Philosophy & Technology (2025) 38:62
https://doi.org/10.1007/s13347-025-00860-1

RESEARCH ARTICLE

# "Does Black Box AI In Medicine Compromise Informed Consent?"

**Samuel Director**[1]

Received: 25 June 2024 / Accepted: 20 February 2025


## Abstract

Recently, there has been a large push for the use of artificial intelligence in medical settings. The promise of artificial intelligence (AI) in medicine is considerable, but its moral implications are insufficiently examined. If AI is used in medical diagnosis and treatment, it may pose a substantial problem for informed consent. The short version of the problem is this: medical AI will likely surpass human doctors in accuracy, meaning that patients have a prudential reason to prefer treatment from an AI. However, given the black box problem, medical AI cannot explain to patients how it makes decisions, yet such an explanation seems to be required by informed consent. Thus, it seems that doing what is best for patients (treatment via AI), even if patients want to permit this, might be prohibited by medicine's commitment to informed consent. Conflicts between beneficence and autonomy are not new, but medical AI poses a novel version of this conflict, because this problem is one in which even if the patient says they want to use their autonomy to receive better care, the commitment to autonomy (via informed consent) seems to block them from doing so. Given this dilemma, should we abandon informed consent, or should we not use medical AI? My thesis is that we can have our cake and eat it too; we can use opaque AI in clinical medicine and retain our commitment to informed consent, although it may require revising our understanding of informed consent. Specifically, it will require us to distinguish between two levels of consent (higher-order and first-order consent).

**Keywords** Consent · Information · AI · Explainability · Medical AI

Recently, there has been a large push for the use of artificial intelligence in medical settings. The promise of AI in medicine is considerable, but its moral implications are insufficiently examined. If AI is used in medical diagnosis and treatment, it may pose a substantial problem for informed consent. The short version of the problem

✉ Samuel Director
samjdirector@gmail.com

[1] Jepson School of Leadership Studies and Philosophy, Politics, Economics & Law, University of Richmond, Richmond, VA, USA

Published online: 13 May 2025

is this: medical AI will likely surpass human doctors in accuracy, meaning that patients have a prudential reason to prefer treatment from an AI. However, given the black box problem, medical AI cannot explain to patients how it makes decisions, yet such an explanation seems to be required by informed consent. Thus, it seems that doing what is best for patients (treatment via AI), even if patients want to permit this, might be prohibited by medicine's commitment to informed consent. Conflicts between beneficence and autonomy are not new, but medical AI poses a novel version of this conflict, because this problem is one in which even if the patient says they want to use their autonomy to receive better care, the commitment to autonomy (via informed consent) seems to block them from doing so. Given this dilemma, should we abandon informed consent, or should we not use medical AI? My thesis is that we can have our cake and eat it too; we can use opaque AI in clinical medicine and retain our commitment to informed consent, although it may require revising our understanding of informed consent. Specifically, it will require us to distinguish between two levels of consent (higher-order and first-order consent). To clarify, I intend this paper to be about the practice of using AI to diagnose patients and suggest treatment plans, not the practice of using AI to actually perform surgery (e.g. AI that uses robotics to perform surgery).

This paper will proceed in the following order. In section I, I explain in more detail the promise of medical AI. In section II, I show that in the view of many philosophers, black box AI in medicine is incompatible with our current understanding of informed consent. In section III, I consider and reject explainable AI as a solution to this problem. In section IV, I advance a solution to this problem in the form a revision to our understanding of informed consent (which I call higher-order consent) that allows us to maintain our commitment to informed consent *and* use AI in clinical settings. Although I intend this paper to be about broad issues of morality, not tethered to any particular country's legal codes, I would be remiss to not mention that my perspective on informed consent does come from an Anglo-American tradition of analytic philosophy and is perhaps most compatible with the legal systems of North America and the UK.

One final clarification: there is a distinction between consent and valid consent, where valid consent is understood to be consent that succeeds at altering the normative landscape (e.g. waiving a right against intervention, etc.). Unless otherwise stated, all uses of "consent" in this paper refer to valid consent.

## 1 The Promise of Medical AI

Although much of it is still nascent, the potential uses of AI in medicine are substantial. A review article in *Nature* identified numerous potential uses for AI in clinical practice, including "disease diagnosis, interpretation of patient genomes, treatment selection, automated surgery, patient monitoring, [and] patient risk stratification for primary prevention" (Yu et al., 2018, 720). My specific interest in this paper concerns the use of AI in diagnosis and treatment recommendation in clinical medicine.

Within diagnosis and treatment, AI has many applications. AI can be used to predict and alert doctors to the presence of hypoxaemia, which is "an unwanted

physiological condition known to cause serious patient harm during general anaesthesia and surgery" and "is associated with cardiac arrest, cardiac arrhythmias, postoperative infections and wound healing impairments, decreased cognitive function and delirium, and cerebral ischaemia" (Lundberg et al., 2018, 749). Lundberg et al. (2018) report that machine learning algorithms can enable anesthesiologists to predict hypoxaemia twice as accurately:

> The [AI] system, which was trained on minute-by-minute data from the electronic medical records of over 50,000 surgeries, improved the performance of anesthesiologists by providing interpretable hypoxaemia risks and contributing factors...Our results suggest that if anesthesiologists currently anticipate 15% of hypoxaemia events, with the assistance of this system they could anticipate 30%, a large portion of which may benefit from early intervention (749).

AI can also be used to diagnose diabetic retinopathy. The FDA approved an AI for this purpose, which has been in use at the University of Iowa (He et al., 2019, 34). Abràmoff et al. (2018) report that this AI (with 819 patients) was 87.2% sensitive (meaning able to correctly identify cases where the person does not have the condition) and 90.7% specific (meaning able to identify cases where the person does have the condition), while "previous studies have shown that board-certified ophthalmologists that perform indirect ophthalmoscopy achieve an average sensitivity of 33%, 34%, or 73% compared to the same ETDRS standard" (Abràmoff et al., 2018, 3). This is a stunning difference in abilities between human doctors and medical AI, one which both inspires confidence in the abilities of AI and some skepticism about the abilities of human doctors (in this domain). Abràmoff et al. (2018) describe this as the "ability to bring specialty-level diagnostics to primary care settings," with the possibility "to help prevent vision loss in thousands of people with diabetes annually" (1).

More generally, AI has huge implications for any area of medicine that relies on imaging to make diagnoses, which includes "radiology, ophthalmology, dermatology and pathology" (Yu et al., 2018, 722). In dermatology, "convolutional neural networks trained on 129,450 clinical images achieved dermatologist-level accuracy in diagnosing skin malignancy" (Yu et al., 2018, 723). Furthermore, "the deep-learning algorithm outperformed the average dermatologist in a comparison of the algorithm predictions and the assessments by 21 dermatologists on a set of photographic and dermoscopic images" (Yu et al., 2018, 723).[1] Some in the literature have gone as far as to suggest eliminating entire specialties. Chockley and Emanuel (2016) have said that "the ultimate threat to radiology—the one that could actually end radiology as a thriving specialty—is machine learning" (1415).[2]

[1] Here, Yu, Beam and Kohane are referring to Esteva et al. (2017).

[2] It is important to note that there are those who argue that medical AI is not as impressive as it has been claimed to be. For example, a 2019 meta-analysis "found the diagnostic performance of deep learning models to be equivalent to that of health-care professionals. However, a major finding of the review is that few studies presented externally validated results or compared the performance of deep learning models and health-care professionals using the same sample. Additionally, poor reporting is prevalent in deep learning studies, which limits reliable interpretation of the reported diagnostic accuracy" (Liu et al., 2019, e271).

The motivation for using medical AI is that it (among many other things) is cost cutting and produces better patient outcomes. If we replace certain features of clinical medicine with AI, we can eliminate the costly labor of technicians, hopefully reducing the cost of medicine. Additionally, (and more important), AI can produce better medical outcomes for patients than human doctors can. As Yu, et al. put it, "AI systems have specialist-level performance in many diagnostic tasks…[and] can better predict patient prognosis than clinicians" (722).

I'll remain neutral on what the future of AI medicine would look like in the clinic. It could be just AI, AI and humans, etc. As will become clear, my argument is contingent on a future scenario in which it is very likely that treatment via human doctor will remain an option for all who object to treatment via AI.

## 2 The Problem for Informed Consent

In this section, I argue that medical AI clashes with current doctrines of informed consent. For the sake of argument, I assume that (at some point) medical AI will surpass human doctors in many fields and will be reliably able to diagnose and suggest treatments that produce better medical outcomes for patients than human doctors could. While this is a large assumption, I make it because I want to determine if *AI itself* poses a problem for informed consent, not if the potential inaccuracy of AI poses a problem. As I will set up the problem, *if* opaque AI really is better for patients than human doctors, we need to confront how AI being a black box interferes with consent.[3] If it turns out that AI in medicine isn't actually better for patients, then we don't need to ask the question about consent. My suspicion is that most consent-based objections to AI in medicine really reduce to a worry that AI is not medically better for patients, not from an actual conflict with informed consent. Another issue I will bracket is how an AI would incorporate patient input into the process of shared decision making. For the sake of the argument, I'm assuming that medical AI will have the ability to incorporate a patient's preferences and beliefs into its treatment recommendation.[4]

To be abundantly clear, I am not claiming that medical AI is currently better than human doctors or that it will be at any near point in the future. I want to investigate the ethical implications of the possible future world in which AI does surpass doctors. I bring this up, because it is tempting to reply that medical AI isn't or won't soon be better than human doctors. That may be true, but it would not affect my argument unless there is compelling reason to think that it's extremely unlikely that AI will surpass humans in medicine.

[3] Gray Grant, Behrends, and Basl (2025) defend explainability (not just in medical AI) on the grounds that opaque AI is not always more accurate, in part because it violates procedural duties that make it likely to ignore relevant evidence and to include evidence that ought to not be included.

[4] See Holm (2022) for a discussion of shared decision making and medical AI. Recent work on using AI to predict the preferences of incompetent patients may also be relevant here; see (Earp et al. 2024).

One might wonder how issues of algorithmic bias and fairness relate to this issue. Given that the question of diagnosis and treatment is not, necessarily, a question of distributing resources across competing participants, fairness questions are less relevant to this issue. However, issues of bias are extremely relevant. An algorithm may be biased in a way that skews its recommendations along racial lines. For example, if an algorithm thinks that African Americans live shorter lives on average or have higher pain tolerance than other ethnicities, it might suggest treatments leading to worse medical outcomes for African Americans. This is an extremely important problem. However, it is beyond the scope of this paper to propose an account of how AI can avoid biased outputs. For the sake of argument, I assume that there will eventually be a way to eliminate data from algorithms that might lead to racially unequal outcomes. Additionally, it is still an important question whether a well-functioning, non-biased but opaque algorithm would violate consent.[5]

If it's true that medical AI will produce better outcomes for patients, then it seems that informed consent poses a barrier to the best outcome for the patient. Valid informed consent must be informed. But, given the black box nature (whether in principle or in practice) of AI, doctors would be unable to give the patient an explanation of why the AI made a particular diagnosis and treatment recommendation (at least not an explanation of the kind that they should be able to give if it were their own reasoning process that led to the treatment and diagnosis). Given this, the patient's right to an explanation is frustrated, she cannot make an informed decision, and thus she cannot validly consent. So goes the problem. As Schiff and Borenstein (2019) put it:

> Relevant information includes the purpose of the treatment, its potential benefits and risks, and possible alternative treatment options. Yet the novelty and technical sophistication of an AI device places additional demands on the informed consent process. When an AI device is used, the presentation of information can be complicated by possible patient and physician fears, overconfidence, or confusion. Moreover, 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...The result can be a system largely unintelligible to humans beyond its most basic inputs and outputs (139–140).

I will address the nature of the black box problem more in the next section when explainability comes up; for now, it will suffice to say that when black box AI is involved, patients are cut off from receiving an adequate explanation (according to current academic/Anglo-Saxon standards) of how their treatment decision was justified. Without this explanation, their consent is not sufficiently informed and thus not valid.

One might wonder if patients in this scenario (who know that treatment via AI is likely better for them) could waive their right to an explanation. This would be a way

[5] For recent discussions of algorithmic bias and fairness, see Milano and Prunkl (2024), Di Bello and Gong (2023), Babic and Johnson King (2023).

to say that the patient gave valid consent to the AI treatment plan. However, current conceptions of informed consent in medicine would not allow this, as we require that patients are at least able to receive an explanation of their diagnosis and treatment, if they want to. We may allow patients to waive their rights to an explanation of their treatment, but we presently require that they have the possibility of accessing this right. In other words, our current understanding of informed consent would allow a patient to waive their right to an explanation (although some question even that), but this is only in a context where this patient could, if they desired, receive an explanation. When it comes to opaque AI, the problem is not whether patients can waive a right to an explanation that they otherwise could easily access; the question is whether consent can be informed in a case where patients could not access an explanation of their diagnosis and treatment plan even if they wanted to. It is not all that controversial to suggest that, sometimes, patients may autonomously waive their right to an explanation.[6] However, that occurs in a context where a patient *could* receive an explanation if they so desired. In the context of opaque AI, no such explanation can be given, even if the patient wants one.

The previous paragraphs may have dealt too quickly with what bioethicists call the *right to not know.*[7] Although it often comes up in the context of genetic testing, the basic idea can generalize to any medical context. The idea is that patients have a right to not know certain pieces of medical information that they select to not know, for whatever reason. This fits into the dialectic, because it may be a way to say that opaque AI does not actually conflict with current practices surrounding informed consent. If patients can already waive their right to be informed, how is treatment via opaque AI different from this, as they could simply waive their right to be informed in this case as well?

The debate on the right to not know is extensive, with some arguing that patients lack this right altogether, while others argue that patients do have this right, so long as exercising it won't cause harm to third parties.[8] However, the right to not know debate comes in a context where it is a given that the patient can know the information if they want to. As mentioned in the previous paragraph, when it comes to opaque AI, this is not an option at all. In that manner, appealing to the right to not know will not make treatment via opaque AI compatible with our current standards of informed consent.

Several have argued that black box AI is incompatible with informed consent and thus that we should prohibit black box AI. Kundu (2021) claims that "relying on devices whose logic is opaque violates principles of medical ethics…Patients' autonomy and informed consent may be jeopardized if they are unable to question

[6] As Eyal (2019) says, "what we may call 'waiveable waivers' of informed consent rights ('Please spare me the full disclosure or decide in my name unless I ask to resume full disclosure or active control') seem relatively easy to accept and fairly consistent with autonomy.".

[7] The debate on the right to not know begins with Ost's 1984 article.

[8] Adorno (2004) defends the latter position, namely that there is a right to not know but that it is not absolute and is violable when doing so would protect third parties. Others (e.g. Hull 1985) in addition to Ost have argued that patients lack the right to not know.

the AI system" (1328). In a related vein, Maclure (2021) argues that explainable AI in medicine is a requirement of Rawlsian views of public reason; as she says,

> Regulators ought to impose an explainability requirement upon the (public and private) organizations that choose to delegate decision-making to AI programs when the wellbeing, rights, and opportunities of citizens are at play. For example, if IBM's Watson Health makes a surprising diagnosis or recommends an unconventional treatment, the medical team has the duty to explain to the patient why the diagnosis was made or the treatment recommended. This is required by the norm of informed consent (12).[9]

Wadden (2021) argues that opaque AI clashes with standard conditions for informed consent:

> Opaque systems...pose a particular problem for disclosure and understanding, both of which are part of the informational components of informed consent. Indeed, opacity makes it difficult for a healthcare professional to disclose why a system makes a decision, and it stands in the way of a patient understanding why the system makes its decision. A patient can reasonably be deemed competent enough to make a decision and able to voluntarily make a choice, but without proper information they cannot provide informed consent (decide and authorize). At best, in these circumstances, we might be able to claim the patient is assenting, but this does not hold the same legal or procedural weight as informed consent (96).[10]

Watson et al. (2019) endorse a similar line of argument:

> One frequently cited obstacle to machine learning's wider clinical adoption is a lack of understanding among patients and doctors about how predictions are made. This is especially true of some top performing algorithms, like the deep neural networks used in image recognition software. These models may reliably discriminate between malignant and benign tumours, but they offer no explanation for their judgments...If doctors do not understand why the algorithm made a diagnosis, then why should patients trust the recommended course of treatment? Is informed consent even possible without some grasp of how the model reached its conclusion? (2)

In the context of patient-centered medicine, Christian Bjerring and Busch (2021) argue that:

> In striving to establish a state of shared mind between patient and practitioner, exchange of information, explanation, and understanding become key components...But when we include black-box AI systems into the mix, it becomes potentially impossible to uphold this practice. The core reason is simple: since

[9] See Da Silva (2023) for a response to Maclure (2021).

[10] Wadden ultimately defends a "grey box" solution, which involves AI systems that "include some combination of opacity and transparency" (2021, 94).

> black-box AI systems do not reveal to practitioners how or why they reach the recommendations that they do, then neither can practitioners who rely on these black-box systems in decision-making—assuming that they honor their epistemic obligation—explain to patients how and why they give the recommendations that they do. Yet, for patients to make decisions in autonomous and rational ways, it is a requirement that they have "the capacity to make sense of the [medical] information presented and can process it rationally to reach a decision that furthers their health care goals" (Bernat & Peterson, 2006, p. 88).[11]

To sum this section up, patients have a right to an explanation, and opaque AI seems incapable of giving such an explanation.[12] And, even if patients have a right to not know the information that would be provided in an explanation, opaque AI presents an importantly different case in which patients can't know this information even if they wanted to. For these reasons, opaque AI conflicts with our current practices surrounding consent.[13] This means that pursuing what, by stipulation, is best for patients would require medical interventions that lack the valid consent of patients.

## 3 Explainability to the Rescue?

As it stands, it seems that if one is an anti-paternalist committed to autonomy over beneficence, then autonomy (via informed consent) is winning at the expense of beneficence (via the use of medical AI). As an anti-paternalist, I am committed to autonomy winning in a *genuine* conflict between autonomy and beneficence. Ideally, we can navigate the tension in such a way that there is no genuine conflict.

An initially tempting way to dissolve the seeming conflict between autonomy and beneficence is explainable AI. Like all terms in AI discourse, "explainability" can mean many different things; I am using the term to mean "a characteristic of an AI driven system allowing a person to reconstruct why a certain AI came up with the presented predictions" (Amman et al., 2).[14] Along similar lines, He et al. (2019) describe transparent AI as meeting the requirement that "humans should be able to

[11] Christian Bjerring and Busch (2021) 360–361. The authors are quoting from Bernat and Peterson 2006.

[12] An anonymous referee pointed out to me that the discussion so far seems to be motivated by a very strong view of how much information is necessary for a patient to be informed. Typically, we don't need doctors to tell us how they reach all of their decisions in order for us to be sufficiently informed. This is true even for stricter views about disclosure (e.g. Faden and Beauchamp 1986). I share the view that what I'm describing as the standard conception of informed consent is rarely ever lived out in practice. For that reason, I argue at length later on that the status quo in medicine supports the use of AI already.

[13] This version of the problem differs from the problem addressed by Cohen (2020); Cohen (2020) is concerned with whether patients have a right to know that AI was used in their diagnosis (1441). My question is, even if patients know that AI was used, can they have sufficient understanding of its recommendations to give informed consent.

[14] Amman et al. note that "explainability has many facets and, unfortunately, the terminology of explainability is not well defined. Other terms such as interpretability and/or transparency are often used synonymously" (2).

understand or interpret how a given technology reached a certain decision or prediction" (He et al., 2019, 31). In other words, an explainable medical AI would be able to explain, to some degree sufficient to meet the standard of explanation for the patient, how and why it reached its diagnosis and treatment recommendation. If it were truly the case that all AI could be sufficiently explainable in clinical terms that could be intelligible to both doctors and patients, this would clearly be the best system and would pose no problem for informed consent. However, there are reasons to think that this is not really a solution to the problem.

It's helpful to distinguish between two different kinds of opaqueness (the opposite of explainability) in AI: (1) *in-practice opaqueness* and (2) *in-principle opaqueness*.[15] If a system is in-practice opaque, this means that we could explain how it makes decisions if we had unlimited time and computational power; however, given the realities of the clinic and the fact that doctors are not computer scientists, these AI will almost always not be explainable to patients. Yet, they *could* be explained. If a system is in-principle opaque, this means that even with unlimited time and computational power, we *could not* understand why it reached its decision. The claim 'opaque medical AI is incompatible with informed consent' could be understood on either view of opaqueness. In-principle opaqueness (by definition) rules out explainability, so only in-practice opaque AI systems would allow patients to receive an explanation and thus be a potential solution to our problem. The question is whether explainable/in-practice opaque AI would solve the problem by allowing us to use AI in medicine while fully respecting informed consent. Unfortunately, this approach (using explainable/in-practice opaque AI) does not solve the problem.

It's likely that a medical AI that is explainable will be less accurate than an in-principle opaque AI. This means that explainability comes at the cost of high performance and accuracy, ultimately producing a worse medical outcome for patients. This claim is echoed by many working on the technical side of AI. As He et al. (2019) say, "in certain instances, enforcing transparency and interpretability can potentially result in decreased accuracy or predictive performance of a model" (2019, 31). Ursin, Timmerman, and Steger (2022) add that "there is a tradeoff between accuracy and explicability: the more explicable an AI system is, the less accurately it performs (144). As Babic et al. (2021) say, "especially in cases with massively high dimensionality—such as image recognition or genetic sequence

[15] As Nicholson Price (2017) explains: "algorithms can be opaque for multiple reasons. Sometimes, algorithms are nontransparent because, while they may rely on explicit rules, those rules are too complex for us to explicitly understand—for example, patients whose measurements place them in a particular region of n-dimensional (where n is large) characteristic-space are at a higher risk of stroke. In particular, these rules may be impossible to explain or to understand by following the process of scientific/medical discovery: mechanistic lab experiments followed by confirmatory clinical trials. Other times, the relationships used in a blackbox algorithm are literally unknowable because of the machine-learning techniques employed—that is, no one, not even those who programmed the machine-learning process, knows exactly what factors go into the ultimate decisions. A key distinguishing feature of black-box algorithms, as the term is used here, is that it refers to algorithms that are inherently black box (i.e., their developers cannot share the details of how the algorithm works in practice)—rather than to algorithms that are deliberately black box (i.e., their developers will not share the details of how the algorithm works)" (430).

analysis— limiting oneself to algorithms that can be explained sufficiently well may unduly limit model complexity and undermine accuracy" (286). In short, explainable AI will be worse for patients than in-principle opaque AI would be.[16]

For explainable AI to solve the problem, it would have to get us *both* autonomy and beneficence, but it seems that explainable AI merely presents another way of preferring autonomy *over* beneficence. Suppose that explainability (for the time being or forever) comes at the cost of accuracy and ultimately leads to worse outcomes for patients (compared to in-principle opaque AI). If that is the case, patients themselves have a prudential reason to prefer in-principle opaque models over explainable models. Unless patients want a treatment that is less likely to help them than another available option, they should prefer the in-principle opaque AI. Of course, explainable AI may still be more accurate than a human doctor. But, the point stands that patients have a prudential reason to prefer the *best* medical treatment. Again, if explainable and highly accurate AI were possible, it would clearly be the optimal option. The conflict between informed consent and medical AI would entirely dissolve if this were an option. What I'm suggesting is that absent that, we can still get the benefits of in-principle opaque AI in medicine even if we cannot achieve high functioning and explainable AI. My argument for this claim comes in the next section.[17]

## 4 Higher-Order Informed Consent as a Solution

In the previous section, I argued that explainability, if possible, will come at the cost of better patient outcomes and is thus not a solution to our problem. Ideally, a solution to the problem would achieve both a commitment to informed consent and a commitment to beneficence. Here, I'll argue that even if we are faced with either kind of opaqueness (in-practice or in-principle) in medical AI, patients can autonomously make the decision to receive treatment and diagnosis from such an AI.

So, where does informed consent fit? I defend the view that we can get the diagnostic benefits of AI while maintaining a commitment to informed consent if we allow for a kind of *higher-order informed consent*. By this, I mean a kind of consent where patients may not be informed about the first-order details of how the AI reached its diagnosis and treatment plan but are informed of higher-order information about the reliability of the AI itself.

[16] Not everyone working in this field agrees that this tradeoff obtains in reality. See Kernbach et al. (2022) (258).

[17] An anonymous referee pointed out to me that even if opaque AI leads to better diagnosis and treatment, there may be informational benefits of explainable AI which also enhance patient wellbeing. For example, if an explainable AI can give us more general information about the cause of illness, this may help us to better treat the patient. As Pruski points out (although he casts doubt on this claim), "AI explainability…may potentially advance science by revealing causal relationships that will allow us to further improve healthcare" (2024, 481). I am assuming, for the sake of argument, that opaque AI will one day be all-things-considered better for patients than explainable AI, even if explainable AI may have this advantage.

In discussions of informed consent, bioethicists have gone to great lengths to specify what information patients must know to count as being sufficiently informed. However, the philosophical discussion of informed consent has thus far not distinguished between two different kinds of information that might be relevant for a patient. What I will call *first-order medical information* is information about the doctor's process of clinical reasoning. For example, "your white blood cell count is elevated, which indicates condition X, which I will prescribe drug Y to solve, because drug Y has been shown to alleviate condition X because of reasons A, B, and C." Compare this to what I will call *higher-order medical information*, which is information about the doctor themself or the treatment itself. For example, a doctor saying only the following, without an explanation as to why or any discussion of causal mechanisms would count as higher-order information: "in all the cases of X I've treated before, 99% have gone on to live healthy lives" or "drug Y is likely to cure you, but we don't have a causal explanation for why this is the case." In canonical treatments of informed consent, this distinction between higher-order and first-order medical information is absent.[18] Consider the following authoritative documents that pertain to consent:

> **The Nuremburg Codes:** the person involved...should have sufficient knowledge and comprehension of the elements of the subject matter involved, as to enable him to make an understanding and enlightened decision. This latter element requires that, before the acceptance of an affirmative decision by the experimental subject, there should be made known to him the nature, duration, and purpose of the experiment; the method and means by which it is to be conducted; all inconveniences and hazards reasonably to be expected; and the effects upon his health or person, which may possibly come from his participation in the experiment (1448).
>
> **The Declaration of Helsinki 2013:** In medical research involving human subjects capable of giving informed consent, each potential subject must be adequately informed of the aims, methods, sources of funding, any possible conflicts of interest, institutional affiliations of the researcher, the anticipated benefits and potential risks of the study and the discomfort it may entail, post-study provisions and any other relevant aspects of the study. The potential subject must be informed of the right to refuse to participate in the study or to withdraw consent to participate at any time without reprisal. Special attention should be given to the specific information needs of individual potential subjects as well as to the methods used to deliver the information.[19]

[18] Although proving a negative is difficult, in my career researching informed consent, I've never seen anyone invoke this distinction.

[19] Perhaps it's not fair to use standards of consent for research ethics to apply to clinical medicine. I disagree; the standard of consent should not change (or at least should not be more relaxed) in clinical medicine as opposed to research trials. In other words, the risks are potentially higher in research, given the novel nature of the intervention being tested. If we allow the Nuremberg Codes (1996) and the Declaration of Helsinki to be the guidelines for risky consent (research), then they surely must be sufficient (although likely not necessary) for less risky consent (clinical medicine).

In their extremely influential textbook, Beauchamp and Childress (2001, p. 81) say that:

> Professionals are generally obligated to disclose a core set of information, including (1) those facts or descriptions that patients or subjects usually consider material in deciding whether to refuse or consent to the proposed intervention or research, (2) information the professional believes to be material [they go on to list 3 more conditions not relevant for the question at hand].

These sources are extremely influential in the history of informed consent, and none of them distinguish between which level of information must be given to patients. Additionally, although it's too long to reprint here, the Council for International Organizations of Medical Sciences (2016) lists numerous items research participants must be informed about to give valid consent. Nowhere on their list is any distinction made about levels of information.[20] Additionally, the American legal system seems to focus on first-order information and to not require higher-order information; as Cohen (2020) says, "The substantial 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" (1435).

Despite the distinction between higher-order and first-order medical information being absent from various philosophical and policy statements concerning what information is required for consent, it seems clear to me that the implicit understanding of informed consent among bioethicists is that, when possible, patients must be informed of the first-order information. In other words, bioethicists would likely regard access to first-order medical information as a necessary condition for consent being sufficiently informed. Clearly, the model of the patient as the recipient of information which they use to make an autonomous and informed choice is in tension with a world in which patients are only being told higher-order information. The aspiration of the informed consent movement was to depart from a model of the doctor-patient relationship in which patients deferred to the doctor and to move instead to a model where the patient uses the clinical information given by the doctor to make an informed decision about their care. A medical system in which doctors only tell patients higher-order medical information (with no chance for patients to ask "why" questions) would much more resemble the kind of "trust me, I know what's best" view that the informed consent movement is a response to. Not to belabor the point, but it bears repeating. Although bioethicists have not made the distinction between higher-order and first-order medical information, it seems very clear to me that they regard first-order information as being required for informed consent. If the requirement of informed consent were to permit higher-order information to count as sufficient for informed consent, then patients could be sufficiently informed even if they were only told "I don't know the details of why the treatment works, but I know that it works in 99% of similar cases."[21] Proponents of informed consent as it

[20] See Council for International Organizations of Medical Sciences (2016) (103–106).

[21] A referee helpfully pointed out that SSRIs and paracetamol are examples of treatments that we have good reason to think are effective, but we do not know why.

is currently practiced would not regard this as sufficient information, indicating that they instead view first-order information as necessary for informed consent.

In our current understanding, it would not be enough if the patient knows that the doctor is very reliable but knew nothing about the reasoning in support of the treatment. Higher-order consent, as I've proposed it, violates this view by suggesting that patients don't need to be informed about the first-order reasoning of their doctors, so long as they are informed about the higher-order reliability of their doctors or of the treatment itself.

Against this backdrop, my proposal is that if we allow higher-order medical information to be sufficient for consent to be informed, then patients can still be sufficiently informed even if doctors use in-principle opaque medical AI. In a world where medical AI is widely used, we would have access to higher-order medical information about how accurate and reliable an algorithm is, and if this information is sufficient for consent, then patients can consent to being treated by opaque AI.[22] We can express this view as the following thesis:

> **Higher-Order Consent:** if patients are only informed of higher-order medical information about their treatment/diagnosis (and cannot access any first-order medical information about their treatment/diagnosis), they can still give valid, informed consent.

Compare this with the view that I take it that most bioethicists (implicitly) endorse:

> **First-Order Consent:** valid, informed consent requires that patients have the option to receive an explanation of their treatment/diagnosis worded in first-order medical information.[23]

I am suggesting that we reject the First-Order Consent thesis and instead endorse the Higher-Order Consent thesis, meaning that we can regard higher-order medical information as sufficient for consent being informed.[24] The view goes further than

[22] This would be in concert with Kiener's (2021) claim that patients should be informed "that even the best AI comes with the risks of cyberattacks, systematic bias, and a particular type of mismatch between AI's implicit assumptions and an individual patient's background situation" (705).

[23] This could of course be translated by the doctor into terms that the patient would understand.

[24] Steinberg (2024) defends the view that consent to AI is a form of uninformed consent but can still be valid consent under the right conditions. My view differs from this, in that I still view my proposal as a version of informed consent. I've just revised what kind of information is required for informed consent. However, in the world I'm envisioning, patients are not radically ignorant about the reliability of the AI that is recommending treatments to them. They are informed about its high degree of reliability, etc. On Steinberg's own analysis of informed consent, higher-order consent counts as informed consent. Steinberg's view is that "consent is informed when two conditions apply: 1. The agent has at least one of the following: (a) All the available information relevant to her decision (and she understands it); or. (b) The ability – in terms of epistemic access, sufficient time, and cognitive resources – to collect, understand, and consider available information relevant to her decision; or. (c) Available mediating resources or persons that can compensate for any insufficiency in (a) or (b) (e.g. guardian, fiduciary, professional advisor). 2. In deciding whether to consent, the agent can choose autonomously whether to use (1a), (1b), or (1c), or all, or none" (4). In cases of higher-order consent, the agent possesses (a) and (b). She possesses the relevant information (which is higher-order information in this case). If she thinks that first order information is relevant, then she fails to satisfy (a). But, my goal is just to argue that those patients for

defending the aforementioned right to not know as it is traditionally understood and claims that patients have the right to not know even in contexts where knowing is not possible.[25]

For the rest of this section, I will argue for the Higher-Order Consent thesis on three different grounds. First, I appeal to the purpose of consent. Second, I appeal to theories of disclosure in bioethics. Third, I appeal to the fact that many features of medicine seem to already operate as if the Higher-Order Consent thesis were true.[26]

Before continuing, I should add that the following discussion envisions a world in which patients have the option to be treated by the method of their choice, whether that be opaque AI, explainable AI, or humans. I am not suggesting that endorsing my view requires that we allow only opaque AI to treat patients; I am merely arguing that opaque AI is compatible with informed consent.

i) The Purpose of Consent

I'll start with my argument based on the purpose of consent. Broadly speaking, there are two views about what the purpose of consent is, and both support the Higher-Order Consent thesis. According to one view, the purpose of consent is to *protect autonomy*, and according to the other view, the purpose of consent is to

Footnote 24 (continued)
whom higher-order information is sufficient may consent, so this won't hurt my argument. She also, by stipulation, has the ability to access the information about the reliability of the AI in question. Overall, this means that my view is distinct from Steinberg's because my view is a version of informed consent, while Steinberg (2024) defends uninformed/imperfectly informed consent.

[25] I should be clear that opponents of the traditional right to not know will not be persuaded by my view. I'm fine with that implication, as I'm not swayed by their concerns. I think objections to the right to not know misunderstand the point of consent. Opponents of the right to not know argue that patients are obligated to know all of the information so that they can choose autonomously; on my view, the point of consent is to allow patients to make the choices they want to make. At the point where we are saying that consent only permits certain kinds of decisions by competent adults, we are missing the point of consent's autonomy protecting role. Arguments against the right to not know in contexts where using this right would harm third parties are compatible with my view. Ost claims that "the right to be informed… is a mandatory right-i.e., it is not an option right which one may or may not exercise; rather it is a right which we are obligated to exercise" (307). As Harris and Keywood (2001) argue, "where I give someone (against their will) reliable information about themselves or their condition which is relevant to decisions they must make I may violate a liberty they assert but I do not violate their autonomy, for the information I give them is necessary for their autonomous decision making" (419).

[26] In a related vein, Keren and Lev (2022) have recently defended the possibility of consent when a patient is in a state of "suspending ignorance," which occurs when the patient has suspended judgment about a belief relevant to their treatment. My proposal is different in that suspension of belief is not involved. In the cases I'm envisioning, we are stipulating that (if they are aware of all of the evidence) the patients believe that the treatment from the opaque AI is better for them but accept that they cannot access the AI's first-order rationale. In my proposal, there is no belief on which the patient is suspending judgment. Additionally, Steinberg (2024) does raise the distinction between higher and first order information and consent as it relates to AI. However, Steinberg (2024) applies this distinction differently than I do. As Steinberg (2024) discusses higher-order consent, he discusses an agent having higher-order knowledge that they are ignorant about the first-order information. As he says, "Having higher-order knowledge about first-order ignorance only means that the individual can know that she cannot know risk-relevant information about her decision" (14). I agree with this, however what I'm proposing is not about an agent's awareness of her own ignorance. I discuss higher-order consent in the context of an agent knowing certain relevant pieces of information about the treatment itself (e.g. its average success rate, etc.).

*promote wellbeing*.[27] Obviously, consent can do both things at once, but the proponents of these views intend to isolate what is intrinsically important about consent, as the goals of wellbeing and autonomy can easily come apart. I'll argue that both views of the purpose of consent support the view that patients may opt for treatment from an in-principle opaque AI.

Suppose that the purpose of consent is to protect our autonomy by allowing competent and free adults make the choices that they desire to make. As stipulated, we are assuming that treatment via in-principle opaque AI is in the medical best interests of the patient. If patients' goal is to pick the treatment with the highest possible chance for the best medical outcome (which is certainly not everyone's goal), then such patients will want to be treated by opaque AI and would be fine with receiving only higher-order medical information. If we were to deny such individuals this option, we frustrate the point of consent. In the name of informed consent, we should not tell patients that they lack the autonomy to waive certain rights when they deem it beneficial to do so. If the purpose of consent is to promote autonomy, it is disrespectful to autonomy to prevent a competent patient from pursuing an option they believe to be better for them. Again, I'm not envisioning a world in which anyone is being forced to be treated by an AI; human doctors are still available.[28]

If the purpose of consent is to promote wellbeing, and if it's correct that treatment via opaque AI produces the best medical outcome for patients, then the argument goes through quite easily. Allowing treatment via opaque AI would enable patients to achieve the function of consent by doing what is likely to best advance their wellbeing. Again, not allowing treatment via opaque AI would actively frustrate the goal of consent, as this would hinder an agent from maximizing their own wellbeing.

Thus, on both possible views of the purpose of consent, treatment via opaque medical AI can satisfy the function of consent. And, preventing such a treatment option would actually frustrate the purpose of consent!

ii) Theories of Disclosure

Now, I argue that according to the plausible theories of what disclosure for informed consent requires, treatment via in-principle opaque AI is compatible with informed consent.

Per Beauchamp and Childress (2001), the main views about disclosure are: (1) the professional practice standard, (2) the reasonable person standard, and (3) the subjective standard. The professional practice standard holds that the relevant information that doctors must disclose is determined by the "professional community's customary practices," which means that "professional custom establishes the amount and kinds of information

[27] As Beauchamp and Childress (2001) say, "since the mid-1970s, the primary justification advanced for requirements of informed consent has been to protect autonomous choice" (77). Elsewhere, Beauchamp (2010) puts it like this: "the purpose of consent provisions is not protection from risk, as many earlier federal policies seemed to imply, but the protection of autonomy and personal dignity" (9). Other philosophers have taken the view that the point of consent is to promote wellbeing. As Stacey Taylor (2004) argues, "the ethical foundation of informed consent is really concern for human well-being" (384).

[28] My view is in concert with Ploug and Holm's (2020) argument that patients have the right to not be treated or diagnosed by AI.

to be disclosed" (Beauchamp & Childress, 2001, p. 82). Thus, whatever it is customary to disclose is what doctors must to disclose. According to the reasonable person standard, "we must determine the information to be disclosed by reference to a hypothetical reasonable person" (Beauchamp & Childress, 2001, p. 82). If a reasonable person would judge a piece of information as relevant, then that information is relevant. And, if a reasonable person would judgea piece of information to not be relevant to their decision, then doctors need not disclose that information. Lastly, the subjective standard "judges adequacy of information by reference to the specific informational needs of the individual person, rather than the hypothetical 'reasonable person'" (Beauchamp & Childress, 2001, 83). In short, whether a piece of information is relevant for a patient is determined by the patient herself, but doctors are only obligated to disclose information that the patient asks for or which they could reasonably determine is relevant for the patient.

I take it that the professional practice standard is implausible, as it seems to commit a clear is-ought error of inferring from the way clinicians act to the way clinicians ought to act. The two plausible views are the reasonable person standard and the subjective standard. Both views have the implication that informed consent is compatible with being treated and diagnosed by an opaque AI.

A reasonable patient would want to receive the best care, regardless of how much first-order information they receive. If they are aware that AI care is most likely to achieve that, then a reasonable person would likely waive their right to an explanation so that they can receive the best medical care available. Suppose a patient has two options:

> **Treatment 1:** she can have a first-order explanation of the treatment and diagnosis, but she has compelling reason to think that there are better treatments available to her.
> **Treatment 2:** she cannot have a first-order explanation of the treatment and diagnosis, but there is compelling reason to think that this treatment will be more effective than treatment 1.

Treatment 2 is the more rational option to pick and thus would be selected by the reasonable patient (assuming that such a patient values health over access to information). Consider the previously mentioned use of AI to diagnose diabetic retinopathy. AI is substantially better than an ophthalmologist at identifying diabetic retinopathy. The AI (on 819 patients) was 87.2 sensitive (meaning able to correctly identify cases where the person does not have the condition) and 90.7% specific (meaning able to identify cases where the person does have the condition). This dramatically outperforms humans: "Previous studies have shown that board-certified ophthalmologists that perform indirect ophthalmoscopy achieve an average sensitivity of 33%, 34%, or 73%" (Abràmoff et al., 2018, 3). If the accuracy of the AI is truly that much better than that of human doctors, it seems plausible (all else equal) that a reasonable person would prefer the AI.

The previous paragraph needs a brief clarification. I do not endorse uniqueness about rationality, which is the view that from any body of evidence, there is only one rational conclusion to be inferred. As Matheson (2011) puts it, uniqueness holds that "For any body of evidence E and proposition P, E justifies at most one doxastic

attitude toward P" (360). I think that, depending on their priors, different agents (all of whom are reasonable persons) can rationally infer different conclusions. So, I am not claiming that choosing Treatment 2 is the only reasonable option, just that it is a reasonable option for someone who wants to receive the best medical outcome possible.

The fact that Treatment 2 seems rational leads me to speculate that most objections to AI in the clinical setting are really about a fear that the AI is not reliable, not about a concern that it jeopardizes informed consent. In other words, the consent-related concerns that many philosophers have about using AI in medicine are difficult to explain as a real concern for consent; instead, I'll tentatively suggest that the concerns many have about clinical AI are better explained by a background intuition that medical AI isn't actually better than doctors.

As a subjective matter, we'd have to ask individual patients if, given the facts about medical AI, they would prefer treatment via AI (without a first-order explanation) or from a human (with a first-order explanation). According to the subjective standard, what information a patient needs for consent is whatever information they want. If a patient wants only higher-order information, then their consent can be informed with only that information.

In short, both plausible views about what information ought to be disclosed to patients support the Higher-Order Consent thesis.

iii) The Status Quo Supports the Higher-Order Consent Thesis

Although my view conflicts with our rhetoric around informed consent, I'm not sure it's actually in conflict with already well-accepted medical practices. The mechanics of diagnostic tests are not usually explained to patients, and patients don't know how these tests reach a conclusion. They are told the outcome of the test. Those tests are not black boxes, but they may as well be to the patient. What's relevant to the patient is that they are accurate.[29] We don't usually have our doctor's success rate with us as we choose. So, with AI, we would have a piece of information we don't have with the doctor.[30] Patients often don't ask, nor would they always understand, the doctor's reasoning toward their diagnosis/treatment. But, so long as they are informed about the doctor's reliability, then it seems like they are sufficiently informed to give valid consent.

An objection by analogy enters here. One might object that using opaque AI in clinical medicine would justify absurd practices, such as a doctor using a Magic 8-Ball to make diagnosis and treatment decisions. Like the AI, the 8-Ball is entirely opaque to us. Clearly, current medical standards would not allow this. However, this analogy is easily set aside; as Cohen (2020) points out:

[29] I draw inspiration for this point from (Amman et al. 4–5).

[30] As Kawamleh (2023) says, "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," but "such information is in fact available for epistemically opaque AI systems through empirical validation" (910–911).

> With the Magic 8-Ball or Astrology...not only can the physician not explain why it works, but she also has no epistemic warrant that it works. Conversely, with medical AI/ ML (especially its more opaque forms), the argument is that it is more like Aspirin, in that the former may be true but the latter is not—assuming that physicians have good reason to believe the AI/ML is likely to lead to better decisions. The epistemic warrant for that proposition need not be firsthand knowledge—we might think of medical AI/ML as more like a credence good, where the epistemic warrant is trust in someone else...this is true of most FDA-approved pharmaceuticals. The physician is likely quite ignorant of the underlying trial design or results that led FDA to believe that the drug was safe and effective, but her knowledge that it has been FDA-approved supplies the necessary epistemic warrant (1444).

In short, the analogy fails, because we have no reason to believe that the recommendations of 8-Balls are likely to produce good medical outcomes. If, instead, it were an opaque 8-Ball that had shown itself to be highly reliable in this domain, it seems much more reasonable to use it.[31] Cohen (2020) suggests that this is more or less our situation with regard to many common pharmaceutical interventions; we know that they work but don't know why. London (2019) makes this point quite powerfully in his defense of opaque AI in medicine. In brief, London (2019) argues that just as AI can be opaque but effective, so too are many medical interventions:

> Although medicine is one of the oldest productive sciences, its knowledge of underlying causal systems is in its infancy; the pathophysiology of disease is often uncertain, and the mechanisms through which interventions work is either not known or not well understood. As a result, decisions that are atheoretic, associationist, and opaque are commonplace in medicine...As counterintuitive and unappealing as it may be, the opacity, independence from an explicit domain model, and lack of causal insight associated with some of the most powerful machine learning approaches are not radically different from routine aspects of medical decision-making. Our causal knowledge is often fragmentary, and uncertainty is the rule rather than the exception. In such cases, careful empirical validation of an intervention's practical merits is the most important task (17–18).

Patients already give valid, informed consent to doctors based on high-order information alone. As long as the patient knows that the doctor is credentialed, reliable in the past, etc., this seems perfectly fine. Allowing the use of opaque medical AI would merely be an extension of this existing practice. Patients can know the reliability of the AI, how it works, etc.[32]

[31] In a recent article in this journal, Schmidt, Martin Putora, and Fijten (2025) disagree with my claim that reliability is sufficient in these cases.

[32] Pruski (2024) has made a similar argument: "This point about reliability being more important than explainability is also highlighted by the fact that even in recent history we were not able to reliably explain how commonly used medications worked despite having good evidence that they did" (482).

One might worry that appealing to the status quo is not sufficient to justify a practice. After all, many unjust practices have been the status quo at certain points. I completely agree and don't think that this argument on its own is sufficient to justify the use of AI in healthcare. I only mean this argument as a response to those who seem to treat AI in medicine as if (from the standpoint of consent) it would be a massive departure from current practice and thus requires special justification. My goal is to undermine this claim by pointing to how similar the status quo is to AI treatment.

To sum up my argument, the purpose of consent, theories of disclosure, and the status quo all support the Higher-Order Consent thesis. If this thesis is true, then treatment and diagnosis from in-principle opaque medical AI is compatible with informed consent. My suggestion is that willing patients who are aware of the reliability of medical AI are making consensual decisions when they agree to waive their right to a first-order explanation of their diagnosis and treatment. To suggest that consent requires explainability of first-order information about diagnosis and treatment is to insist that patients must settle for worse medical outcomes.[33]

## 5 Objections

i) Modal Explanations?

One might object that an important difference between the status quo and the potential AI future is that, currently, patients can ask for an explanation. Whereas, in the AI world, this is not even an option. In a sense, the status quo allows for modal explanations, which are explanations a patient could ask for even if they don't do so. Modal explanations are not possible under the opaque AI system. As we discussed earlier in the paper, informed consent does not require that a patient understands the diagnosis, but it may require that the patient *can* have an explanation if they want one. In other words, the patient must have the ability to receive an explanation, even if they ultimately waive this ability. On this standard, opaque AI does not pass, while human doctors do.

I have two responses. First, it's far from true that doctors are able to give a fully explainable line of reasoning for how they reach their decisions, even if patients ask. There's a sense in which the demand for explainability in medicine is not motivated for doctors or AI. We don't typically require that doctors be fully explainable to patients, nor do we expect patients to be able to even understand the information if it were explained to them. Additionally, all human minds are opaque to a certain degree.[34] Second, if the choice is between better outcomes and a patient's right to a

[33] London (2019) echoes this point: "Any preference for less accurate models…carries risks to patient health and welfare. Without concrete assurance that these risks are offset by the expectation of additional benefits to patients, a blanket preference for simpler models is simply a lethal prejudice" (18).

[34] Some (e.g. Kawamleh, 2023, 903) have gone as far as to say that human minds are black boxes. That seems too strong. Even if humans lack full transparency, we can give a rough explanation of our thought process in a way that is not possible with opaque AI. For a detailed discussion of post-hoc rationalizations in humans (which bears on whether we are opaque to ourselves), see Doris (2015).

modal explanation, it seems clear that patients will want better outcomes. Respect for informed consent requires that we respect their decision. If patients choose to receive treatment that makes modal explanations not possible, this is their autonomous choice. At the risk of repetition, we should not allow our commitment to informed consent to prevent competent adults from seeking high quality care that they want.

ii) Pigeons and Trust?

I'll close this section with a discussion of a fascinating objection proposed by Alvarado (2022). Alvarado (2022) cites evidence which suggests that pigeons can be trained to be highly reliable at radiology; he then uses this as an analogy for why we ought to prefer explainable AI over opaque AI, even if the latter is more accurate:

> Few would feel as confident calling for a significant role for pigeons in assisting in diagnosis, let alone calling for the replacement of radiologists with pigeons. This is not just because of some irrational prejudice towards pigeons, but rather because there is something highly suspect, I argue, about replacing human expertise with an opaque, albeit accurate, methodology simply on the basis of predictive success (123–124).

Alvarado (2022) argues that this is the case because "in the context of epistemically demanding inquiry…non-epistemic reasons [practical reasons] are simply not good reasons for the adoption of novel artifacts" (132). Instead, Alvarado (2022) suggests that "when it comes to the introduction of novel technologies and the trust one can allocate to them within the context of formal inquiry, such as scientific endeavors (in which one seeks to understand and not merely to observe)" (132), we should appeal to epistemic reasons (reasons to believe a proposition is true). As he says, "citing the many lives that could be saved by accepting a proposition, or the efficiency of doing so, does not address the proposition's relation to truth and therefore does not constitute an epistemic reason to adopt it" (132).

I confess that I find this puzzling. I agree that, ideally, we should be able to know why a new medical device works. However, if the tradeoff is between having this knowledge or saving lives, I would choose the latter. I'm willing to concede that this means patients might not have strong reasons to trust opaque AI, but this perhaps shows that patients have the autonomy to receive treatment from sources that they, strictly speaking, don't trust.[35] Perhaps the catchier way to put it is that if pigeon radiologists reliably produced more accurate results than human radiologists, I say bring on the pigeon radiologists! Additionally, I'll note that while the goal of medical research is to find the truth, the goal of many patients is purely pragmatic: to achieve the best medical outcome (unless doing so compromises some more

[35] For those who find this too concessive, Durán and Jongsma (2021) have argued that "the reliability of algorithms provides reasons for trusting the outcomes of medical artificial intelligence" (329). Baron (2025) reaches a similar conclusion, that "for notions of trust that are appropriate for AI, explainability is not a necessary condition" (1). For a reply to Baron (2025), see Fan (2025). Budnik (2025) has also recently argued that "instead of trusting AI systems, we should strive to make them reliable" (1). As a referee helpfully pointed out to me, it may still be possible for the patient to have trust in the humans who created the AI.

important goal, e.g. religious integrity). Thus, if we agree that medicine is not only a truth-seeking enterprise, then I'm not sure it's a problem for pragmatic reasons to justify the adoption of a novel medical technology.

## 6 Conclusion

I've argued that medical AI (even if it's in-principle opaque) does not threaten informed consent in clinical medicine. Importantly, the view I defend in this paper is compatible with continuing to prioritize research into explainable AI. All I mean to defend is that informed consent and opaque AI are compatible, but this is consistent with informed consent being better achieved by explainable AI. It would clearly be the best option if we could have highly accurate and explainable AI in medicine.

We might reasonably wonder what medicine would look like were this view to be adopted. The goal of this paper is not to work out the logistical details of how this might all play out in a clinical setting. However, I would be remiss if I didn't address this issue. Some may worry that what I'm suggesting is a dangerous return to the old paternalism that characterized medicine prior to the informed consent movement. Fortunately, this is not the case.[36] There's nothing paternalistic about my proposal, as patient autonomy is the deciding factor all the way down. It is patients who are waiving a right to a first-order explanation, not doctors mandating this or withholding it from them when they would otherwise want to receive it. Several outcomes are possible, from completely replacing doctors with AI to using AI merely as a diagnostic aid. My own view is that the outcome with no human doctors remaining in medicine is not a good one. Although, such an outcome is compatible with respect for informed consent, so long as everyone genuinely prefers this outcome. But, consent is not the only value in medicine. We want patients to be comfortable, which is likely better achieved with humans still involved in the process. Perhaps my view would open the door to moving from, "patients may pursue AI care consistent with informed consent," to "patients can give informed consent to an AI, so let's only have AI doctors from here on." I do not think this is likely to happen. There will certainly still be demand for human doctors. Many people will not trust AI, and many will think that it's not really more reliable than human doctors.[37] In a market system, that demand will furnish a supplier. However, if the demand for human doctors decreases sufficiently, we may reach a point where a market no longer furnishes them. On my view, such an outcome is compatible with informed consent, but individuals who still insist on exercising their right to first-order explanation

[36] Additionally, Steinberg (2024) argues that the informal norms of medicine as a profession are set up so as to protect the expression of autonomy, even if traditional informed consent is not being followed. Steinberg (2024) makes this point in the context of a proposal about consent under radical ignorance, but a similar point could apply to the use of higher-order consent. Barocas and Nissenbaum (2014) defend a similar but importantly different point (2014a, 2014b).

[37] Evidence suggests that patients will be hesitant to use medical AI: "large-scale adoption of AI hinges not only on adoption by healthcare systems and providers but also on patient utilization, and patients are reluctant to utilize medical AI" (Cadario, Longoni, and Morewedge 2021, 1636).

of their treatment would be entitled to treatment via human doctors or explainable AI.[38] As I've gone to great pains to emphasize, what patients receive in terms of information should be respectful of what they want to receive. Just as insisting that they must receive first-order information does not respect their consent, so too does insisting that they can only receive higher-order information. In that regard, I differ from Cohen (2020), who claims that "if using AI/ML really produced better patient outcomes across the board, then it seems desirable for it to become the standard of care. As a matter of informed consent, ethically or legally, it is not clear why we should shed tears if in such a world, patients do not have access to the non-AI/ML approach" (1463).

**Acknowledgements** Thanks are due to David Boonin, James Stacey Taylor, Andrew J Cohen, Lauren Hall, Connor Kianpour, Jess Flanigan, Joe Millum, Blake Harris, Garrett Mindt, Steven Gubka, Stephen Hoover, and audience members at the 2023 PPE Society Meeting. Thanks are also due to Emily Director for her continued support, especially for times when I am being opaque.

**Author contributions** I am the sole author of this paper.

**Funding** I received a $100 honorarium for presenting an earlier version of this paper at an online workshop hosted by the Institute for Humane Studies. I also received a $300 honorarium for presenting this paper at a lecture series hosted by the Rochester Institute of Technology.

**Data Availability** N/A.

## Declarations

**Ethics Approval** No research on human subjects was conducted for this paper.

**Competing Interests** I declare no competing interests.



## References

Abràmoff, M. D., Lavin, P. T., Birch, M., et al. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine, 1*, 39. https://doi.org/10.1038/s41746-018-0040-6

Adorno, R. (2004). The right to not know: an autonomy based approach. *Journal of Medical Ethics, 30*(5), 435–439.

[38] For a contrary view on this point, see Pruski (2024, 484).

Alvarado, R. (2022). Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI. *Bioethics, 36*, 121–133.
Babic, B., et al. (2021). Beware explanations from AI in health care. *Science, 373*(6552), 284–286.
Babic, B., & Zoë Johnson, K. (2025). Algorithmic fairness and resentment. *Philosophical Studies 182*, 87–119. https://doi.org/10.1007/s11098-023-02006-5
Barocas, S., & Nissenbaum, H. (2014). Computing ethics big data's end run around procedural privacy protections. *Communications of the ACM, 57*(11), 31–33.
Baron, S. (2025). Trust explainability and AI. *Philosophy & Technology, 38*, 4. https://doi.org/10.1007/s13347-024-00837-6
Beauchamp T & J. Childress. (2001). *Principles of Biomedical Ethics*. 5th Edition. Oxford, UK: Oxford University Press.
Beauchamp, T. (2010). *Standing on Principles: Collected Essays*. Oxford University Press.
Bernat, J. L., & Peterson, L. M. (2006). Patient-centered informed consent in surgical practice. *Archives of Surgery, 141*(1), 86–92.
Earp, B. D., Sebastian Porsdam Mann, Allen, J., Salloch, S., Suren, V., Jongsma, K., & Matthias Braun, et al. (2024). A personalized patient preference predictor for substituted judgments in healthcare: Technically feasible and ethically desirable. *The American Journal of Bioethics 24*(7), 13–26. https://doi.org/10.1080/15265161.2023.2296402
Budnik, C. (2025). Can we trust artificial intelligence? *Philosophy & Technology, 38*, 10. https://doi.org/10.1007/s13347-024-00820-1
Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour, 5*, 1636–1642. https://doi.org/10.1038/s41562-021-01146-0
Chockley, K., & Emanuel, E. (2016). The end of radiology? three threats to the future practice of radiology. *Journal of the American College of Radiology, 13*(12), 1415–1420.
Christian Bjerring, J., & Busch, J. (2021). Artificial intelligence and patient-centered decision making. *Philosophy & Technology, 34*, 349–371.
Cohen, I. G. (2020). Informed consent and medical artificial intelligence: What to tell the patient? *Georgetown Law Journal, 108*, 1425–1469.
Council for International Organizations of Medical Sciences. (2016). *International Ethical Guidelines for Health Related Research Involving Humans.* Geneva, Switzerland: Council for International Organizations of Medical Sciences. http://www.cioms.ch/ethicalguidelines-2016/
Di Bello, M., Gong, R., (2023). Informational richness and its impact on algorithmic fairness. *Philosophical Studies* 1–29.
Doris, J. M., (2015) *Talking to our selves: Reflection, ignorance, and agency*. Oxford University Press.
Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*, 329–335.
Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*, 115–118.
Eyal, N. (2019). Informed Consent. *Stanford Encyclopedia of Philosophy* . https://plato.stanford.edu/entries/informed-consent/#Con
Faden, R., & Beauchamp, T. (1986). *A History and theory of informed consent*. Oxford University Press.
Fan, N. (2025). Explainability is necessary for AI's trustworthiness. *Philosophy & Technology, 38*, 17. https://doi.org/10.1007/s13347-025-00847-y
Grant, D. G. Behrends, J, & Basl, J. (2025). What we owe to decision-subjects: Beyond transparency and explanation in automated decision-making. *Philosophical Studies 182*, 55–85.
Harris, J., & Keywood, K. (2001). Ignorance, information, and autonomy. *Theoretical Medicine, 22*, 415–436.
He, J., Baxter, S. L., Xu, J., et al. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine, 25*, 30–36. https://doi.org/10.1038/s41591-018-0307-0
Holm, S. (2022). Handle with care: Assessing performance measures of medical AI for shared clinical decision-making. *Bioethics, 36*, 178–186. https://doi.org/10.1111/bioe.12930
Hull, R. (1985). Informed consent: patient's right or patient's duty? *The Journal of Medicine and Philosophy, 10*, 183–197.
Kawamleh, S. (2023). Against explainability requirements for ethical artificial intelligence in health care. *AI Ethics, 3*, 901–916. https://doi.org/10.1007/s43681-022-00212-1

Keren, Arnon and Ori Lev. (2022). Informed consent, error and suspending ignorance: providing knowledge or preventing error? *Ethical Theory and Moral Practice* 351–368.

Kernbach, J.M., Hakvoort, K., Ort, J., Clusmann, H., Neuloh, G., Delev, D. (2022). The Artificial Intelligence Doctor: Considerations for the Clinical Implementation of Ethical AI." In: Staartjes, V.E., Regli, L., Serra, C. (eds) *Machine Learning in Clinical Neuroscience. Acta Neurochirurgica Supplement*, vol 134. Springer, Cham. https://doi.org/10.1007/978-3-030-85292-4_29

Kiener, M. (2021). Artificial intelligence in medicine and the disclosure of risks. *AI & Society, 36*, 705–713.

Kundu, S. (2021). AI in medicine must be explainable. *Nature Medicine, 27*, 1328.

Liu, X., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet, 1*(6), E271–E297.

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

Lundberg, S. M., Nair, B., Vavilala, M. S., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng, 2*, 749–760. https://doi.org/10.1038/s41551-018-0304-0

Maclure, J. (2021). AI, explainability and public reason: The argument form the limitations of the human mind. *Minds and Machines, 31*, 421–438. https://doi.org/10.1007/s11023-021-09570-x

Matheson, J. (2011). The case for rational uniqueness. *Logos & Episteme,* 2(3), 359–373.

Milano, S., & Prunkl, P. (2025). Algorithmic profiling as a source of hermeneutical injustice. *Philosophical Studies, 182*, 185–203. https://doi.org/10.1007/s11098-023-02095-2

The nuremberg code 1947. (1996). BMJ 313 :1448 https://doi.org/10.1136/bmj.313.7070.1448

Ost, D. (1984). The 'right' not to know. *The Journal of Medicine and Philosophy, 9*(3), 301–312.

Ploug, T., & Holm, S. (2020). The right to refuse diagnostics and treatment planning by artificial intelligence. *Medicine, Health Care and Philosophy, 23*, 107–114.

Price, W. (2017). Nicholson II, regulating black-box medicine. *Mich L Rev, 116*, 421–474.

Pruski, M. (2024). AI-enhanced healthcare: not a new paradigm for informed consent. *J Bioeth Inq, 21*(3), 475–489. https://doi.org/10.1007/s11673-023-10320-0

Schiff, D., & Borenstein, J. (2019). How should clinicians communicate with patients about the roles of artificially intelligent team members? *AMA Journal of Ethics, 21*(2), E138-145. https://doi.org/10.1001/amajethics.2019.138

Schmidt, E., Putora, P. M., & Fijten, R. (2025). The epistemic cost of opacity: how the use of artificial intelligence undermines the knowledge of medical doctors in high-stakes contexts. *Philosophy & Technology, 38*, 5. https://doi.org/10.1007/s13347-024-00834-9

Da Silva, M. (2023). Explainability, public reason, and medical artificial intelligence. *Ethical Theory and Moral Practice, 26*, 743–762.

Steinberg, E. (2024). AI, radical ignorance, and the institutional approach to consent. *Philosophy & Technology, 37*(101), 1–26.

Taylor, J. S. (2004). Autonomy and informed consent: a much misunderstood relationship. *Journal of Value Inquiry, 38*(3), 383–392.

Ursin, F., Timmermann, C., & Steger, F. (2022). Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? *Bioethics, 36*, 143–153. https://doi.org/10.1111/bioe.12918

Wadden, J. J. (2021). What kind of artificial intelligence should we want for use in healthcare decision-making applications? *Canadian Journal of Bioethics, 4*(1), 94–100.

Watson, D., et al. (2019). Clinical applications of machine learning algorithms: Beyond the black box. *The British Medical Journal, 364*, 1–4.

World Medical Association (2024) "Wma declaration of helsinki – ethical principles for medical research involving human subjects." https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/#:~:text=In%20medical%20research%20involving%20human%20subjects%20capable%20of%20giving%20informed,risks%20of%20the%20study%20and

Yu, K.-H., Beam, A., & Kohane, I. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering, 2*, 719–731.