leads to the predictable result that most of human activity fails to meet such standards. Yet, we do not want to characterize most health care providers as "unethical" for failing to provide such information—especially when it is not legally required. Ascriptions of unethical behavior usually refer to a violation of a legal requirement or responsibility.

Second, while maximal interpretations may be useful for defining philosophical ideals, they are limited in their utility for providing practical guidance to healthcare providers seeking to uphold their professional responsibilities.

Third, maximal interpretations risk setting up ideals for explainability that do not and cannot inform practice or

[4] I will primarily use the terms explainability and transparency, though the literature vacillates between a family of terms including: transparency, interpretability, surveyability, explicability, etc. There have been serious efforts to distinguish between these different terms [19] and to highlight the importance of not conflating these terms (Herzog 2022). For my purposes, they will function in similar ways—to either mitigate or eliminate epistemic opacity in AI applications. As such, I will follow scholars, like Ursin et al. [31], who include these terms under the umbrella concept of "explicability" or "explainability" and focus on "explainability". The finer distinctions are valuable but beyond the scope of my more general argument that epistemic opacity does not violate patients' right to informed consent.

[5] See for example: London [18], Zerilli et al. (2019), and Duran and Jongsma (2021).

patient-centered care (see [6]; Bjerring and Busch 2021). It is not cleared how detailing the risks of cyber-attack for a given AI system used in diagnostics would facilitate a patient's medical decision-making or a health care provider's ability to communicate information that is material to a patient's decision to undertake one medical treatment or another. Rather than comparing AI systems' explainability to some idealized diagnostic process, AI systems' explainability (or lack thereof) should be evaluated based on practical considerations [14]. Given that my aim is to define ethical standards that are informative and useful to practitioners and patients, I will be arguing against the ethical demand for explainability through the lens of existing legal requirements in the U.S. I take the legal requirements for informed consent as defining a minimal criteria for ethical healthcare and providing practical guidance in health care contexts.

If the use of epistemically opaque AI systems prevented one from meeting legal requirements for informed consent, then it would clearly be unethical to use such systems in clinical settings. But, as I will show, the use of epistemically opaque AI systems does not violate existing legal—and thus minimal ethical—requirements for informed consent. Moreover, most scholars mistakenly locate the source of ethical concerns with medical AI in the "essential" or "deep" epistemic opacity of medical AI systems. However, as I will show, epistemic opacity is a wide-spread and trenchant feature of expertise, both human and machine, and clinical reasoning. However, the epistemic opacity of human experts does not pose barrier to meeting legal and ethical requirements for informed consent. Thus, the Explainability Imperative effectively demands a higher standard of transparency for medical AI systems than for human experts in ordinary medical contexts. This double standard is not rooted in existing legal requirements for informed consent and must be supported on independent grounds.

1.4 Roadmap

I will start the discussion by examining legal standards of informed consent, and then move on to more general philosophical considerations.

In the second section, I will describe the legal requirements for informed consent, how informed consent is documented by human medical professionals, and what constitute legal violations of a patient's right to informed consent.

In the third section, I will describe how various experts have defined explainability for medical AI systems and argued for the Explainability Imperative in healthcare.

In the fourth section, I will describe an epistemically opaque AI system for diagnosing eye diseases whose use is compatible with informed consent requirements. In this case, explainability is not necessary for informed consent.

In the fifth section, I will describe and respond to possible arguments for nonetheless upholding the Explainability Imperative and holding medical AI systems to a higher standard of transparency than human experts in ordinary medical contexts.

2 Informed consent

Informed consent is a patient right and guiding legal and ethical concept in healthcare. As an instrument, informed consent is intended to enable patients to conduct their own evaluation of the risks and benefits associated with a medical intervention and to make informed decisions about their own bodies, health, and future. This patient right has a counterpart in the form of a physician's responsibility to understand and explain the risks and benefits of medical interventions in a way that a patient can understand. We can think of informed consent as a type of explainability requirement in healthcare. Just as EU residents have a right to an explanation in the form of "meaningful information about the logic involved" in automated decision-making (General Data Protection Regulation (GDPR) [8], Articles 13–15), a U.S. patient has a right to "meaningful information about the logic involved" in their medical diagnosis and treatment (Astromske et al. [1], 516).

2.1 Information requirement

While informed consent involves other legal requirements (e.g., having the capacity to make decisions and being free to do so from other's manipulation and control), I will be focusing on the information requirement for informed consent: the patient should be adequately informed of the risks and benefits of a given medical intervention. Most courts have interpreted the Information Requirement as including information about the patient's diagnosis and treatment as well as the risks and benefits of: treatment, alternative treatments and no treatment. Hence, the standard of information disclosure is called "standard risk-and-benefit disclosure" (Sawicki [25], 831).

Importantly, this is limited to information material to the procedure itself. The majority of courts in the US have "rejected the notion that the failure to disclose the physician's experience or qualification breaches the duty of informed consent, on the theory that only information about the procedure itself is material" (Cohen [4], 1435). In fact, one court rejected the patient claim of a tort when the patient specifically asked about the physician's qualification and was misled (Cohen [4], 1435). This was also the conclusion reached for cases of "ghost" surgeries, overlapping surgeries in which a different surgeon wraps up parts of an operation that was initiated and conducted by the primary surgeon.

Consider the case of Perna vs. Pirozzi in which the plaintiff (Perna) requested a specific surgeon (Pirozzi) and Dr. Pirozzi's name was on the consent forms. Nonetheless, the operation was ultimately performed by someone else and Dr. Pirozzi was not present. Post-surgical complications arose and the plaintiffs sued claiming that they had consented to Dr. Pirozzi carrying out the surgery, no one else. However, the New Jersey Supreme Court found that there was no violation of informed consent in this case (Cohen [4], 1437).

2.2 Violations of informed consent

So what *does* constitute a violation of informed consent? Legally, a breach of duty to secure informed consent requires a demonstration of:

> "(1) failure to disclose a specific risk in violation of the governing standard; (2) materialization of that risk; (3) "causation"—that is, if the risk had been disclosed, the patient, or a prudent person in the patient's position, would not have proceeded as she did; and (4) that no exception, like emergency, excuses the failure to disclose" Wilson [34]

In short, the information required concerns the risks and benefits of: treatment recommended, alternative treatments possible, and no treatment. It does not include information about the reliability of the expert, human or machine, that produces a diagnosis or recommends a treatment or even carries out the medical intervention.

Importantly, informed consent standards are only relevant for interventions that place the patient at high risk. Interventions with minimal risk—such as observational studies, diagnostic screening, or benign procedures like blood draws—usually do not require disclosure or understanding.[6] As such, we ought to distinguish between informed consent for the diagnostic screening itself, such as informed consent for the AI system used to screen for age-related macular degeneration (AMD), and informed consent for a medical intervention based on the results of the medical AI, such as retinal injections based on an AI diagnosis of AMD. Concerns about informed consent given the advent medical AI concern the latter case.

2.3 Evidence of informed consent: medical documentation

The solicitation of informed consent requires that the health care provider explain and justify their expert opinion and recommendation. In the US, this is usually enforced through the requirement for medical record documentation on the part of a health care provider[7]. This record "shall contain sufficient information…to support the diagnosis and justify the treatment, and…shall also contain evidence of appropriate informed consent for any procedure or treatment for which it is appropriate" Powell [23].

The general structure of a medical note that is common in the US and English-speaking context is captured by the acronym Subjective, Objective, Assessment, and Plan (SOAP) whereby the note begins with a report of the subjective statements of symptoms and experiences in the patient's own words followed by a report of objective findings that include data from diagnostic testing, physical exam, etc. The health care provider then provides their expert assessment of the evidence, integrating subjective and objective reports, to explain how they arrived at a diagnosis. Finally, the health care provider describes the plan of treatment which includes medications prescribed, further testing requested, recommendations, etc. In the assessment portion of the medical note, the health care provider should describe evidence of informed consent for any medical procedure or treatment proposed. This usually involves a discussion of the risks and benefits associated with different treatment options, an opportunity for the patient to ask questions and receive answers that inform their decision-making process, and the different types of uncertainty about the medical condition, diagnosis, or treatment.

The following is a sample SOAP note from ophthalmology (Fig. 1).

In this note, the doctor diagnosis the patient with macular degeneration based on the presence of macular drusen.[8] They also eliminate the possibility of alternative eye diseases such as retinal detachment, choroidal neovascularization, macular edema, and hemorrhage based on the available subjective and objective data. The doctor then recommends that the patient immediately follow up if anything changes in her vision.

The physician communicates to the patient signs and symptoms of related eye diseases that may develop (such as hemorrhage), indicating that no such signs are apparent at the current time, and (presumably) offers the patient an opportunity to ask any questions about her ongoing macular degeneration and possible future treatment options. This satisfies the information requirement for the patient, especially considering this is merely a diagnostic visit, with no substantial treatment recommended or adopted. Thus, the type of information required is minimal as compared to the type of information that must be communicated to solicit the patient's informed consent to

[6] Some exceptions exist, including diagnostic screening for STDs (such as HIV) and genetic tests which, in many jurisdictions, do require the patients' consent.

[7] Such records are examples of peer-to-peer explanations of the sort that Holzinger et al. [11] seek to define for AI in the medical domain.

[8] Drusen are yellow deposits under the retina that are made up of lipids and proteins.

Fig. 1 Source: https://www.medicaltranscriptionsamplereports.com/ophthalmology-soap-note-sample-report//

SUBJECTIVE: The patient was seen in followup today. She is a pleasant (XX)-year-old with a history of moderate dry macular degeneration. The patient reports some slight blurring of vision at near.
OBJECTIVE: Visual acuity, uncorrected, is 20/30-2 OD, 20/50 pinhole, 20/40 OS. Intraocular pressure is 19 mmHg OU. Anterior segment examination shows 2+ NS, OD and a PCIOL OS. Dilated funduscopic examination reveals macular drusen, OU. There is a large drusenoid pigment epithelial detachment in the left eye. There is no evidence of any macular edema, hemorrhage or subretinal fluid.
ASSESSMENT AND PLAN: Moderate dry macular degeneration, both eyes, with drusenoid pigment epithelial detachment, left eye. The patient appears stable from a retinal standpoint. We do not see any evidence of choroidal neovascularization, macular edema or hemorrhage. We did review signs and symptoms of these, and she does know to call immediately if she does have any distortion or vision changes. We have asked her to return in six months for a followup.

Fig. 2 Source: https://www.mtexamples.com/ophthalmology-eye-exam-chart-note-medical-transcription-sample-reports/

Ophthalmology Chart Note Sample Report #5
PROCEDURE: Laser suture lysis, left eye.
Risks and benefits of the procedure were explained to the patient. The patient expressed understanding of these risks and signed a consent form.

DESCRIPTION OF PROCEDURE: At noon, the patient was seated at the frequency doubled YAG laser and a drop of proparacaine was instilled in the left eye. The area of the flap suture was visualized with the Hoskins lens. The suture was cut with two applications of laser at 450 mW, green wavelengths, 0.1 second duration, 50 micron spot size. Conjunctiva was Seidel negative afterwards. The patient tolerated the procedure well.

significant medical intervention. As I will argue below, in this type of situation a diagnosis by a physician is equivalent in effect and value to the patient as a diagnosis provided by an AI system, if both are known to meet acceptable medical standards for reliability in their diagnoses. In both cases, the explanation consists of indicating the presence of relevant signs of macular degeneration—macular drusen.

Now, let us turn to consider a sample note for a case in which a substantial treatment is recommended, and more information must be provided to secure a patient's informed consent. In the following case, a patient is advised to undergo laser suture lysis to the left eye (Fig. 2).

In this case, there is explicit mention of a conversation between the physician and patient about the risks and benefits of the procedure. The patient indicates they understand these risks. Furthermore, it is reported that the patient signed a consent form. The information provided to the patient centers on the risks and benefits of the procedure which is in line with the "standard risk-and-benefit disclosure" that is legally required of physicians.

3 Explainable AI for informed consent

It is usually presumed that the use of an epistemically opaque AI system will leave experts and users unable to provide and receive the sort of information and assessment described in the previous section and which is legally and ethically required of physicians. This leaves human experts unable to assess the reliability of the machine's diagnoses. For example, Grote and Berens [9] claim that epistemic opacity effectively reduces the evaluation of medical AI systems by medical experts to a guessing game:

> "in the clinical setting, if an algorithm provides the clinician with a prognostic/diagnostic decision, the rationale of that very decision remains elusive…without knowing why a decision has been made, its evaluation turns into a guessing game for the clinician." (Grote 2021, 5)

But exactly what sort of explanation or information is ethically required but unavailable for epistemically opaque AI systems? Grote and Berens [9] argue that information about the system's confidence in a diagnosis is critical for informed consent, yet unavailable in the case of epistemically opaque AI systems. Schiff and Borenstein describe the type of information that is relevant, necessary, but opaque to experts and users:

> "For instance, can physicians or others understand why the AI system made the prediction or decision that led to an error, or is the answer buried under unintelligible layers of complexity? Will physicians be able to assess whether the AI system was trained on a data set that is

> representative of a particular patient population? And will physicians have information about comparative predictive accuracy and error rates of the AI system across patient subgroups? In short, if physicians do not fully understand (yet) how to explain an AI system's predictions or errors, how could this knowledge deficit impact the quality of an informed consent process and medical care more generally?" (Schiff and Borenstein [26], 140).

The inability to get this type of information from an epistemically opaque AI system is problematic because this information is taken to be necessary to both (a) justify a diagnosis or treatment plan for medical documentation records and to (b) secure the informed consent of the patient who is subject to the AI application in the course of clinical care. I now turn to describe in greater detail two accounts that argue for the Explainability Imperative in healthcare.

3.1 Patient-centered explainability

Wadden [33] argues that the ethical use of AI in healthcare, especially with respect to informed consent rights, requires both explainability and understandability. For Wadden, the black box problem in AI is fundamentally a problem of reduced understanding. An AI system is a black box insofar as it prevents "an individual person's understanding of why the system makes its recommendations" and, conversely, an AI system is less epistemically opaque and preferable insofar as it facilitates an individual's ability to "understand what is happening in healthcare" (Walden 2021, 3). This poses a legal and ethical problem because such understanding "is an essential feature of a patient's autonomy and capacity to consent to treatment" (Wadden [33], 3).

For, Wadden, AI in healthcare must be explainable and understandable to both medical experts and patients in order to uphold informed consent standards in healthcare (Wadden [33], 4). What sort of information about the AI system is needed to secure this understanding, and thus informed consent? Wadden adopts an "average patient standard" whereby

> "what information is needed will differ among individuals…the average patient would need to know enough to understand recommended courses of action, to be able to ask meaningful questions about this information, and to be able to make a decision in light of the benefits and harms of each option. If they can do this, then they understand the system enough for it to be useful" (Wadden [33], 4).

Wadden is careful to clarify that this does not mean the patient needs to have the same level of understanding as the expert,

> "Rather, like any other medical consenting practice, the patient needs to be able to understand the risks, harms and benefits, be able to rephrase and discuss the options and information given, be able to understand the outcomes and what these would mean for adjustments to life, and so on based on the explanation offered to them by a specialist. In so doing, the level of understanding required for AI procedures would be equivalent to the level of understanding we require of capable patients in any other hospital setting" (Wadden [33], 4).

The point is that explanations should enable users to understand the risks and benefits of treatment, alternative options, and the practical implications of treatment on the patient's life and future. This is in line with the type of an information a patient would need to provide their informed consent in any other healthcare setting.

3.2 Physician-centered explainability

For others, like Holzinger et al. [12], ethical AI requires user-relative understandability and causality as well as more objective measures like reproducibility and (re-)traceability. Holzinger et al. [12] are primarily concerned with how deep learning systems will interface with human experts, in this case physicians and healthcare providers who need to secure informed consent from their patients. This is clear from their highlighting "peer-to-peer explanation as it is carried out among physicians during medical reporting" as the type of explanation they are interested in for explainable AI in healthcare (Holzinger et al. [12], 2). Their analysis reflects their concern with cases where deep learning AI systems complement or overrule human physicians, human experts' abilities to understand AI systems, and what explanations of an AI system are adequate for the purposes of human experts. In other words, they are interested in characterizing what it is for an AI system to be explainable to a human expert.

It is within this context that Holzinger et al. argue for the Explainability Imperative in the medical domain: "medical professionals must have a possibility to understand how and why a machine decision has been made" (Holzinger et al. [12], 2). For Holzinger et al., explanation "means to provide causes of observed phenomena in a comprehensible manner through a linguistic description of its logical and causal relationships" so that explanations facilitate causal reasoning and understanding (Holzinger et al. [12], 4). In other words, Holzinger et al. require a causal model of the AI system's diagnostic reasoning with a mapping from the AI system's features to a causal model that human agents can understand. Thus, explainability "in a technical sense highlights decision-relevant parts of the used representations of

the algorithm and active parts in the algorithmic model, that either contribute to the model accuracy on the training set, or to a specific prediction for one particular observation" (Holzinger et al. [12], 3). Therefore, explanations must provide information about: the causes of a phenomenon, how the algorithm represents the data/phenomenon internally, what features of the representation contribute to the AI model's final prediction or improve model performance, as well as the causal/logical relationships between the phenomenon modeled, the AI's representation of the phenomenon, and the final prediction. If we are considering an AI model for the detection of an eye disease, like diabetic retinopathy (DR), then explanations must provide information about the known causes of DR, how an AI system represents DR internally (via neural network layers, nodes and links that define and use features of training dataset), the salient features of the input data that contribute to the final prediction and which improve the model's predictions and overall performance. This information is needed for robust causal explanations and, as a result, human expert understanding (Holzinger et al. [12], 3). A good explanation measures high in "causability" which means it produces causal understanding in a human expert.

Human experts' causal explanation and understanding of an AI system is deeply connected to human experts' ability to retrace and reproduce the AI decision-making process. Holzinger et al.'s "retraceability" requirement entails that "human experts must still have a chance, on demand, to understand and to retrace the machine decision process" (Holzinger et al. [11], 3). Through retraceability, human experts can meaningfully interact with the machine's thought process: "humans must be able to understand, to re-enact, and to be able to interactively influence the machine decision process" (Holzinger et al. [11], 16). This calls for the development of explainable AI methods and models "necessary to reenact the machine decision-making process, to reproduce and to comprehend both the learning and knowledge extraction process" (Holzinger et al. [12], 2).

3.3 Summary

Regardless of how narrowly or how broadly explainability is construed, there is a clear demand for explainable, and thus ethical, AI in healthcare. On both accounts, explainability is necessary for the ethical use of AI in healthcare. On Wadden's account, explanation is necessary (though not sufficient) for understandability of both patients and physicians where understandability requires a patient to have information about risks/benefits, consequences of, and alternatives to the medical treatment proposed. On Holzinger's account, explainability is necessary and defined by reference to causal information. Such explanations are then evaluated by their ability to facilitate expert understanding, reproducibility, and effective interaction with the AI system. This echoes Schiff and Borenstein's claim that "for an informed consent process to proceed appropriately, it requires physicians to be sufficiently knowledgeable to explain to patients how an AI device works, which is rendered difficult by the black-box problem" (Schiff and Borenstein [26], 140, emphasis added). Thus, physicians' robust causal knowledge of the AI system (as Holzinger et al. [12] require) enables physicians to uphold their responsibilities for securing informed consent from patients and to facilitate patient understanding (as Wadden [33] requires). What unites these different approaches to explainability in healthcare AI is that they are intended to avoid or mitigate the epistemic opacity of the AI system at hand.

A natural follow-up question is: how does a human diagnosis, with its associated assessment and report, compare with that of an AI system? To address this question, I now turn to describe a 'black box' deep learning system used to diagnose diabetic retinopathy and other eye diseases based on medical images. I will first describe how the system functions and the type of explanations it can generate. I will then compare the assessment provided in the case of human experts (in Sect. 2) with the sort of assessment this epistemically opaque AI can produce. I will show that an epistemically opaque AI systems can provide post-hoc explanations that justify their diagnoses and/or recommendations just as human experts do. I conclude that the use of such systems does not violate a patient's right to informed consent.

4 An example of ethical and opaque medical AI: Ting et al. [28]

Ting et al. [28] develop a deep learning system (a convolutional neural network) to detect and diagnose referable diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, retinopathy of prematurity (ROP), etc. The deep learning system takes a patient's retinal image as input data and effectively compares the image to a representation of eye disease that the system learned from multiple large datasets of images from patients with and without three diseases: diabetic retinopathy, glaucoma, and macular degeneration. The system was then tested by comparing its diagnoses on new data and images with those of professional graders (retinal specialists, general ophthalmologists, trained optometrists) for both the Singapore Integrated Diabetic Retinopathy Program (SIDRP) and ten external data sets (494, 661 retinal images) from 6 different countries (Singapore, China, Hong Kong, Mexico, USA, and Australia) over a 5-year period. Ting et al.'s [28] DLS exhibited high and clinically acceptable performance in terms of sensitivity and specificity (> 90%) for identifying DR and related eye diseases using retinal images from diverse populations

Table 4. Primary Validation Dataset Showing the Area Under the Curve, Sensitivity, and Specificity of the Deep Learning System vs Trained Professional Graders in Patients With Diabetes, SIDRP 2014-2015, With Reference to a Retinal Specialist's Grading

	Value (95% CI)[a]		
	Deep Learning System	Trained Professional Graders	*P* Value[b]
Referable diabetic retinopathy[c]			
Area under the curve[d]	0.936 (0.925-0.943)		
Sensitivity, %	90.5 (87.3-93.0)	91.2 (88.0-93.6)	.68
Specificity, %	91.6 (91.0-92.2)	99.3 (99.2-99.4)	<.001
Vision-threatening diabetic retinopathy[e]			
Area under the curve[d]	0.958 (0.956-0.961)		
Sensitivity, %	100 (94.1-100.0)[f]	88.5 (75.3-95.1)	<.001
Specificity, %	91.1 (90.7-91.4)	99.6 (99.6-99.7)	<.001

[a] Eyes were the units of analysis (n = 35 948). Asymptotic 95% CI was computed for the logit of each proportion and using the cluster sandwich estimator of standard error to account for possible dependency of eyes within each individual (exception, sensitivity calculation for the deep learning system).

[b] *P* value was calculated between the deep learning system vs trained professional graders using the McNemar test.

[c] Referable diabetic retinopathy was defined as moderate nonproliferative diabetic retinopathy, severe nonproliferative diabetic retinopathy, proliferative diabetic retinopathy, diabetic macular edema, and ungradable eye.

[d] Cluster-bootstrap, biased-corrected 95% CI was computed for each area under the curve, with individual patients as the bootstrap sampling clusters.

[e] Vision-threatening diabetic retinopathy was defined as severe nonproliferative diabetic retinopathy and proliferative diabetic retinopathy.

[f] Exact Clopper-Pearson left-sided 97.5% CI was calculated owing to estimate being at the boundary.

Fig. 3 A summary of Ting et al.'s deep learning system comparing its performance relative to the performance of professional graders

(Ting et al. [28], 2211). The DLS was comparable to and, in some cases, outperformed[9] professional graders' performance in detecting and classifying eye diseases (Ting et al. [28], 2220). The results are summarized in the table below (Fig. 3).

4.1 Limitations: epistemic opacity

However, a key limitation of this study is that

> "the DLS uses multiple levels of representation to analyze each retinal image without showing the actual diabetic retinopathy lesions (e.g., microaneurysms, retinal hemorrhages). These data points can possibly be the shape or contour of the optic disc or tortuosity or caliber of the retinal vessels [rather than lesions]. Such black-box issues may have an effect on physicians' acceptance for clinical use" (Ting et al. [28], 2221–2222).

Medical experts detect and diagnose DR by looking for specific retinal features such as retinal lesions, cotton wool spots, or hemorrhages in the retinal image. In contrast, deep neural networks like the one used in this study are trained on millions of retinal images and pick up on the *implicit* features of the data which are most predictive of diabetic retinopathy rather than identifying *explicit* clinical features of the retinal image which physicians usually identify, share, and use to diagnose a patient. Such models are black boxes. The worry is that neural networks may use certain features as predictors of DR that are not clinically or causally connected to DR, such as "the shape of contour of the optic disc" or the race of the patient population (different background color of the retina) or variability in pupil dilation.

Subsequent studies conducted by Dai et al. [5] expressly frame their work as improving upon the system developed by Ting et al. [28]. Ting et al.'s deep learning system was "trained directly end-to-end from original fundus images to the labels of DR grades" and, as a result, "might fail to encode the lesion features due to the black box nature of deep learning" (Dai et al. [5], 7). In contrast, Dai et al.'s model leverages lesion features as prior knowledge which make their system interpretable, and thus more transparent, in comparison to the "black box nature" of Ting et al.'s study (Dai et al. [5], 7).

To summarize, Ting et al. acknowledge that the black box nature of their system is a central limitation of their study and subsequent studies, like that of Dai et al. [5], explicitly endorse their systems as superior because of their improved explainability in comparison to the black box deep learning system developed by Ting et al. [28]. Thus, the epistemic opacity (or transparency) of a medical AI system is taken to be a key limitation (or advantage) of an AI system for use in clinical settings.

4.2 Machine malpractice?

Given the epistemic opacity of Ting et al.'s DLS, would its deployment in clinical practice constitute a violation of a patient's right to informed consent? Or undermine a physician's ability to secure a patient's informed consent? A concrete way to explore this question is to compare the type of information that the epistemically opaque AI system can provide to justify its diagnosis with (1) the type

[9] Ground truth was based on the determination of retinal specialists with over 5-years experience in diabetic retinopathy grading. The DLS was shown to outperform two trained senior professional graders (non-retinal specialists) with over five years experience by reference to the grading of the retinal specialists. For example, two trained graders and the DLS were given a retinal image to grade. The DLS outperformed the two trained graders because it gave the correct grading more often than the trained graders, where the correct grading was determined by the retinal specialists' grade.

of information that a human expert provides in the assessment portion of the medical documentation record and with (2) the type of information legally required for informed consent.

I take it that there is no prima facie reason why an AI system that is used either autonomously or as a complementary tool would preclude the ability to populate either the subjective or objective portion of the SOAP medical note. The subjective part merely reports the patient's subjective statements of symptoms and experiences. This can still be provided, reported, and encoded as another form of input data. The objective portion of the note reports the various objective data like the results of various tests and medical imaging etc. In this case, it would include the various input data like the retinal images used by the AI system to classify the eye disease. Again, the use of an AI system does not fundamentally change the contents or possibility of populating this part of the note. Finally, the plan/treatment is to be proposed by the healthcare provider given the diagnostic results of the AI system. This is unchanged whether the diagnosis is reached by a machine or human expert. If you are diagnosed with macular degeneration, the recommended treatment is retinal injections regardless of who made the diagnosis. The question of the reliability of the diagnosis is a separate question that is addressed by the assessment portion of the note in which the physician or expert system justifies the diagnosis.[10] I will assume for the moment that that question is answered and focus instead on comparing the assessment portion of the medical note for a human vs. machine expert.

Recall that in the sample ophthalmological SOAP note discussed earlier (Sect. 2.3), the doctor diagnosed the patient with macular degeneration in both eyes. In the assessment portion of their note, they justify their diagnosis in two ways: (1) they refer to the presence of macular drusen in the medical images provided and (2) the absence of indicators for alternative eye diseases like macular edema, hemorrhage, etc. This assessment is a form of post-hoc reasoning used to support the treatment plan by highlighting the salient evidence for the diagnosis in the medical image—the presence of drusen in a retinal image—and ruling out alternative diagnoses. This type of information is also taken to satisfy the information requirement for informed consent.

Can a similar justification be provided by an epistemically opaque AI system like that developed by Ting et al. [28]? Yes.

The AI system developed by Ting et al. [28] classifies a retinal image as displaying age-related macular degeneration (AMD) based on the following definition:

> "Referable AMD was defined as the presence of intermediate AMD (numerous medium-sized drusen, 1 large drusen ≥125 μm in greatest linear diameter, non-central geographical atrophy, and/or advanced AMD central geographical atrophy or neovascular AMD) according to the Age-Related Eye Disease Study grading system" (Ting et al. [28], 2221).

Therefore, if the AI system detects "referable AMD" then the "explanation" would refer to the relevant indicator or evidence for this diagnosis, i.e., the presence of "numerous medium-sized drusen". Moreover, the AI system also checks and rules out alternative eye diseases like glaucoma. The final justification of the AI decision or diagnosis has two elements: (1) the presence of some indicator connected to the disease state and (2) the elimination of alternative possible diagnoses.

Can a physician who uses this epistemically opaque (characterized as such by both its developers and competitors) AI system nonetheless meet the information requirement? Yes. First, the type of justification that is available from this sort of system is analogous to the sort of justification that a human expert may provide in their medical note, namely: the presence of drusen and elimination of alternative possible diseases. The diagnosis of macular degeneration by an AI system would be justified by reference to the definition of AMD given in the assessment or explanation. For example, if the diagnosis was of "referable AMD" then the assessment could describe the relevant indicator or evidence for this diagnosis as the presence of "numerous medium-sized drusen". This indicator, the presence of drusen, justifies the machine's diagnosis just as it did in the first assessment provided earlier by a human expert. The justification of the diagnoses is similar for human and machine experts. This is contrary to the claims of Muller (2021) that informed consent requires the justification of a judgment and that such a justification is impossible for epistemically opaque AI applications.

Furthermore, this AI system also checks and rules out alternative eye diseases like glaucoma. Again, it is possible to provide the same type of post-hoc justification for the decision of an epistemically opaque AI system as the types of justification provided by human experts. Notably, we do not have information of how the AI system represents the disease internally or the step-by-step process by which it reached the diagnosis. But such information is neither provided by nor required for the human expert making the same diagnosis. We do not have a description of the internal mental representation of disease by the human physician or a step-by-step description of their clinical reasoning. While it is not possible to provide the sort of detailed information and causal explanation required by proponents of explainable AI in healthcare (such as Holzinger et al. [12]), it is

[10] This is in line with Muller's (2021) construal of the explainability required by AI regulations (like GDPR) as reflecting demands for justification first and foremost.

nonetheless possible to explain and understand the diagnosis of an epistemically opaque AI system to the same extent that it is possible to explain and understand the diagnosis of a human expert.

Importantly, both medical notes presented earlier (Sect. 2.3) are taken to provide evidence of informed consent to different degrees reflecting the significance of the medical intervention (diagnostic or surgical procedure). In neither case is there any report of the probabilities related to confidence levels, accuracy and error rates for the diagnosis rendered or the efficiency and complication rates for the treatment recommended. Human experts do not, and are not required to, provide reports of probabilistic confidence intervals for their diagnoses.

Grote and Berens [9] have argued that information about the system's confidence in a diagnosis is critical for informed consent, yet unavailable in the case of epistemically opaque AI systems. Namely, Grote (2021) objects that AI systems are (currently) unable to report a probability distribution of the model's uncertainty when it comes to a specific diagnosis. Ideally, a trustworthy AI system "states that a given image shows disease X, while reporting an uncertainty of 0.3" (Grote 2021, 337). Such precise probabilistic reports of uncertainty (and thus confidence) are required for trustworthy medical AI systems. This requirement is rooted in a certain assumption about human clinical reasoning. Namely, Grote and Bergen's [9] appear to envision that when a clinician is, for example, making a diagnosis of a skin disease, "After assessing the evidence, she concludes that the patient has disease *x*, where she has a confidence of 0.8 in her proposition" (Grote and Bergen's [9], 207). However, the epistemic opacity of an AI system poses a problem because "As the patient is not provided with sufficient information concerning the confidence of a given diagnosis or the rationale of a treatment prediction, she might not be well equipped to give her consent to treatment decisions" (Grote and Bergen's [9], 208).

However, we do not require, and often times do not receive, such precise probabilistic reports of confidence or uncertainty levels from human experts. For example, when a health care provider diagnoses a patient with monkey pox, the patient may ask how sure the provider is. In response, the healthcare provider may respond "It is more likely that you have monkey pox than some other disease" or "It is more likely that you have monkey pox than not". In other words, the confidence level, if reported, is reported in general and relative terms where the contrast class is alternative diseases or no disease. The health care provider does not, and usually cannot, provide a probabilistic report such as "My uncertainty in this proposition describing your diagnosis of monkey pox is 0.2 and my confidence is 0.8". If we are to set similar standards for AI systems, then we should require that an AI system report whether it is more confident in one diagnosis than alternative diagnoses or no disease—not provide a probability distribution of its uncertainty. In fact, such precise probabilistic reports of confidence and uncertainty are of limited utility given that, as Grote and Bergen's [9] point out, "the confidence score reported by the algorithm and that of a clinician may be given on the same scale… but mean very different things…and human self-reported confidence can be quite substantially differ [sic] from mathematical notions of confidence" (Grote and Berens [9], 207).

Yet, if we consider qualitative and relative likelihoods of diagnoses, AI systems can provide such information, despite their epistemic opacity. For example, the DLS developed by Ting et al. [28] is trained on images displaying a variety of eye disease and no disease. The DLS then diagnoses eye disease by first detecting a feature that indicates the presence of eye disease and then identifying the disease by ruling out alternative eye diseases. Given the DLS's diagnosis, it is possible to report that the DLS is both more confident than not that an eye disease is present in a retinal image and the DLS is more confident that it is one disease than some other disease. This report is based on the DLS's superior performance in detecting and diagnosing eye diseases as measured by sensitivity and specificity. Through testing and validation, Ting et al. [28] demonstrate that, given a DLS's diagnosis of eye disease X, it is more likely that X is present than absent (the null hypothesis) and it is more likely that the patient has disease X (e.g., diabetic retinopathy) than some other disease Y (e.g., glaucoma). Thus, insofar as we do hold AI systems to similar standards as human experts, AI systems are able to adequately provide information about qualitative and relative confidence levels.

Similarly, Schiff and Borenstein [26] question whether human experts would be able to understand why an AI made a certain decision, whether it was trained on a representative sample, the comparative predictive and error rates of the AI system across patient subgroups, or how to explain an AI system's errors and failures. A knowledge deficit concerning these details was taken to support the claim that epistemically opaque AI systems undermine the informed consent process and medical care generally. Yet this "knowledge deficit" also occurs with respect to human experts. Human experts and patients usually do not know the error rates of their colleagues (or doctors), the exact clinical reasoning process by which they arrived at a (perhaps mistaken) diagnosis, or whether their colleague received adequate training to treat patients from different subgroups.

Furthermore, recall that this type of information about the past performance or reliability of the human medical expert carrying out a treatment (or, less significantly, diagnosis) is not legally required for informed consent (Sect. 2). Nonetheless, such information is in fact available for epistemically opaque AI systems through empirical validation and Ting

et al. [28] report this information concerning sensitivity, specificity, area under curve, etc.

Moreover, the human assessment (Sect. 2.3) does not include an explanation of the pathophysiologic mechanism underlying the disease, macular degeneration (in the first case), or an explanation of why suture lysis is necessary after trabeculectomy (in the second case). Such information is not required for either the patient or for peer-to-peer explanations of the sort exemplified by the medical note. Hence, the absence of any discussion of the step-by-step reasoning process by which the doctor reached their diagnosis. The clinical reasoning of the author is not transparent and transparency is not needed to satisfy the requirements of a peer-to-peer explanation or to secure informed consent. There is no linguistic argument provided to justify the diagnosis apart from noting the presence of macular drusen and having eliminated alternative eye diseases like a retinal detachment.

Some have claimed that "the level of understanding required for AI procedures would be equivalent to the level of understanding we require of capable patients in any other hospital setting" (Wadden [33], 4). However, we do not ordinarily require human experts to provide explanations that describe "every relevant epistemic step of the process by which an input [retinal image] is transformed into an output [diagnosis]" (Humphreys 2009, 618) or to make transparent each step by which they arrived at a diagnosis. Thus far, it seems that (contrary to popular belief) epistemic opacity is not an obstacle for ethical AI based on at least one metric of ethical medical practice—informed consent—and it is possible to secure informed consent with the use of an epistemically opaque AI system. In other words, explainability is not necessary for informed consent.

5 Reconsidering the EI

So why hold machine experts to stricter standards of explainability (and transparency) than human experts? In this section, I will evaluate potential arguments for nonetheless upholding the Explainability Imperative and adopting stricter explainability standards for machine experts than human experts in healthcare.

The primary argument for stricter transparency requirements is that, given their epistemic opacity, AI systems can fail in unexpected ways. This is especially troublesome in high-stakes contexts (like healthcare) where errors can lead to death or significant and irreversible harm to a patient.

I will first argue that empirical validation techniques suffice to address this concern. I will then argue that such risks of error are on par, if not higher, for human experts given the opacity of human minds and rates of human error and bias. Yet, there is no corresponding Explainability Imperative in the ordinary clinical context.

5.1 Empirical validation for informed consent

Concerns about AI systems using unreliable correlations for prediction are legitimate and Caruana's [24] example of an AI system that associated asthma with lower probability of death (POD) should give us pause. Yet, there is no reason to believe that this error could not have been probed and discovered through empirical validation techniques. It is often presumed that a model would learn such (unreliable) correlations, be found adequate on common measures of predictive performance and testing, and wreak havoc in real-world applications. However, had the model described by Caruana et al. [24] been tested across datasets from different hospitals and patient populations, asthmatic and otherwise, it would have clearly been found inadequate. Its predictions of lower POD for asthmatic patients would be at odds with observational data and expert knowledge of higher POD for asthmatic patients who contract pneumonia.

Human experts can rigorously test how successful a system is at correctly diagnosing human patients, across a range of conditions, and evaluate the robustness of the system's performance given changes to key parameters. In doing so, they discover and highlight the fragility of systems whose predictions are based on unreliable indicators and guide the development of more robust models. Epistemic opacity does not preclude the possibility of this type of empirical validation and testing.

Consider the case of Ting et al. [28] in which the risk of different biases is effectively probed through empirical testing despite the epistemic opacity of the deep learning system at hand. They evaluate their neural network by comparing its diagnoses with those of professional graders (retinal specialists, general ophthalmologists, trained optometrists) for both the SIDRP and ten external data sets (494, 661 retinal images) from 6 different countries (Singapore, China, Hong Kong, Mexico, USA, and Australia) over a 5-year period. In doing so, Ting et al. [28] effectively test the generalizability of their neural network model to different ethnicities, retinal images produced by different cameras, and against real-world DR screening programs—not just publicly available datasets (Ting et al. [28], 2214). In other words, they effectively investigate potential sources of error (such as racial bias) and test the robustness and generalizability of their epistemically opaque system against changes in cameras, demographic subgroups, dataset type and source, etc. In doing so, they are able to effectively rule out certain sources of error or bias and characterize the range in which their system is effective. A system that can effectively generalize across such diverse populations is, by definition, reliable.

5.2 Information for explanation vs. consent

In fact, this information from empirical validation is exactly the type of information needed to characterize the risk and reliability of using such a system—the type of information that is actually needed to facilitate patients' decision-making and informed consent. Patients need to know the risks (a system's error rates) and benefits (a system's accuracy) to assess the reliability of an AI system's diagnosis to support making potentially life-altering medical decisions. The epistemic opacity of the system does not prevent experts from either discovering or communicating this information to patients even if they do not understand the underlying causes or reasons why a system succeeds or fails. While detailed "under the hood" causal explanations that answer how or why questions (of the sort required by Holzinger et al. [12]) may be necessary for experts' purposes of error-correction and model development, such information is beyond what is needed for patients' purposes and informed consent.

The information requirement for informed consent can nonetheless be satisfied without detailed information about how or why the system failed. Note that causal explainability, in terms of technical explanations of why an AI system reached its decision, does not provide the information needed for informed consent, namely information about risks and benefits of treatment, alternative treatment options, treatment consequences, etc. Yet, this is precisely the sort of information that empirical validation and testing can provide, without explainability.

Arguably, such explainability is also beyond what is legally required for informed consent in ordinary medical settings. Recall that the information requirement for informed consent does not include information about the physician's reliability, error rate, past experience, etc. There is a significant body of case law demonstrating that such knowledge about the performance and skill of the diagnostic agent or surgeon is unnecessary for informed consent, even when a patient signs an informed consent form that specifies the agent. Rather, the risks and benefits that must be known pertain to the treatment and only insofar as it is a substantial medical intervention (which excludes most diagnostic screening technologies). While a patient needs to know the risks and benefits of an eye surgery or treatment (like retinal injections) which a medical expert recommends based on a diagnosis made using an AI system, the patient does not need to know the risks and benefits pertaining to the agent, human or machine, making the diagnosis. This is an important distinction. Information about the reliability or potential error of a machine—which Schiff and Borenstein [26], Wadden [33], and Holzinger et al. [12] demand—lies beyond the scope of the legal information requirement for informed consent. All that is legally required is information about the risks and benefits of treatments which a patient undergoes based on that diagnosis. A healthcare provider can provide information about the risks and benefits of treatment that is recommended based on AI diagnostics or even that which is directly recommended by an AI system. Epistemic opacity does not pose an obstacle to this type of information.

Furthermore, there is no reason to believe that such systems cannot autonomously describe the known risks and benefits associated with a given treatment recommendation. In fact, a large part of the attraction to medical AI systems is the potential for personalized medicine using AI systems that can identify risks and benefits based on individual patient data. AI systems can be designed to recommend treatments and include a description of specific contraindications for a treatment given patient data, potential adverse reactions related to treatment, a comparison of treatment options, a patient's risk of disease, etc. [29]. In other words, AI systems can provide information about risks and benefits related to treatment options that are tailored to the individual's health data—just the sort of information needed for informed consent.

5.3 A double standard

XAI proponents who define explainability in terms of transparency or intelligibility are holding machine experts to a stricter standard than human experts. This double standard is a problem because we generally believe that "frameworks established for artificial intelligence (AI) should echo the principles and standards to which physicians are held" (McCoy et al. [20] 211, 254). Perhaps one can motivate holding AI systems to a higher standard of transparency than human experts by arguing that human experts are either more reliable than AI systems or fundamentally more transparent (at least with respect to their clinical reasoning) than AI systems. Both claims are mistaken. Human experts' reasons are also epistemically opaque to themselves, and their performance is often less reliable than that of AI systems.

Many clinicians' decisions, and the cognitive processes by which they reached said decisions, are neither transparent nor explainable to themselves, much less others. Peter Carruthers persuasively argues that "Our common-sense conception of the transparency of our own minds is illusory… On the contrary, for the most part our own thoughts and thought processes are (in a sense) opaque to us" (Carruthers [2], xii). Knowledge of our own thoughts and thought processes, e.g., deciding or diagnosing, are acquired in the same way as knowledge of the mental states and thoughts of other people. As such, the same interpretive challenges that arise for understanding others arise in cases of self-knowledge. The same skepticism about whether a person's verbal or written rationalization for a particular decision genuinely reflects the neuronal or mental pathway by which

they arrived at that decision also arises in cases of self-report on our own decision-making processes. Just as there are doubts that post-hoc explanations or explainable models of epistemically opaque AI systems may not be factive, but merely approximate (Rudin 2019), similar doubts arise with respect to the reasons experts give for their own decisions.

Carruther's argument largely centers on the opacity of our minds to ourselves. We do not have reliable first-person knowledge of our own mental states. Rather, most explanations of first-person mental states and processes are actually post-hoc rationalizations that are vulnerable to a whole host of cognitive biases. This is the case for both experts and non-experts, with experts being especially vulnerable to certain biases, like primacy bias, and certain misleading forms of rationalization.

The first-person opacity of individuals, experts included, has important consequences for second-person knowledge of expert reasoning and judgment. Consider the following example. I go see a clinical expert, *A,* and present my symptoms; they arrive at a diagnosis. I then ask expert *A* to explain their diagnosis, which they may do by highlighting some salient indicator or symptom I have. Expert *A* explains that they observed inflammation in a certain part of my body and a high white blood cell count. They subsequently eliminated some alternative possible diseases, like gastritis and menstrual cramps. Thus, they concluded that it is likely I have a case of appendicitis.

This is a form of post-hoc rationalization or an explanation of why expert *A* came to a judgment or made a decision. However, it may very well be the case that expert *A* did not actually engage in complex analytic reasoning about the probabilities of different symptoms or diseases, eliminating alternative possibilities, etc. to arrive at their diagnosis, as they explained. Instead, they relied on their intuitive reasoning process and arrived at their diagnosis based on an intuitive judgment about the harms of delaying treatment vs. administering unnecessary treatment. Moreover, this intuitive judgment may be largely based on a primacy bias because the experts' last patient experienced severe complications associated with delayed treatment for appendicitis. The regret associated with that decision is a key driver for the current diagnosis and treatment recommendation. Therefore, an explanation that provides patients with the analytical and clinical reasons for a judgment (such as the results of a given medical test) is factually inaccurate.

Keep in mind that expert *A* is not intending to deceive anyone by failing to report the intuitive basis of their judgment or their susceptibility to primacy bias. Expert *A* themselves does not realize that such affect-based subjective judgments are the actual reasons behind their clinical judgment. This is because, as Carruthers argues, they have no more access to their own mental states or processing than they have into someone else's mental state and processes. Just as we cannot know how or why an expert arrived at a decision, they cannot know how or why they themselves arrived at a decision despite it being their own reasoning and judgment in question. Both the expert's and the patient's knowledge of how or why the expert reached their diagnosis is unreliable. This because the explanation provided is based on expert *A*'s first-person knowledge and report of how and why they reached a clinical judgment and this report does not track the actual mental states and processes involved in expert *A*'s clinical diagnosis. Thus, the opacity of an expert's mind to themselves poses a challenge for the explainability of an experts' judgments to others, both patients and peer experts.

This example makes clear how Carruther's account about first-person access to mental states bears on explainability requirements for informed consent between two people. A patient or peer expert bases their second-person knowledge of human diagnostic reasoning on a medical expert's first-person knowledge or self-report of how they reason from a clinical indicator to a diagnosis. After all, it is not possible for patients to objectively identify, much less trace, a path from the expert's initial visual input to the mental state representing a final diagnosis in a given expert's mind. Rather, patients often rely on experts' report of their reasons for a given diagnosis. These reasons are taken to provide an explanation of the diagnosis. This explanation is satisfactory for informed consent to the degree that it includes information that is material to the patient's decision-making. It is the presumed reliability of human experts' self-report of reasons for a diagnosis that grounds most demands to explainable AI. Yet, if the actual reasons behind an experts' diagnosis or judgment are opaque to the expert themselves, then the explanations an expert provides to a patient will be unreliable.

Carruthers' philosophical theory of self-knowledge, and its implications for second-person knowledge of expert reasoning, is empirically supported by studies of the explainability of human decision-making. For example, doctors undergoing residency training in radiology cannot be taught a rule-based system by which to classify medical images. Rather, much like a learning algorithm, they are exposed to many examples and told where to look and what to look for. Importantly,

> "The underlying diagnostic image patterns are too abstract and variable, and the similarity between cancerous and non-cancerous image patterns are too subtle, for rule-based decision-making. Instead, the radiologist must learn from a sufficiently large number of labeled examples as to what constitutes possible cancer and what does not. *But expert radiologists typically find it all but impossible to put into words, or explain, to their patients, insurance companies or even fellow experts, exactly how they arrived at the decision* in

> precise enough terms so that another person can arrive at the same conclusion based on the same underlying data (Sevilla and Hegde, 2017)" (Hegde and Bart [10]; emphasis added).

In other words, human experts like radiologists would be unable to provide explanations that meet Holzinger et al.'s [12] or Wadden's [33] explainability conditions for informed consent. They cannot ensure reproducibility or understandability. The thought process by which an expert radiologist decides whether a medical image displays cancer is incredibly complex and opaque to both the radiologist and others. This complexity is because

> "If we were able to open up the thought process of the typical physician deciding what surgical technique to use or whether to recommend a particular patient to go on PrEP, we would find a lot of potential inputs. The physician may be drawing on a varied assortment of vague memories from a medical school lecture, what the other doctors during residency did in such cases, the latest research in leading medical journals, the experience with and outcomes of the last 30 patients the physician saw, etc. It is beyond cavil that a physician who fails to describe each of these steps of reasoning does not violate the law of informed consent" (Cohen [4], 1442).

Given this complexity, it is impossible (as Hegde and Bart [10] found) for a physician to provide the sort of comprehensive and explicit explanation that some demand of ML algorithms. This is due, in part, to the sheer complexity of the various inputs and how they interact in the decision-making process, and in part due to the opacity of a physician's own mind to itself. A human expert would no more be able to tease apart these inputs and organize them in a manner that clearly delineates which inputs cause the final diagnosis in which way to ensure an accurate self-report of their clinical reasoning than an epistemically opaque AI system. This type of opacity extends to other types of decision-making in medicine and represents the fallacy of what Carruthers calls, the mental transparency assumption: the assumption that knowledge of our minds are transparent to themselves in a way that they are not transparent to other people (Carruthers [2], 11). As a result, knowledge of our own mental states is privileged, certain and authoritative. This assumption is clearly mistaken. The opacity of human minds may explain why, in practice, medical experts nonetheless disagree with each other about diagnostic findings, variability among experts may reflect the fundamentally interpretive nature of expert self-knowledge in decision-making.

Interestingly, human explanations of critical health care decisions made are often held to a lower standard despite being equally (if not more) fallible than the decisions reached by medical AI applications. After all, the primary motivation for adopting medical AI applications is because of their improved performance relative to human doctors. For example, Esteva et al. [7] report, ML algorithms outperform dermatologists at classifying skin cancer. Furthermore, as Grote and Berens [9] report, diagnostic errors remain quite high in the U.S. with an average of 10% of patient deaths being due to diagnostic error and 5% of health care seeking adults being subject to diagnostic error. Given the realities of clinical uncertainty, time constraints, and risks associated with collecting information and medical treatment, an ML algorithm that can outperform clinicians using complex data sets in less time, with smaller error rates, and less cognitive biases is desirable. Yet, such algorithms with their superior performance and reduced error rates are being subject to much stricter explainability requirements than those applied to fallible human experts. The key motive for the explainability imperative is fear of algorithmic error or failure that results in patient harm. However, ordinary human clinical judgment also fails and harms patients in a variety of ways (including death). Thus, AI systems should not be subject to unrealistic and strict double standards when it comes to explainability requirements. As Astromske et al. [1] put it, "advancements of medical diagnostics tools should not change the level of explanation that is currently expected from the physicians as required by law" (Astromske et al. [1], 516).

To sum, in ordinary medical contexts, explainability is not required for informed consent or otherwise. We do not need information about how a human medical expert reached a decision. We do accept post-hoc explanations of why the expert reached a decision. Similar sorts of post-hoc explanations (that do not require "opening the black box" or transparency of some sort) are available for epistemically opaque AI systems. If we are to hold AI systems to the same standards as those upheld in ordinary clinical contexts, then the requirement for explainability in healthcare AI need to be either relaxed or replaced. Alternatively, some argument needs to be made to warrant a stricter Explainability Imperative for medical AI systems than human experts. I have considered two possible arguments based on the comparable reliability and transparency of AI systems and human experts. However, these arguments fail because AI systems are just as, if not more, reliable than human experts; human experts are also fallible in high-stakes contexts; and the presumption of greater human transparency is fallacious.

6 Concluding remarks

Some have argued for strict explainability requirements for medical AI systems, what I have called the Explainability Imperative. The concern is that epistemically opaque AI

systems violate patient's right to certain types of information needed for informed consent. There are two key assumptions at play in this argument: (1) there is a distinct type of epistemic opacity associated with AI systems that calls for strict(er) explainability requirements and, conversely, (2) patients ordinarily need and do have access to detailed explanations of diagnoses for informed consent. Neither of these assumptions hold. Human medical experts and their diagnostic reasoning is just as epistemically opaque as medical AI systems. The explanations provided by epistemically opaque AI systems are on par with human explanations of medical diagnoses. In addition, more detailed explanations or transparency is not necessary to meet the information requirement for informed consent nor can human experts provide such explanations.

Importantly, it is doubtful that human experts have that sort of accurate or reliable insight into their own diagnostic criteria or can communicate such complex and implicit forms of reasoning. If human experts' own clinical reasoning was readily transparent to themselves and easily and frequently communicated to patients, then explainable AI would be a straightforward research project: simply unpack the explanations which human experts provide to patients and implement them in an algorithm.[11] For example, "if *X* is the diagnosis, provide the patient with *Y* explanation". Yet explainable AI is challenging, and complex algorithms are needed to recover the same diagnostic skills as human experts. This gives us reason to suspect that human experts' clinical reasoning involves tacit and complex cognitive criteria. If so, then human experts can no more "explain" their reasoning to the satisfaction of XAI proponents than a medical AI system. In short, it would be impossible for human agents to meet the requirements for informed consent. This is clearly not the case. Therefore, it must be that such explanations, which neither human nor machine agents can provide, are not required for informed consent.

If true, the Explainability Imperative has several important consequences. If explainability is a necessary condition for the ethical deployment of AI in clinical settings, then the use of epistemically opaque algorithms is by default unethical. However, I have demonstrated that the use of epistemically opaque AI systems does not necessarily violate a patient's right to informed consent. It is possible to meet the legal information requirement for informed consent using epistemically opaque AI systems. Moreover, empirical validation techniques can better provide the information needed for informed consent than AI explanations. I conclude that explainability is neither necessary (nor sufficient) for the ethical use of medical AI systems. As such, the Explainability Imperative is found to be, thus far, unwarranted.

[11] I thank Eric Winsberg for bringing this point to my attention.

Funding The author did not receive support from any organization for the submitted work.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

1. Astromskė, K., Peičius, E., Astromskis, P.: Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. AI Soc. **36**(2), 509–520 (2021)
2. Carruthers, P.: The Opacity of Mind: An Integrative Theory of Self-Knowledge. OUP Oxford, Oxford (2011)
3. Char, D.S., Abràmoff, M.D., Feudtner, C.: Identifying ethical considerations for machine learning healthcare applications. Am. J. Bioethics **20**(11), 7–17 (2020). https://doi.org/10.1080/15265161.2020.1819469
4. Cohen, I.G.: Informed consent and medical artificial intelligence: What to tell the patient? SSRN Electron. J. (2020). https://doi.org/10.2139/ssrn.3529576
5. Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., Liu, R., et al.: A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat. Commun. **12**, 3242 (2021). https://doi.org/10.1038/s41467-021-23458-5
6. Durán, J.M., Jongsma, K.R.: Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J. Med. Ethics **47**(5), 329–335 (2021). https://doi.org/10.1136/medethics-2020-106820
7. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist–level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017). https://doi.org/10.1038/nature21056
8. General Data Protection Regulation (GDPR). General data protection regulation (GDPR) – official legal text. Accessed Jun 3, 2022. https://gdpr-info.eu/
9. Grote, T., Berens, P.: On the ethics of algorithmic decision-making in healthcare. J. Med. Ethics **46**(3), 205–211 (2020). https://doi.org/10.1136/medethics-2019-105586
10. Hegdé, J., Bart, E.: Making expert decisions easier to fathom: on the explainability of visual object recognition expertise. Front Neurosci **12**, 670 (2018). https://doi.org/10.3389/fnins.2018.00670
11. Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B.: What Do We Need to Build Explainable AI Systems for the Medical Domain? Dec 28, 2017. https://doi.org/10.48550/arXiv.1712.09923.
12. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Data Min. Knowl. Discov. **9**(4), e1312 (2019). https://doi.org/10.1002/widm.1312. (**Wiley Interdisciplinary Reviews**)
13. Kaminski, M.E.: The right to explanation, explained. Berkeley Technol. Law J. **34**(1), 189–218 (2019). https://doi.org/10.15779/Z38TD9N83H
14. Kempt, H., Heilinger, J.-C., Nagel, S.K.: Relative explainability and double standards in medical decision-making. Ethics Inf. Technol. **24**(2), 1–10 (2022). https://doi.org/10.1007/s10676-022-09646-x
15. Krishnan, M.: Against interpretability: a critical examination of the interpretability problem in machine learning. Philos.

Technol. **33**(3), 487–502 (2020). https://doi.org/10.1007/s13347-019-00372-9
16. Kundu, S.: AI in medicine must be explainable. Nat. Med. **27**(8), 1328–1328 (2021). https://doi.org/10.1038/s41591-021-01461-z
17. Lipton, Z.C.: The Mythos of Model Interpretability. Jun 10, 2016. https://doi.org/10.48550/arXiv.1606.03490
18. London, A.J.: Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent. Rep. **49**(1), 15–21 (2019). https://doi.org/10.1002/hast.973
19. Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining Explanations in AI." In Proceedings of the Conference on Fairness, Accountability, and Transparency, 279–88. FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3287560.3287574.
20. McCoy, L.G., Brenna, C.T.A., Chen, S.S., Vold, K., Das, S.: Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J. Clin. Epidemiol. **142**, 252–257 (2022). https://doi.org/10.1016/j.jclinepi.2021.11.001
21. Ophthalmology Eye Exam Chart Note Medical Transcription Sample Reports. Accessed May 15, 2022. https://www.mtexamples.com/ophthalmology-eye-exam-chart-note-medical-transcription-sample-reports/
22. Ophthalmology SOAP Note Sample Report. Accessed May 15, 2022. https://www.medicaltranscriptionsamplereports.com/ophthalmology-soap-note-sample-report//
23. Powell, S.: "Medical Record Completion Guidelines," Aug 24, 2011, 11. https://www.mclaren.org/uploads/public/documents/macomb/documents/medical%20staff%20services/ms%20Medical%20Record%20Completion%20Guidelines.pdf
24. Caruana, R., Lou, Y., Gehrke, J., Koch, P.: Intelligible Models for HealthCare | Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–30. Sydney, Australia (2015). https://doi.org/10.1145/2783258.2788613
25. Sawicki, N.N.: A common law duty to disclose conscience-based limitations on medical practice. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 2017. https://papers.ssrn.com/abstract=3038016
26. Schiff, D., Borenstein, J.: How should clinicians communicate with patients about the roles of artificially intelligent team members? AMA J Ethics **21**(2), E138–E145 (2019). https://doi.org/10.1001/amajethics.2019.138
27. Somashekhar, S.P., Sepúlveda, M.-J., Puglielli, S., Norden, A.D., Shortliffe, E.H., Rohit Kumar, C., Rauthan, A., et al.: Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann. Oncol. **29**(2), 418–423 (2018). https://doi.org/10.1093/annonc/mdx781
28. Ting, D.S.W., Yim-Luicheung, C., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., et al.: Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA **318**(22), 2211–2223 (2017). https://doi.org/10.1001/jama.2017.18152
29. Uddin, Mohammed, Yujiang Wang, and Marc Woodbury-Smith. 2019. "Artificial Intelligence for Precision Medicine in Neurodevelopmental Disorders." NPJ Digital Medicine 2 (November): 112. https://doi.org/10.1038/s41746-019-0191-0.
30. Ursin, F., Timmermann, C., Orzechowski, M., Steger, F.: Diagnosing diabetic retinopathy with artificial intelligence: What information should be included to ensure ethical informed consent? Front. Med. (2021). https://doi.org/10.3389/fmed.2021.695217
31. Ursin, F., Timmermann, C., Steger, F.: Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? Bioethics **36**(2), 143–153 (2022). https://doi.org/10.1111/bioe.12918
32. Vincent C. Müller. 2021. "Deep Opacity Undermines Data Protection and Explainable Artificial Intelligence." In Overcoming Opacity in Machine Learning, 1–21. http://explanations.ai/symposium/AISB21_Opacity_Proceedings.pdf#page=20.
33. Wadden, J.J.: Defining the undefinable: the black box problem in healthcare artificial intelligence. J. Med. Ethics. (2021). https://doi.org/10.1136/medethics-2021-107529
34. Wilson, Robin Fretwell. 2016. The Promise of Informed Consent. Edited by I. Glenn Cohen, Allison K. Hoffman, and William M. Sage. Vol. 1. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199366521.013.53.