Philosophy & Technology (2025) 38:5
https://doi.org/10.1007/s13347-024-00834-9

RESEARCH ARTICLE

The Epistemic Cost of Opacity: How the Use of Artificial Intelligence Undermines the Knowledge of Medical Doctors in High-Stakes Contexts

Eva Schmidt[1,2] · **Paul Martin Putora**[3,4] · **Rianne Fijten**[5]

Received: 19 August 2024 / Accepted: 13 December 2024

Abstract
Artificial intelligent (AI) systems used in medicine are often very reliable and accurate, but at the price of their being increasingly opaque. This raises the question whether a system's opacity undermines the ability of medical doctors to acquire knowledge on the basis of its outputs. We investigate this question by focusing on a case in which a patient's risk of recurring breast cancer is predicted by an opaque AI system. We argue that, given the system's opacity, as well as the possibility of malfunctioning AI systems, practitioners' inability to check the correctness of their outputs, and the high stakes of such cases, the knowledge of medical practitioners is indeed undermined. They are lucky to form true beliefs based on the AI systems' outputs, and knowledge is incompatible with luck. We supplement this claim with a specific version of the safety condition on knowledge, Safety*. We argue that, relative to the perspective of the medical doctor in our example case, his relevant beliefs could easily be false, and this despite his evidence that the AI system functions reliably. Assuming that Safety* is necessary for knowledge, the practitioner therefore doesn't know. We address three objections to our proposal before turning to practical suggestions for improving the epistemic situation of medical doctors.

Keywords Medical AI · Safety Condition · Explainable AI · Artificial Intelligence · Black-box AI · Healthcare

✉ Eva Schmidt
eva.schmidt@tu-dortmund.de

[1] Department of Philosophy and Political Science, TU Dortmund, Dortmund, Germany

[2] Lamarr Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany

[3] Department of Radiation Oncology, Kantonsspital St. Gallen, St. Gallen, Switzerland

[4] Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

[5] Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Reproduction, Maastricht University Medical Center, Maastricht, Netherlands

Published online: 13 January 2025

1 Introduction

In the field of medicine, in recent decades, a fundamental shift towards evidence-based medicine has occurred. This shift has been accompanied by an objective approach, often relying on statistical methods. These statistical tools have become gradually more complex, which has recently been enabled by an increase in computing power, with the rising prominence of artificial intelligence (AI) in medicine. The use of AI has led to impressive results in various fields in medicine. Some of the new AI methods are used to detect or classify what is there (e.g., pulmonary nodules) (Baldwin et al., 2020), others to predict how things will progress (e.g., a disease) (Dembrower et al., 2020), and still others to recommend certain behaviors (e.g., a treatment) (Liu et al., 2018). In some instances, the combination of human and machine approaches has yielded the best results (Liu et al., 2018). Many of these applications are image- based, yet their applications and impressive results are not limited to this domain. For instance, AI-based applications are becoming more proficient in predicting health-related outcomes such as diagnosing diseases like lung cancer (Patel et al., 2019) or predicting cancer recurrence for a specific patient (Van Booven et al., 2021).[1]

In contrast to traditional statistical methods, it is often difficult to understand how highly complex and non-linear AI systems work, not only for a practitioner wanting to apply the model and the patient to whom it applies, but sometimes even for the engineers developing the system.[2] We can say that some AI systems are, to a certain extent, opaque. Sometimes the opacity of a system is due merely to the fact that the system is proprietary, but other times the reason is that its model involves sub-symbolic representations incomprehensible even to experts, or that it is too large or too complex to grasp (Bathaee, 2018; Ghassemi et al., 2021; Mann et al., 2023). Importantly, it is often not only the (global) inner workings of such systems that are hard to understand, but also the (local) causes of their individual outputs, for instance, why a certain CT scan was classified as depicting a tumor, or why a certain medical treatment was recommended for a patient (Reyes et al., 2020).

We think that the often superior detection and prediction capabilities of AI systems together with their opacity raise important philosophical, specifically epistemological questions in the medical context. Our concern is that, even if they work well, the opacity of AI systems may undermine users' – in particular, medical doctors' – knowledge acquired on the basis of these systems' outputs. Worries along these lines have been discussed by several philosophers of science, who predominantly, but not exclusively, ask how the opacity of AI systems in scientific inquiry affects the ability of scientists to explain and understand natural phenomena (e.g. Sullivan, 2022, Duede, 2023, Boge, 2022, Creel, 2020z & Beisbart, 2024). As our focus on the epistemic consequences of AI opacity indicates, we share the broad concerns of these authors. However, as will become clear in the following, our paper has a different

[1] See Topol (2019) for a detailed overview of applications.

[2] A quick note on terminology: We speak here of 'AI systems' to pick out AI-driven applications, typically programmed with the help of recent methods such as machine learning (ML), which are especially powerful and flexible problem-solvers.

focus: We are interested in scenarios in which medical practitioners, not scientists, aim for knowledge of a particular item of information, where their ability to integrate this information with further knowledge is substantially limited.

We will present a case that clarifies our concern (Sect. 2). We will next examine philosophical approaches to knowledge which might be used to substantiate the intuition that knowledge is undermined in our case, and argue that a version of the safety account of knowledge – Safety* – is the best fit (Sect. 3). Then, we will address three objections to our proposal (Sect. 4). In the conclusion, we will suggest practical ways of improving the epistemic situation of medical doctors. Our overall goal, then, is to flesh out an epistemological worry about the use of opaque AI systems by appeal to philosophical theorizing; and, in a more pragmatic vein, ways to alleviate the worry.

2 The Intuitive Case for Undermined Knowledge

Cancer Risk

> Imagine the scenario: An opaque AI system, which has learned to predict the risk of breast cancer recurring within two years after surgery on the basis of next generation sequencing data, is used to identify rare, at-risk patients, who qualify for additional (adjuvant) treatment. This treatment takes the form of endocrine therapy, chemotherapy, or radiotherapy and reduces the risk of recurrence over a course of two decades.[3] A medical doctor, call him MD, receives the system's prognosis and decides on its basis whether to prescribe the treatment to the relevant patient. The system's previous performance in trials was satisfactory, and when MD used it on earlier occasions, he couldn't find anything amiss. This is evidence that MD has, which makes it very likely that this is a well-working AI system. On a particular occasion, the system correctly predicts that a certain patient P is not at risk. On the basis of this, MD comes to believe that P is not at risk, and his belief is true: P has a very low risk of her breast cancer recurring.

In *Cancer Risk*, should we say that MD *knows* that P is not at risk? This question has two aspects. The first aspect is, how should the content 'P is not at risk' be construed? Is it the kind of thing that can be known, in principle? The second aspect is whether in the specific situation described, MD is in a position to know. Turning to the *first* aspect of the question, let's start with the content of the AI system's output. We take it that, in providing its prognosis, it relies on objective correlations between the DNA structure of patients and their developing breast cancer, on how frequently patients with such-and-such a DNA have had recurring breast cancer after treatment. The AI system's output then asserts an objective probability for a certain patient with such-and-such a DNA to have breast cancer again within two years. In our scenario, the specific output is that patient P is not at risk. Think of the system determining that

[3] These are common forms of therapy for adjuvant treatment of breast cancer, among others. See e.g. Dong and Gewirtz (2022) or Gradishar et al. (2021).

P's risk is below a certain threshold, say 0.05, and on that basis providing the output that she is not at risk.[4] We propose that MD's belief that P is not at risk takes up the content outputted by the system. We wish to remain neutral on whether the content of MD's resulting belief then involves an objective or a subjective probability.

But can it truly be said that someone knows a probabilistic content? Following Moss (2013, 2016), we insist that it can. As Moss shows, we do indeed naturally say things like, 'she knows that her specimen most likely belongs to *G hackmani*' (cf. Moss, 2013, p. 7) or 'he knew that it was probably time to repaint the fence'. Moss spells out probabilistic contents in terms of sets of probability spaces, which involve probability functions defined over sets of possible worlds (Moss, 2016, Chap. 1 and Chap. 2). She argues that belief with such contents constitutes knowledge under the same conditions as belief with non-probabilistic contents, e.g. when the subject is not in a Gettier case or a fake barn case (Gettier, 1963; Goldman, 1976). Correspondingly, MD's risk assessment – his belief that P is not at risk, or has a very low risk of recurring breast cancer – can be taken as a belief with a probabilistic content, which ascribes a low probability to the possibility that P will have breast cancer.[5] And this is something that MD can know, in principle.

In light of this result, is MD intuitively in a position to know in *Cancer Risk*? (This is the *second* aspect of our question.) Note the following relevant facts of the scenario. *First*, since the AI system is opaque, MD does not know, and is not able to understand, how the system works or why exactly it provides its predictions. Even more discomforting, not even its developers fully understand the system's inner workings or would be able to tell why a certain output was provided (Bathaee, 2018).

Second, MD cannot immediately tell by the effects of his treatment on his patients (or lack thereof) whether the system's predictions are correct. If P were at risk, but did not receive any additional treatment due to a false prediction, it would take months or even years before she would exhibit any symptoms. Moreover, that P is at risk does not imply that she will definitely develop cancer. It would take even longer for MD to notice that the system *systematically* falsely classifies at-risk patients as not at-risk, the system was not working well in general, or for a specific sub-population. In this hypothetical situation, the malfunctioning of the system would become apparent only years later after a large number of patients who received no additional treatment will have fallen ill again sooner than expected. This contrasts with uses of other AI systems – say systems classifying cat images – where it is easy to tell by the results whether an output is correct.

Third, even well-tested AI systems have been known to have bugs, to malfunction, or to exhibit bias. Here are some possible causes for concern: Assume that the algorithm was trained in Asia and later applied in Europe without any additional adjustments for underlying factors such as genetic or environmental differences between these populations. It may be that these differences between populations influence the performance of the AI system. That is, because of a genetic difference, the algo-

[4] We thank Eric Raidl and Florian Boge for discussion.

[5] Note further that it is highly plausible that we can have (at least probabilistic) knowledge of the future, as I can know that there will be no snow today (July 17) or that my daughter will most likely graduate from high school next year.

rithm might misclassify a much larger percentage of at-risk patients as not at risk (or vice versa) when used in Europe than it did during its training phase on an Asian population. Transfer of AI tools across different populations is a real-world problem, as evidenced by the studies in Kaushal et al. (2020), by AI systems used to detect intracranial hemorrhages (O'Connor, 2021), or by Watson for oncology, which is used world-wide despite being trained by doctors at one US hospital (Ross & Swelitz 2017). Further, it could be that because of some bias in the training data, the AI system regularly classifies a subgroup of the population with certain features as not at risk even if they are at risk (Fazelpour & Danks, 2021; Kordzadeh & Ghasemaghaei, 2021; Garcia, 2016). For instance, patients from a poor socio-economic background may have been severely underrepresented in the training data, and as a consequence the AI system may systematically under- or overestimate their risk. (For an illustrative case, see Obermeyer et al., 2019.) Despite an AI system's validation, there may be some internal error or bug that renders the system unreliable. Examples of such flawed medical AI systems include NarxCare, which helps US medical institutions decide who to give pain medication by providing an opioid overdose risk score (Szalavitz, 2021), the Epic Early Detection of Sepsis model (Richardson, 2021), or US systems for allocating healthcare to chronically sick people (Lecher, 2018). Next, some AI systems continue learning based on new data while in use in their application environment. It is possible that at some point, the system starts performing worse, not better, at detecting at-risk patients. As our examples show, these are not outlandish science fiction philosophy stories, but very real possibilities.[6]

Fourth, a lot is at stake in the situation, both for P and for MD. If P were a high-risk patient, not receiving the additional treatment would negatively impact her health. And MD might face serious legal repercussions if he did not provide P with an appropriate treatment. On the other hand, it would also be highly problematic to prescribe an invasive cancer treatment like chemotherapy to a patient who has a low cancer risk.[7]

Taken together, these facts about *Cancer Risk* (opacity, correctness-blindness, potential flaws, high stakes) intuitively support the following conclusion: Sure, MD

[6] An anonymous reviewer points out that the problems for the use of AI tools in medicine might be even more fundamental: Are the techniques on which such tools rely even in principle able to model a complex biological reality (e.g. Guzman-Alvarez, 2023)? In response, note that studies show that many AI systems used in medicine are highly reliable, and sometimes have greater precision than human doctors. As examples, see the studies cited in Bjerring and Busch (2021, 350) as well as Tschandl et al. (2020) or Tu et al. (2024). Moreover, our argument here can be read as follows: Despite the examples of failures of AI systems that we just listed, given the empirical evidence, we grant that AI systems in medicine often work really well. We go on to argue that nonetheless the use of opaque AI systems may undermine the knowledge of medical doctors. Those who share the reviewer's worry can give our argument a conditional reading: *If* AI systems used in medicine are highly reliable, then there are nonetheless contexts where medical doctors cannot know on the basis of their outputs.

[7] Readers may at this point already have objections to our description of the case (as well as to the inferences we are about to draw from it). In Sect. 4, we address objections that rely on analogies with arguably opaque cases of own expertise, expert testimony, technical instruments, and randomized control trials. We also return to the issue of whether medical doctors typically have or need knowledge. In Sect. 5, we address the concern that medical doctors have additional information available concerning their patients when they make treatment decisions. We focus on the bare case as described to elicit clear intuitions on how the use of opaque AI systems might affect knowledge.

has good evidence that the AI system is working reliably, and he is right about this. Against this background, the fact that the system's output is that P is not at risk is good evidence that, indeed, P is not at risk. MD believes this on the basis of the evidence. However, *if* the opaque system weren't working well and its output were false – which is not a far-fetched possibility – *he would have no way of telling*. MD is utterly unable to distinguish his actual situation, in which the AI system functions well and correctly classifies P as not at risk, from a counterfactual situation in which it is malfunctioning and misclassifies P. MD is thus entirely blinded to whether the system works well or not. If the system were malfunctioning and if its output were false, MD would form the belief that P is not at risk anyway.[8]

In light of this, it seems intuitively that in *Cancer Risk*, despite MD's good evidence that P is not at risk, he is *lucky* that he ends up with a true belief based on the AI system's output. But as Pritchard (2005, 2016) and Zagzebski (1994) have forcefully argued, knowledge is not compatible with luck. So, given that it is a matter of luck that MD's belief is true, he doesn't know that P is not an at-risk patient. To illustrate the relevance of luck, consider the following standard case from the epistemological literature:

Broken Clock

> Smith walks down the stairs and looks at the clock in the hallway at 3:00. It indicates that it is 3:00. However, the clock stopped working exactly 24 h ago, so that it merely happens to indicate the correct time right now. Smith believes that it is 3:00, and justifiably so, thanks to his visual experience of the clock. Since it really is 3:00, his belief is true.[9]

In *Broken Clock*, it is pure luck that Smith ends up with a *true* belief – had he come down the stairs and looked at the clock a few minutes earlier, his belief would have been justified, yet false. He got lucky by looking at the clock at the one time that it gave the correct time. So, intuitively, he doesn't know that it is 3:00. Similarly, MD is lucky that his belief is true: He is completely blinded to whether the AI system is well-working or malfunctioning (a possibility that is not too far-fetched), and so would form the belief that P is not at risk even if it were malfunctioning and the belief were false. And so he doesn't know. This result, if correct, has potentially far-ranging implications for the use of opaque AI systems in medicine. We will turn to two of them in Sect. (4). For now, suffice it so say that it seems worrisome if medical doctors don't *know* the relevant facts when advising patients or making far-reaching decisions. On the extreme end, such a worry has been articulated by philosophers as the claim that it is irrational to reason or to act on the basis of a fact unless you know that fact (Hawthorne & Stanley, 2008; similarly, Fantl & McGrath, 2002). If this view is correct, medical doctors like MD – who cannot acquire knowledge from the outputs

[8] Note that the counterfactual scenario that is worrisome to our mind is *not* one in which a generally well-working AI system, as a rare exception, gives an incorrect output. Even a well-working system will make false predictions on rare occasions. But medical doctors are well aware of this and intuitively this is not threatening to their knowledge like the general malfunctioning of a system.

[9] Readers should beware that Broken Clock is not completely analogous with Cancer Risk. In particular, the condition Safety*, to be introduced below, is met in Broken Clock, but not in Cancer Risk.

of AI systems – cannot rightly use the output of opaque AI systems as a basis of their decisions and actions in many contexts.[10]

To sum up, the intuition we have invoked is that in a medical context in which (1) an opaque AI system is employed that (2) is not immune to errors and (3) whose outputs cannot be checked for correctness by its user, while (4) a lot hinges on them making the right decision, these factors undermine the user's knowledge. This is so despite the user's evidence that the system works well, and despite the fact that it indeed works well. The intuition arises because the user has no way of distinguishing their actual situation, in which all is well, from a counterfactual but not far-fetched situation in which the system malfunctions (all while they are supposed to be making a high-stakes decision). If the system were malfunctioning and the user's belief were false, they would have no way of telling. In *Cancer Risk*, as far as MD can tell, he would believe that the patient is not at risk on the basis of the system's output even if the system were malfunctioning and the belief were *false.* So his belief is true only by luck, and he does not have knowledge.[11]

We believe that the intuition raised by *Cancer Risk* indicates a challenge for at least some uses of opaque AI systems in medicine. Indeed, we think that it may be partially responsible for a certain skepticism and hesitancy towards AI in the healthcare domain (Hengstler et al., 2016; Nadarzynski et al., 2019; McCradden et al. 2020). However, intuitions only carry so much weight in philosophical reasoning. We need to supplement this intuition with a philosophical account of knowledge on which, in scenarios as the one we describe, subjects lack knowledge. We do so by presenting a modal condition on knowledge which is not met by MD in *Cancer Risk*. Specifically, we think that it is a version of the safety condition on knowledge – Safety* – that is plausibly not met by MD, and that this can explain the intuition that he doesn't know.

3 Safety*

We have emphasized above that MD's problem is that, *if* the AI system *were* to malfunction, MD *would* have no way of telling. Even *if* the system malfunction*ed*, MD *would* have the same evidence that it works well and no evidence available to the contrary, and so he *would* rely on it and form beliefs about whether his patients are at risk on the basis of its output. This is the ground of our worry that MD doesn't know.

Our use of subjunctive, or counterfactual, conditionals to formulate the problem suggests a modal approach to MD's epistemic situation – that what undergirds the

[10] We will not, however, rely on this controversial claim in the following.

[11] We note points of contact with Lackey's (2011) argument that in cases of *isolated second-hand knowledge* subjects do know, but cannot assert what they know, and also cannot reason or act from their knowledge. (One of her examples is that of an oncologist who learns about a patient's medical condition only very briefly from a reliable and competent medical student's testimony, and has no further background knowledge available about the patient's condition. Here, the oncologist cannot appropriately assert to the patient that he has pancreatic cancer, and neither can she appropriately start surgery on the patient (Lackey, 2011, p. 255, 266).) Cancer Risk is plausibly a case of isolated second-hand belief. However, MD's epistemic situation is worse than that of the subjects in Lackey's examples. Lackey's oncologist could in principle find out from the medical student, or by studying the patient's files, on what grounds the diagnosis was made. MD is barred from doing the same in our example.

problem that MD intuitively faces is that he fails to meet a modal condition on knowledge. Let us explain.

Broadly, modality is about what is necessary and what is possible, including counterfactual ways the world could have been, but isn't actually. Possible ways the world could be, or could have been, can be conceived as possible worlds (where these are thought of as *complete* ways the world might be). For instance, there is the possible world in which Queen Elizabeth is still alive in 2024, and the possible world in which humans never evolved. These examples of possible worlds show that they can be more or less similar – closer to or farther away from – the actual world, depending on how much needs to be changed about the actual situation to end up in these possible worlds. For instance, it should be intuitive that less has to be changed about the actual world to move to a possible world where the Queen is alive in 2024 than to move to one without humans.

Counterfactual conditionals like the ones stated above can be analyzed by way of possible worlds. (See Starr, 2022 for a critical overview.) To give another example, say we want to evaluate whether the following counterfactual conditional is true: If humans didn't emit CO2, there would be no climate crisis. We can understand this as making a claim about how the world would be if we were to change one specific aspect of it (viz. human CO2 emissions). Correspondingly, to evaluate whether the conditional is true, we focus on possible worlds in which we imaginatively change human CO2 emissions to zero (and nothing else!), and then consider whether in such possible scenarios, there is a climate crisis. That is to say, we look at the possible worlds most similar to the actual world in which humans don't emit CO2, and consider what is true of these worlds. If we subtract humans emitting CO2, the cause of the climate crisis disappears; there is no climate crisis in close-by possible worlds without human CO2 emission. So the counterfactual conditional is true. One might object that we can imagine that there is a climate crisis even if humans emit no CO2, for example because volcano eruptions emit the same amount of CO2. However, this more outlandish possible world is beside the point for evaluating the counterfactual, as it is not one of the most similar worlds to the actual world. We can imagine all kinds of crazy additional changes, but what we want to know is whether the climate crisis would be happening, all else being equal, if humans didn't emit any CO2.

Now modal conditions on *knowledge* use the toolbox of possible worlds to spell out how knowledge excludes luck. As stated above, in order to know, it cannot be a matter of luck that your belief is true. Take Smith from *Broken Clock*. That he ended up with a true belief by pure luck can be understood to say that he could easily have ended up with a false belief – that there are close-by possible worlds in which his belief is false. Vice versa, it is plausible that when a subject knows, her belief couldn't easily have been false, so that it is true in the possible situations that are most similar to her actual situation. As a matter of fact, this is the core idea of the *safety* condition on knowledge, which we here propose to use to substantiate the intuition that MD doesn't know.

In recent decades, Safety has been the focus of much discussion (Pritchard, 2007; Sosa, 1999). To say that a belief is safe is to say that the belief *couldn't easily have been false*.[12] Here is a standard way of phrasing the condition:

Safety If S were to believe that p, then it would not be false that p.

Proponents of Safety typically hold that a subject knows a proposition p *only if* her belief is Safe, i.e. it couldn't easily have been false. Safety is conceived as a necessary condition on knowledge (Williamson, 2000; Sosa, 1999) – a condition that a belief has to meet to count as knowledge.[13]

To illustrate, return to *Broken Clock*. Could Smith's belief that it is 3:00 easily have been false? Had Smith come down the stairs just a few minutes earlier or later – which could have easily happened – he would have ended up with a false belief. (For then he would have looked at the broken clock and believed that it was 3:00, though it wasn't.) There isn't much we would have to change about the scenario to get his belief to come out false. So, the belief could have easily been false – it is unsafe and does not amount to knowledge. Safety captures the idea that knowledge is incompatible with luck. It is lucky of Smith to have formed a true belief exactly in that his belief isn't Safe, in that it could easily have happened that it was false.

To evaluate the conditional Safety, we need to examine whether it is true that, if Smith were to believe that it is 3:00, then it would be 3:00 (it would not be false that it is 3:00). To do so, we must consider possible worlds that are very similar to Smith's actual situation and in which Smith also believes that it is 3:00. We need clarity on whether in all of these possible worlds, it is 3:00/his belief is true. In *Broken Clock*, this is not the case: There is a range of very similar possible worlds (those in which he comes down the stairs a few minutes earlier or later) in which he believes it is 3:00, but this is not true. In this sense, his belief could easily be false.[14]

There are different ways to develop safety conditions in detail. We will rely on a version of the principle that draws on Whiting (2020), which we call Safety*. This principle tells us that the subject knows that p only if, *relative to her partial perspective*, her belief that p could not easily be false. Here's the corresponding conditional:

*Safety** If S were to believe that p, then relative to her perspective, it could not easily be false that p.

This proposal makes a difference to what the close-by possible worlds are that determine whether a belief is safe. With Safety*, the relevant possible worlds are those that are similar to the subject's actual situation as she sees it from her partial perspective on the world. The subject's perspective on her situation is partial because she doesn't have full knowledge of the world, but only of a limited range of features

[12] For an overview, see Rabinowitz (2011).

[13] Keep in mind that there can be more than one necessary conditions that a belief has to meet to be knowledge; for instance, it might be that a belief has to be both Safe and Safe* to be knowledge.

[14] Since what is at issue is whether the belief could easily be false, not whether it would be false in far-fetched scenarios or with difficulty, what is relevant are (again) very similar or close-by possible worlds.

of her situation, relative to her evidence about the world.[15] This evidence provides the way the actual possibility is from her perspective. Possible worlds are more or less similar to the actual possibility according to her partial perspective, depending on how much we would have to change the actual situation in her perspective to reach these situations. So in a nutshell, on Safety*, a belief is safe, just in case relative to the believer's perspective, it could not easily be false.[16]

In *Broken Clock*, then, we get the result that Smith's belief is Safe*. Smith's evidence does not include the fact that the clock is broken (he is unaware of this), but it does include the fact that the clock is generally highly reliable – we can imagine that it has never had any problems before – and that it says that it is 3:00. So the actual situation in his limited perspective is one in which the clock is working well. Relative to this perspective, it could not easily happen that Smith falsely believes that it is 3:00 by looking at the clock. From his perspective, the most similar possible worlds where he has the belief are worlds in which it is true. These will include scenarios like the following: Smith walks down the stairs at 3:00, wearing a different shirt. Smith looks at the clock exactly 12 h earlier than in the actual situation. So, relative to his limited perspective, the belief is Safe*. Correspondingly, in Smith's perspective, relative to his evidence, it is not lucky that his belief is true. As far as he can tell, it is a safe bet that it is 3:00.[17]

Applying Safety* to *Cancer Risk* shows in which respect MD's belief, that P is not at risk, is lucky, or so we will argue now. Let us start with MD's evidence and perspective on the world. With that in hand, we can examine whether, relative to it, MD's belief could easily be false. To begin with, his total evidence includes facts about the AI system's satisfactory performance in trials and during previous times he used it. It includes no evidence that the system is malfunctioning. But further, MD's perspective includes the facts that the system is opaque, that he cannot immediately tell by the results whether its outputs are correct, that systems, including ones used in medicine, have malfunctioned for a variety of reasons, and that a lot is at stake for him and for P. Now at first sight, the fact that his perspective includes evidence of the well-functioning of the AI system suggests that the belief is Safe*. Possible worlds in which MD falsely believes that P is not at risk, due to the AI system's malfunctioning, seem to be quite *dis*similar from the actual possibility, as determined by MD's

[15] We follow Whiting in assuming that a subject's evidence is what she knows, so that the subject's perspective is constituted by what she knows.

[16] Safety is normally conceived of in terms of metaphysical possible worlds. Whiting (2020) departs from this by spelling the principle out via epistemically possible worlds. Epistemically possible worlds are complete possibilities that the subject cannot rule out a priori, which are a priori coherent. The nearby worlds are those that are very similar to the epistemic possibility determined by the subject's total evidence (i.e., knowledge); they are the epistemically possible worlds that are most similar to this epistemic possibility. This affects which possible worlds count as close-by. On the metaphysically possible worlds conception, Safety asks us to inspect possible worlds that are, objectively speaking, similar to how the actual world is. On the epistemically possible worlds conception, Safety* has us consider possible worlds that are similar to the actual possibility according to the subject's limited epistemic perspective, which is given by her overall evidence/knowledge.

[17] Note that it is compatible with this that the belief is not Safe (as elucidated earlier) and so that, since Safety is a necessary condition on knowledge, Smith doesn't know that it is 3:00 even though he meets Safety*.

evidence. We would have to change quite a lot about the actual possibility (according to MD's perspective) to move to such a possible world. So *prima face*, it appears that his belief could not easily be false given his perspective.

However, this first impression doesn't hold up once we place the proper weight on the fact that MD's perspective also includes the opacity of the system, the impossibility of checking the correctness of the output directly, and the quite realistic possibilities of its malfunctioning. Given these factors, from MD's perspective the belief that P is not at risk *could* easily be false. For they cast doubt on the power of MD's evidence to support the assumption that the AI system is well-functioning, and thus also on this assumption itself. This doubt is partly constitutive of MD's perspective. The fact that he is entirely blinded to whether the AI system is working well or not (which is what reasonably raises MD's doubt) acts as a defeater of his evidence for the claim that the system is well-working. And if his evidential support for this claim breaks away, then he doesn't know it. If his knowledge doesn't include the fact that the system is well-functioning, then given what he knows, the possibility that the system is *mal*functioning and its output is false is not far off. And so neither is the possibility that P is at risk despite the AI system's output. So overall, in light of MD's doubt, a possible world in which the system malfunctions and MD's belief, based on its output, is false, is not outlandish after all. That is to say, there is a very similar, close-by possible world in which MD's belief is false. As far as MD knows, the belief could easily be false – it is lucky, from MD's perspective, that it is true. The belief is not Safe* and thus is not knowledge.

One might object that even so, the possible world in which the AI system is malfunctioning is quite different from MD's actual possibility from his perspective, given his good evidence that the system works well. Thanks to this evidence, it is *not* lucky that MD's belief that is based on the system's output is true. Maybe the possible world isn't extremely outlandish, but it's still not something that could easily happen. To respond, we emphasize again that a lot is at stake in MD's actual situation: If MD falsely forms the belief that P is not at risk, and thus P doesn't get additional treatment, this may have severe consequences for P's health. Further, MD might get in trouble for medical malpractice. Plausibly, in situations in which it is vital that the subject's belief is true, we have higher standards of Safety*.[18] That is, even if a belief could not quite *so* easily have been false, we are still inclined to say that it isn't Safe*, simply because the stakes are high. So, it is still lucky from MD's perspective that his belief is true, and it is not Safe*.

[18] In other words, we find it plausible that whether a belief Safe*, and thus whether it is knowledge, can vary with how much is at stake in a situation. This is a claim discussed by epistemologists as 'pragmatic encroachment' (e.g. Fantl & McGrath, 2002).

Table 1 Medical contexts involving reliable but opaque belief-forming methods

Method	Elicited belief	Reliability of method	Opacity of method	Potential flaws
0. Opaque AI system	P is not at risk.	The AI system reliably assesses the cancer risk of patients. It met the standards during trials.	It's not accessible to which patterns the system responds, or why it categorizes P as being low-risk.	Well-known bugs, biases, malfunctions in AI systems generally.
1. Perceptual expertise	This is a benign melanocytic nevus.	The medical doctor is a skin cancer expert. She reliably distinguishes benign from malignant skin conditions by visual inspection.	It's not accessible to the medical doctor how she can visually tell that something is a benign skin condition – she just can.	Well-known biases and flaws in human perception and reasoning.
2. Expert testimony	P is not at risk.	The colleague on whom the general practitioner relies is a renowned and highly reliable breast cancer expert, as well as a trustworthy informant.	It's inaccessible to the general practitioner both what his colleague's reasons are for categorizing P as low-risk and how her reasoning process works.	Well-known biases and flaws in human reasoning.
3. CT scanner	The patient has no brain tumor.	The images created by CT scanners very reliably depict brain structures, including tumors.	The medical doctor doesn't understand the complex internal workings of CT scanners.	As any technical device, (the components of) CT scanners are liable to malfunction sometimes.
4. RCT	Medical treatment T is effective.	RCTs have been widely shown to establish the effects of medical interventions by ruling out irrelevant influences from other factors.	RCTs leave opaque the mechanism by with medical interventions work. So medical doctors cannot tell *why* the treatment is effective.	In recent decades, the replication crisis has called into question the robustness of the results of RCTs.

4 Objections and Replies

Next, we turn to three purported problems for our argument. The *first* problem starts from the worry that our argument overgeneralizes.[19] There are many non-AI cases that appear to involve the four features on which we rely in our analysis – opacity, correctness-blindness, potential flaws, and high stakes. For instance, take a medical doctor who visually inspects a patient's skin and reaches the diagnosis that his mole is a benign melanocytic nevus. In this case, it may be opaque to the doctor how she reached her diagnosis (the mole just *looks* benign); she may be unable to tell directly whether her diagnosis was correct; she may be aware of biases and reasoning errors in humans; and this is a high-stakes situation. And similarly for medical doctors relying on the expertise of their colleagues, on medical machinery such as CT scanners, or on the results of randomized controlled trials (RCTs) (London, 2019; Kempt et al., 2022) – see Table 1. The objection then takes the form of a *reductio ad absurdum*: Our argument implies an absurd widespread skepticism. Consequently, our analysis

[19] We thank Mona Simion for putting the worry to us in this way.

of *Cancer Risk* should be rejected. Medical doctors acquire knowledge in AI-based cases like ours, as they do in many other cases which share the four features.

We limit ourselves to medical contexts with high stakes in which it is not possible for the subject to tell immediately whether the belief in question is correct. In each case to which our argument supposedly overgeneralizes, there is a belief-forming method that a medical doctor can use to acquire relevant beliefs, e.g. that a patient has a certain medical condition, that a patient should receive a particular medical treatment, or that a certain medical treatment is effective. The respective method is highly reliable, but opaque, just like the AI system in *Cancer Risk*, and there are pertinent problems with malfunctions or flaws. We use Table 1 to bring out the structural similarities between *Cancer Risk* (0) and cases to which our argument is supposed to overgeneralize (1)–(4).

In response, we submit that standard cases in which subjects rely on their own perceptual expertise, on expert testimony, on technical devices, or on the results of RCTs differ in relevant ways from *Cancer Risk*. (1) In cases where medical doctors rely on their own perceptual expertise, for instance in determining whether a mole is benign or malignant, the opacity of their belief-forming method (expert perception) is less severe than in cases involving opaque AI systems. In such cases, medical doctors are typically able to make some of their reasoning explicit, for instance by visually double-checking whether the mole has the well-known typical characteristics of malignant melanoma or of a benign melanocytic nevus. Another difference is that, as some philosophers have argued, *self-trust* in one's own capacities, together with the conscientious employment of these capacities, is an essential precondition for having justified beliefs and knowledge in the first place (Zagzebski, 2012; Dormandy, 2020). By contrast, trust in AI systems is not essentially tied up with our human abilities to justifiably believe and know. In virtue of these disanalogies, we cannot generalize without further ado from undermined knowledge in *Cancer Risk* to undermined knowledge in cases of reliance on perceptual expertise. And correspondingly, the competent medical doctor's belief is Safe*: Given her entitlement to self-trust and her (limited) access to how she comes to her diagnosis, there is no justified doubt regarding her own competence from her perspective, and the doctor's belief is true in the most similar possible worlds.

(2) In cases where medical doctors rely on the expertise of others, e.g. when a general practitioner comes to believe that her patient is not at risk for recurring breast cancer because this is what an expert oncologist tells her, the opacity is less severe than in cases of opaque AI systems. Medical experts are typically able to (and are expected to be able to) explain the reasons for their diagnosis or risk assessments to others, including medical doctors (Lackey, 2011): Medical doctors often jointly discuss their reasons for their judgments, and it is not uncommon to ask medical specialists for further explanation of their diagnoses or risk assessments. From an epistemological angle, according to *non-reductionism* about testimony, hearers have a defeasible right to believe what speakers in their community tell them by default, without positive reasons to trust the speakers (Lackey, 2006). This is in disanalogy to AI-supported cases – there is no such default right with respect to technical devices, which are not part of a speaker community and for which we do need evidence of their reliability before we may believe their outputs. These differences between the

cases block any automatic generalization. Again, a medical doctor's belief that P is not at risk formed in this way is Safe*. Because of her access to her colleagues' reasons for their assessment, she is not completely blinded to whether their judgment is well-grounded. This together with non-reductionism about testimony means that there is no good reason for the doctor to doubt her method of belief formation, and so the close-by possible worlds, given her perspective, are ones in which her belief that P is not at risk is true.

(3) Cases in which medical doctors acquire beliefs on the basis of technical devices like CT scanners are not opaque in a threatening sense. At hospitals, there is personnel with technical expertise whose job it is to ensure that all technical devices work flawlessly, and who have substantial understanding of how they work. And typically it is noticeable if a CT scanner, for instance, is not well-calibrated. Moreover, even if such devices sometimes malfunction, this is less worrisome than in the case of AI systems, since hospitals have well-established methods of ensuring that their devices work well, such as regular quality checks and maintenance. (Recall the real-world cases of flawed AI systems – in many of them, one problem is that there is no well-established quality-check system. Also see Topol (2019).) So in the most similar worlds to the actual possibility determined by the medical doctor's evidence, her belief that the patient has no brain tumor is true, and her belief is Safe*.

(4) London (2019) points out, RCTs tell us whether a medical intervention works, but not *how* it works; they do not provide information about the causal mechanisms that make interventions effective. However, we *do* know how RCTs are set up, and under which conditions they provide a good basis for believing in the effectiveness of an intervention. For specific medical treatments, the studies establishing their effectiveness are available in journals, or online. So, as Bjerring and Busch (2021, 364) point out, there is indeed some information made available that can be used by medical doctors to justify their belief that a certain treatment is effective. This is less so in case of opaque AI systems. Their effectiveness is also available in journals or online, and the data required to run the models are known. However, with these opaque systems, medical doctors are entirely blinded to the patterns used to generate the system's output (Bjerring & Busch, 2021, p. 365). So, our argument does not generalize to cases involving RCTs. By the same token, the medical doctor's belief that treatment T is effective is Safe*, since the closest possible worlds from her perspective are ones in which the trials are also set up correctly, and her belief justified by the reliability of the results of the trials is true. Overall, the overgeneralization worry can be blocked for all the examined cases.

The second problem has it that, quite the opposite, our argument *under*generalizes. Isn't *Cancer Risk* a rather exceptional scenario, from which we cannot draw any far-reaching conclusions for the use of AI in medicine? As we see it, the question is whether there are other medical contexts that share the same structural features of (1) opacity, (2) correctness-blindness, (3) potential flaws, and (4) high stakes. If so, the worries about the Safety* of the medical doctors' beliefs and about their knowledge, generalize to these contexts. Concerning (1), modern AI systems are often opaque. Regarding (3), the worry about potential flaws of AI systems is not limited to an individual scenario, but arises for medical AI quite generally. As to (4), clearly many medical contexts involve high stakes.

In light of this, the most problematic structural feature for our argument is (2), correctness-blindness. It is not immediately clear that medical doctors are often unable to check by looking at the results, and in a reasonable time period, whether a certain AI system they use is well-working. However, we believe that by thinking through particular medical decision situations, it becomes plausible that this feature, correctness-blindness, indeed generalizes to a broad range of situations. Examples include hepatitis C testing, high-stakes triage scenarios, patient-facing triage, embryo selection for in vitro fertilization, the detection of intracranial hemorrhages, or the distinction of recurrent brain tumors from radiation necrosis.[20] For illustration, consider a triage scenario involving acute shortage of treatment capacities, in which the deciding medical doctor has to take care of an overwhelming number of patients. A reliable, but opaque AI system is used by her to differentiate between patients who should receive curative care and those who should receive only palliative care. The system relies on a large body of data about a patient's medical history, current medical condition, age, gender, etc. Because of the great number of patients, the doctor lacks the time to check up on the status of individual patients. In a particular case, the AI system gives the (true) output that the patient doesn't need curative care. This is a high-stakes scenario, the system is opaque, and there are the well-known flaws of AI systems used in medicine. But what about correctness-blindness? The medical doctor here is unable to check whether the system's classification of the patient is correct, for in this overwhelming medical emergency, she has no time to check the patient's status herself. Further, we can assume that even a patient who would do fine with curative care will just die if given merely palliative care, so that the medical doctor won't be able to check back after the fact whether a patient really would have needed curative care. In light of this, she has no way of distinguishing her actual situation, in which the system works well, from a counterfactual but not far-fetched situation in which the system malfunctions. Intuitively, then, the medical doctor's belief is not Safe*, and she doesn't know that the patient needs no curative care. We lack the space to develop the same line of reasoning for other cases here. However, due to the structural features plausibly shared by many cases such as this, we believe that the epistemic problem we pinpoint generalizes to a significant number of medical contexts.

To press the *third* problem, our opponents will grant that knowledge is undermined in cases like ours, but insist that this is uninteresting. For knowledge is not needed in medical decision situations. Instead, medical doctors quite often correctly rely on (merely) rational belief or degrees of belief in accordance with their evidence. Medical doctors, together with their patients, often have to make decisions under uncertainty. Their evidence is not strong enough to ensure knowledge, but they have to come to reasonable decisions anyway. For this, they should adopt the belief (or degree of belief) supported by their overall evidence, and then act on that. In *Cancer Risk*, the overall evidence supports the hypothesis that P is not at risk, which MD should believe, at least to a certain degree; and the rational thing to do on this basis is not to prescribe adjuvant treatment. So our argument is successful, but our results

[20] For discussions of the use of AI in the mentioned contexts, see e.g., Liu et al. (2021), Laxar et al. (2023), Ilicki (2022), Afnan et al. (2021), Savage et al. (2024), and Park et al. (2021).

are beside the point. MD doesn't know that P is not at risk, but this is of no practical relevance.

Our response is that, in medical contexts, knowledge is important in at least two practical respects. Most philosophers agree that moral responsibility for an action includes an epistemic condition (Rudy-Hiller, 2022; Duff, 2009).[21] On the standard phrasing of this condition, an agent is responsible for an action only if she knows what she is doing; and she is excused and thus not responsible or blameworthy for an action if she is non-culpably ignorant of the negative consequences of her action (Duff, 2009, p. 979; Wieland, 2017). If medical doctors relying on AI-generated diagnoses or recommendations are prevented from knowing these (e.g.: P is not at risk), they will, for instance, not know whether they are ending medical treatment for a completely recovered patient, or whether they are withholding life-saving treatment from a high-risk patient. Either way, they will be faultlessly ignorant. If patients don't receive the right treatment in such cases, we will thus not be able to hold medical doctors responsible – there is a responsibility gap (Baum et al., 2022). To the extent that it is important that we can apportion moral or legal responsibility for medical decisions, then, it is important that medical doctors have knowledge.

In the same vein, informed consent to medical treatments involves an epistemic condition. According to the standard view, a patient's consent to a medical treatment is valid only if the patient knows – and even understands– relevant facts about her medical condition, the treatment options, possible side-effects of the available treatments, and the like (Eyal, 2019, § 4.2, Keren & Lev, 2022).[22] Now the medical doctor is the person whose task it is to disclose the relevant information to the patient and so to enable her to know the relevant fact and to give informed consent. If the medical doctor's knowledge of the relevant facts is undermined by the use of opaque AI systems, this threatens to undermine the patient's knowledge as well, and so to undermine informed consent. Given that informed consent is a valuable good, it is important that medical doctors know.[23]

Maybe this strong stance doesn't convince our opponents. Both in the debate over moral responsibility and that over informed consent, there is room for discussion how extensive the required knowledge has to be (Keren & Lev, 2022) or whether knowledge is needed to ensure these goods (Rosen, 2008).[24] But even if knowledge is not necessary for moral responsibility or informed consent, and if it can be argued that some weaker epistemic standing, such as justified belief, will suffice, we would like to insist that improving medical doctors' epistemic situations is a good idea. Independently of the issue of knowledge, it is desirable that their beliefs are supported by the best evidence they can get. For believing on better grounds makes for better medical decisions in high-stakes situations. So even those opponents who are

[21] The notion of *mens rea* indicates that the same may hold for legal responsibility (Duff, 2009). We have to leave discussion of this to one side here.

[22] We thank Arnon Keren for helpful discussion of this point.

[23] See also Steinberg (2024) for the worry that radical ignorance in AI contexts undermines informed consent.

[24] Relatedly, consider the nuanced discussion of the responsibility of medical doctors and other players in AI-supported decision-making in Verdicchio and Perin (2022).

not now convinced of the importance of knowledge in medical contexts should agree with the practical measures we suggest in the conclusion.

5 Summing up

We have argued that Safety* can be used to substantiate the intuition that MD's belief that P is not at risk is lucky, and thus not something that MD knows. Relative to MD's perspective, the belief could easily be false, even though MD has evidence that the AI system functions reliably. The system's opacity and the impossibility to check immediately by the results whether its outputs are correct together with the well-known flaws of AI systems, against the backdrop of the high stakes of the situation, reasonably leads to doubt whether the system is well-functioning and whether its output is correct. In light of this doubt, the possibility that P is at risk despite the system's output is not far-fetched. So, from MD's perspective, there is a close-by possible world in which MD's belief is false, and his belief is not Safe* and not knowledge.

Instead of concluding that opaque AI systems shouldn't be used in medical contexts, we want to end this paper on a more positive note, by focusing on ways in which the problem might be overcome. Epistemic problems similar to ours have been discussed with respect to beliefs based on internet searches, social media, or other everyday software applications, where there also are opaque AI systems that provide users with outputs (Dahl, 2018; Grindrod, 2019; Miller & Record, 2013). Despite their different approaches, the authors emphasize two ways to overcome problems for knowledge from the opacity of AI systems: first, *other*, *independent sources* may help the subject to acquire knowledge where an AI system's output alone cannot; second, *making AI systems more transparent* – overcoming some of their opacity – can alleviate the problem. A third way of overcoming epistemic problems discussed in the literature on medical AI involves rigorous validation of AI systems (Durán & Formanek, 2018; Durán & Jongsma, 2021; Topol, 2019; Ghassemi et al., 2021).

As to the first way, we can transfer Miller and Record's (2013) and Dahl's (2018) suggestion that subjects should use other sources, outside those employing personalized AI systems on the internet, to the medical context. As a matter of fact, medical decision-making is a multifaceted process with several factors relevant, such as the availability of treatment options (Panje et al. 2018), a wide range of decision-making criteria (Glatzer et al., 2020), and even emotions that may significantly impact the decision- making process (Treffers & Putora, 2020). On the one hand, medical doctors may have other information available about whether a patient is at risk aside from the outputs of AI systems. In identifiable "broken-leg" scenarios (Meehl, 1957), the physician needs to be able to override the AI system. The "broken-leg rule" proposes judgmental approaches can be more effective than AI systems for rare circumstances, due to their capacity to take into consideration additional information that may be relevant (Neirotti et al., 2021). The Safety* of the beliefs of medical doctors can thus be improved on other grounds.

As to the second way, the problem that we have discussed is partly due to the AI system's opacity (the second worrisome feature of MD's situation). If, in *Cancer Risk*, MD is able to understand the inner workings of the AI system, this puts him

in a position to remove his doubt that maybe the system doesn't work well. If MD has such an understanding of the system, he can check whether it produced a correct output in a particular case. While a complete understanding of the inner workings remains improbable, approaching this with explainable AI may be valuable (Holzinger, 2018; Holzinger et al., 2017). Similarly, if there is an expert who understands the system's inner workings, MD can rely on her expertise to ensure that a particular output is correct. This changes MD's perspective – as far as he can tell, then, his belief that P is not at risk could not easily be false. And so MD can make sure that his belief is Safe*.

Concerning the third way, a medical doctor's epistemic situation would be improved if reliable societal structures for validation and certification were in place. As Grindrod (2019) suggests, subjects *can* properly form beliefs and gain knowledge by relying on instruments whose workings they don't understand. For this, they need to rely on other members of their epistemic community who have a better understanding of these instruments or who have checked by external testing that they work well. That is to say, a medical doctor could rely on experts who thoroughly check and validate the system, and who are thus in a position to certify that the system works reliably, and even that it is periodically checked for sustained reliability across time (London, 2019; Topol, 2019; Ghassemi et al., 2021). This solution calls for a reliable and exceptionless system for validating and certifying AI systems used in high-stakes medical contexts, as required e.g. by the future EU AI Act. It would help to alleviate worries concerning the well-known flaws of AI systems and so contribute to the Safety* of medical doctors' beliefs.

In this way, close analysis of what is epistemically problematic in a hypothetical case like *Cancer Risk* leads to practical suggestions for improving the epistemic situation of actual medical doctors relying on support from AI systems.

Acknowledgements For extremelyhelpful discussion, we would like to thank audiences at the following events: ThomasSchmidt's colloquium, Berlin 2024; the Ethics & Epistemology Research GroupTwente, 2022; the graduate workshop *Philosophy Meets Machine Learning*, Tübingen2022; GAP11, Berlin 2022; the European Epistemology Network Conference, Glasgow2022; the Rhine Ruhr Epistemology Network Meeting, Dortmund 2022; the workshop *Issuesin XAI 4*, Delft 2022; the Theoretical Philosophy Colloquium Dortmund, 2022;the Luxembourg workshop 2022. We are also grateful to our anonymous reviewersfor very useful comments.

Author Contribution All three authors contributed to the article. Eva Schmidt is the first author.

Funding Open Access funding enabled and organized by Projekt DEAL.
This paper was supported by the Volkswagen Foundation, as part of the project Explainable Intelligent Systems (EIS), project number AZ 98510, and by the Federal Ministry of Education and Research of Germany and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany.

Data Availability/Code Availability There are no data, materials, or code to be made available by the authors.

Declarations

Ethics Approval No ethics approval was needed or obtained for this article.

Consent for Publication The authors consent to submit this article.

Financial Interests None.

Competing Interests None.

References

Afnan, M., Liu, Y., Conitzer, V., Rudin, C., Mishra, A., Savulescu, J., & Afnan, M. (2021). Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Human Reproduction Open*, *2021*(4), hoab040. https://doi.org/10.1093/hropen/hoab040

Baldwin, D. R., Gustafson, J., Pickup, L., Arteta, C., Novotny, P., Declerck, J., & Gleeson, F. V. (2020). External validation of a convolutional neural network artificial intelligence tool predict malignancy in pulmonary nodules. *Thorax*, *75*(4), 306–312. https://doi.org/10.1136/thoraxjnl-2019-214104

Bathaee, Y. (2018). The Artificial Intelligence Black Box and the failure of intent and causation. *Harvard Journal of Law & Technology*, *31*(2), 889–938. https://jolt.law.harvard.edu/assets/articlePDFs/v31/The-Artificial-Intelligence-Black-Box-and-the-Failure-of-Intent-and-Causation-Yavar-Bathaee.pdf

Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From responsibility to reason-giving explainable Artificial Intelligence. *Philos Technol*, *35*(12). https://doi.org/10.1007/s13347-022-00510-w

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

Boge, F. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines* 32(1), 43–75. https://doi.org/10.1007/s11023-021-09569-4 (2022).

Creel, K. (2020). Transparency in Complex Computational systems. *Philosophy of Science*, *87*(4), 568–589. https://doi.org/10.1086/709729

Dahl, E. S. (2018). Appraising Black-Boxed Technology: The positive prospects. *Philosophy and Technology*, *31*(4), 571–591.

Dembrower, K., Liu, Y., Azizpour, H., Eklund, M., Smith, K., Lindholm, P., & Strand, F. (2020). Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology*, *294*(2), 265–272.

Dong, P., & Gewirtz, D. A. (2022). Editorial: Risks and benefits of adjuvants to Cancer therapies. *Frontiers in Oncology*, *12*, 913626. https://doi.org/10.3389/fonc.2022.913626

Dormandy, K. (2020). Epistemic Self-Trust: It's personal. *Episteme*, 1–16. https://doi.org/10.1017/epi.2020.49

Duede, E. (2023). Deep learning opacity in scientific discovery. *Philosophy of Science*, *90*(5), 1089–1099.

Duff, A. (2009). Legal and Moral responsibility. *Philosophy Compass*, *4*, 978–986. https://doi.org/10.1111/j.1747-9991.2009.00257.x

Durán, J., & Formanek, N. (2018). Grounds for Trust: Essential epistemic opacity and computational reliabilism. *Minds & Machines*, *28*, 645–666.

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*(5).

Eyal, N. (2019). Informed Consent. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Spring 2017 Edition)*. URL: https://plato.stanford.edu/archives/spr2019/entries/informed-consent

Fantl, J., & McGrath, M. (2002). Evidence, Pragmatics and Justification. *The Philosophical Review*, *111*, 67–94.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*(8), e12760. https://doi.org/10.1111/phc3.12760

Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, *33*(4), 111–117.

Gettier, E. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), E745–E750.

Glatzer, M., et al. (2020). Decision making Criteria in Oncology. *Oncology (Williston Park, N.Y.)*, *98*(6), 370–378.

Goldman, A. (1976). Discrimination and perceptual knowledge. *Journal of Philosophy*, *73*, 771–791.

Gradishar, W. J., Moran, M. S., Abraham, J., Aft, R., Agnese, D., Allison, K. H., & Kumar, R. (2021). NCCN Guidelines® Insights: Breast Cancer, Version 4.2021: Featured Updates to the NCCN Guidelines. Journal of the National Comprehensive Cancer Network, 19(5), 484–493. Retrieved Jul 19, 2024, from https://doi.org/10.6004/jnccn.2021.0023

Grindrod, J. (2019). Computational beliefs. *Inquiry*, 1–22.

Guzman-Alvarez, A. (2023). Deep Learning for Investigating Causal Effects with High-Dimensional Data: Analytic Tools and Applications to Educational Interventions. Doctoral Dissertation, University of Pittsburgh. http://d-scholarship.pitt.edu/44666/ (Accessed Nov. 15, 2024).

Hawthorne, J., & Stanley, J. (2008). Knowledge and action. *Journal of Philosophy*, *105*(10), 571–590.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105–120.

Holzinger, A. (2018). *From machine learning to explainable AI.* Paper presented at the 2018 world symposium on digital intelligence for systems and machines (DISA).

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). *What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923*.

Ilicki, J. (2022). Challenges in evaluating the accuracy of AI-containing digital triage systems: A systematic review. *Plos One*, *17*(12), e0279636. https://doi.org/10.1371/journal.pone.0279636

Kaushal, A., Altman, R., & Langlotz, C. (2020). Geographic distribution of US cohorts used to Train Deep Learning algorithms. *Journal of the American Medical Association*, *324*(12), 1212–1213.

Kempt, H., Heilinger, J. C., & Nagel, S. K. (2022). Relative explainability and double standards in medical decision-making. *Ethics and Information Technology*, *24*, 20. https://doi.org/10.1007/s10676-022-09646-x

Keren, A., & Lev, O. (2022). Informed consent, error and suspending ignorance: Providiong Knowledge or preventing error? *Ethical Theory and Moral Practice*, *25*(2), 351–368.

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, 1–22.

Lackey, J. (2006). Knowing from Testimony. *Philosophy Compass*, *1*, 432–448. https://doi.org/10.1111/j.1747-9991.2006.00035.x

Lackey, J. (2011). Assertion and isolated second-hand knowledge. In J. Brown, & H. Cappelen (Eds.), *Assertion: New Philosophical essays* (pp. 251–276). Oxford University Press.

Laxar, D., Eitenberger, M., Maleczek, M. (2023). The influence of explainable vs non-explainable clinical decision support systems on rapid triage decisions: a mixed methods study. *BMC Medicine* 21, 359 (2023). https://doi.org/10.1186/s12916-023-03068-2

Lecher, C. (2018). *What happens when an algorithm cuts your health care*. URL: https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy

Liu, C., Liu, X., Wu, F., Xie, M., Feng, Y., & Hu, C. (2018). Using artificial intelligence (Watson for Oncology) for treatment recommendations amongst Chinese patients with lung cancer: Feasibility study. *Journal of Medical Internet Research*, 20(9), e11087.

Liu, W., Liu, X., Peng, M., Chen, G. Q., Liu, P. H., Cui, X. W., Jiang, F., & Dietrich, C. F. (2021). Artificial intelligence for hepatitis evaluation. *World Journal of Gastroenterology*, *27*(34), 5715–5726. https://doi.org/10.3748/wjg.v27.i34.5715

London, A. J. (2019). Artificial Intelligence and Black-Box Medical decisions: Accuracy versus Explainability. *Hastings Center Report*, *49*(1), 15–21.

Mann, S., Crook, B., Kästner, L., Schomäcker, A., & Speith, T. (2023). *Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy*. *arXiv:2307.14232*.

McCradden, M. D. (2020). Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: A qualitative study. *CMAJ open*, 8(1),E90–E95.
Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counsel- ing Psychology*, *4*(4), 268.
Miller, B., & Record, I. (2013). Justified belief in a digital age: On the epistemic implications of secret internet technologies. *Episteme*, *10*(2), 117–134.
Moss, S. (2013). Epistemology Formalized. *Philosophical Review*, *122*(1), 1–43.
Moss, S. (2016). *Probabilistic knowledge*. Oxford University Press.
Nadarzynski, T., et al. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital Health*, *5*, 2055207619871808.
Neirotti, P., Pesce, D., & Battaglia, D. (2021). Algorithms for operational decision-making: An absorptive capacity perspective on the process of converting data into relevant knowledge. *Technological Forecasting and Social Change*, *173*, 121088.
O'Connor, M. (2021). Algorithm's 'unexpected' weakness raises larger concerns about AI's potential in broader populations. URL: https://healthimaging.com/topics/artificial-intelligence/weakness-ai-broader-patient-populations
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissectingracial bias in an algorithm used to manage the health of populations. *Science*, *336*, 447–453.
Panje, C. M. (2018). Treatment options in Oncology. *JCO Clin Cancer Inform*, 2(2), 1–10.
Park, Y. W., Choi, D., Park, J. E., et al. (2021). Differentiation of recurrent glioblastoma from radiation necrosis using diffusion radiomics with machine learning model development and external validation. *Scientific Reports*, *11*, 2913. https://doi.org/10.1038/s41598-021-82467-y
Patel, D., et al. (2019). Implementation of Artificial Intelligence techniques for Cancer Detection. *Augmented Human Research*, *5*(1), 6.
Pritchard, D. (2005). *Epistemic luck*. Clarendon.
Pritchard, D. (2007). Anti-luck epistemology. *Synthese*, *158*(3), 277–297.
Pritchard, D. (2016). *Epistemology*. Springer.
Rabinowitz, D. (2011). The Safety Condition for Knowledge. *The Internet Enyclopedia of Philosophy*. URL: iep.utm.edu/safety-c/
Räz, T., & Beisbart, C. (2024). The Importance of Understanding Deep Learning. *Erkenntnis* 89 (5).
Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F. M., von Tengg-Kobligk, H., & Wiest, R. (2020). On the Interpretability of Artificial Intelligence in Radiology: Challenges and opportunities. *Radiol Artif Intell*, *2*(3). https://doi.org/10.1148/ryai.2020190043. e190043.
Richardson, L. (2021). *Artificial Intelligence Can Improve Health Care but Not Without Human Oversight*. URL: https://www.pewtrusts.org/en/research-and-analysis/articles/2021/12/16/artificial-intelligence-can-improve-health-care-but-not-without-human-oversight
Rosen, G. (2008). Kleinbart the oblivious and other tales of ignorance and responsibility. *Journal of Philosophy*, *105*(10), 591–610.
Ross, C., & Swetlitz, I. (2017). IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. URL: https://www.statnews.com/2017/09/05/watson-ibm-cancer/
Rudy-Hiller, F. (2022). The Epistemic Condition for Moral Responsibility. In E. N. Zalta & U. Nodelman (Eds.). *The Stanford Encyclopedia of* Philosophy *(Winter 2022 Edition*). URL: https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic
Savage, C., Tanwar, M., Elkassem, A., Sturdivant, A., Hamki, O., & Smith, A. (2024). Prospective Evaluation of Artificial Intelligence Triage of Intracranial Hemorrhage on Noncontrast Head CT examinations. *American Journal of Roentgenology*. https://doi.org/10.2214/AJR.24.31639
Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical Perspectives*, *13*, 141–153.
Starr, W. (2022). Counterfactuals. In E. N. Zalta & U. Nodelman (Eds.). *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (Eds.), URL: https://plato.stanford.edu/archives/win2022/entries/counterfactuals
Steinberg, E. (2024). AI, radical ignorance, and the Institutional Approach to Consent. *Philos Technol*, *37*, 101. https://doi.org/10.1007/s13347-024-00787-z
Sullivan, E. (2022). Understanding from Machine Learning models. *British Journal for the Philosophy of Science*, *73*(1), 109–133.
Szalavitz, M. (2021). *The Pain Was Unbearable. So Why Did Doctors Turn Her Away*? URL: https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7

Treffers, T., & Putora, P. M. (2020). Emotions as Social Information in Shared decision-making in Oncology. *Oncology (Williston Park, N.Y.)*, *98*(6), 430–437.

Tschandl, P., Rinner, C., Apalla, Z., et al. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, *26*, 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

Tu, T., Palepu, A., Schaekermann, M., et al. (2024). Towards conversational diagnostic AI. *arXiv org*. https://doi.org/10.48550/arXiv.2401.05654

Van Booven, D. J., et al. (2021). A systematic review of Artificial intelligence in prostate Cancer. *Research and Reports in Urology*, *13*, 31–39.

Verdicchio, M., & Perin, A. (2022). When doctors and AI interact: On human responsibility for Artificial risks. *Philos Technol*, *35*, 11. https://doi.org/10.1007/s13347-022-00506-6

Whiting, D. (2020). Knowledge, justification, and (a sort of) safe belief. *Synthese*, *197*(8), 3593–3609.

Wieland, J. W. (2017). Introduction: The Epistemic Condition. In P. Robichaud, & J. W. Wieland (Eds.), *Responsibility: The Epistemic Condition*. Oxford Academic. https://doi.org/10.1093/oso/9780198779667.003.0017

Williamson, T. (2000). *Knowledge and its limits*. Oxford University Press.

Zagzebski, L. (1994). The inescapability of Gettier problems. *The Philosophical Quarterly*, *44*(174), 65–73.

Zagzebski, L. (2012). *Epistemic Authority: A theory of Trust, Authority, and autonomy in belief*. Oxford University Press.