Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI

Juan Manuel Durán ,[1] Karin Rolanda Jongsma [2]

[1]Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands
[2]Julius Center, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

Correspondence to
Dr Juan Manuel Durán, Delft University of Technology, Delft 2600, The Netherlands; j.m.duran@tudelft.nl

Received 19 August 2020
Revised 11 January 2021
Accepted 8 February 2021
Published Online First 18 March 2021

ABSTRACT

The use of black box algorithms in medicine has raised scholarly concerns due to their opaqueness and lack of trustworthiness. Concerns about potential bias, accountability and responsibility, patient autonomy and compromised trust transpire with black box algorithms. These worries connect epistemic concerns with normative issues. In this paper, we outline that black box algorithms are less problematic for epistemic reasons than many scholars seem to believe. By outlining that more transparency in algorithms is not always necessary, and by explaining that computational processes are indeed methodologically opaque to humans, we argue that the reliability of algorithms provides reasons for trusting the outcomes of medical artificial intelligence (AI). To this end, we explain how *computational reliabilism*, which does not require transparency and supports the reliability of algorithms, justifies the belief that results of medical AI are to be trusted. We also argue that several ethical concerns remain with black box algorithms, even when the results are trustworthy. Having justified knowledge from reliable indicators is, therefore, necessary but not sufficient for normatively justifying physicians to act. This means that deliberation about the results of reliable algorithms is required to find out what is a desirable action. Thus understood, we argue that such challenges should not dismiss the use of black box algorithms altogether but should inform the way in which these algorithms are designed and implemented. When physicians are trained to acquire the necessary skills and expertise, and collaborate with medical informatics and data scientists, black box algorithms can contribute to improving medical care.

Linked

► http://dx.doi.org/10.1136/medethics-2021-107352
► http://dx.doi.org/10.1136/medethics-2021-107353
► http://dx.doi.org/10.1136/medethics-2021-107462
► http://dx.doi.org/10.1136/medethics-2021-107463

Check for updates

To cite: Durán JM, Jongsma KR. *J Med Ethics* 2021;**47**:329–335.

BACKGROUND

The use of advanced data analytics, algorithms and artificial intelligence (AI) enables the analysis of complex and large data sets, which can be applied in many fields of society. In medicine, the development of AI has spawned optimism regarding the enablement of personalised care, better prevention, faster detection, more accurate diagnosis and treatment of disease.[1 2] Aside from the excitement about new possibilities, this emerging technology is also paired with serious ethical and epistemic challenges.

Algorithms are being developed in several forms, ranging from very simple and transparent structures to sophisticated self-learning forms that continuously test and adapt their own analysis procedures.[3 4] It is these later types of algorithms that are often referred to as *black box algorithms*.[2 5] At its core, black boxes are algorithms that humans cannot survey, that is, they are epistemically opaque systems that no human or group of humans can closely examine in order to determine its inner states.[6] Typically, black box algorithms do not follow well understood rules (as, for instance, a Boolean Decision Rules algorithm does), but can be 'trained' with labelled data to recognise patterns or correlations in data, and as such can classify new data. In medicine, such self-learning algorithms can fulfil several roles and purposes: they are used to detect illnesses in image materials such as X-rays,[7] they can prioritise information or patient files[8] and can provide recommendations for medical decision-making.[9 10] The training of such systems is typically done with thousands of data points. Their accuracy, in contrast, is tested against a different set of data points of which the labelling is known (ie, done by humans). Interestingly, even if we claim understanding of the underlying labelling and mathematical principles governing the algorithm, it is still complicated and often even impossible to claim insight into the internal working of such systems. Take for example an algorithm that can accurately detect skin cancer in medical images as well as support the diagnostic accuracy of physicians. Physicians may be able to interpret—and even verify in many cases—the outcome of such algorithms.[11–13] But unfortunately, a black box algorithm is opaque, meaning that the physician cannot offer an account of how the algorithm came to its recommendation or diagnosis. This is a challenge for medical practice as it raises thorny epistemological and ethical questions that this article intends to address: Do we have sufficient reasons to trust the diagnosis of opaque algorithms when we cannot entrench how it was obtained? Can physicians be deemed responsible for medical diagnosis based on AI systems that they cannot fathom? How should physicians act on inscrutable diagnoses?

The epistemological opacity that characterises black box algorithms seems to be in conflict with much of the discursive practice of giving and asking for reasons to believe in the results of an algorithm, which are at the basis of ascription of moral responsibility. Concerns relate to problems of accountability and transparency with the use of black box algorithms,[14–17] (hidden) discrimination and bias emerging from opaque algorithms,[18–21] and the raising of uncertain outcomes that potentially undermine the epistemic authority of experts using black box algorithms.[11 22 23] Especially in the field of medicine, scholars have lately shown a preference for arguing that black box algorithms should not be accepted nor trusted as standard practice, principally because they lack features that are essential to

good medical practice. Rudin has even gone further to suggest that black box algorithms must be excluded in high-sensitive practices, such as medicine.[24]

Whereas these moral concerns are genuine, they neglect the epistemological bases that are their conditions of possibility. We propose, instead, that the epistemology of algorithms is prior to, and at the basis of studies on the ethics of algorithms. In other words, we make visible the epistemic basis that—to a certain extent—governs normative claims. However, this epistemic trust does not come with normative justification, and therefore, justified actions cannot be based on this knowledge alone. Our strategy consists in showing first that *computational reliabilism* (CR) offers the right epistemic conditions for the reliability of black box algorithms and the trustworthiness of results in medical AI; second, we show that prominent ethical questions with regard to decision-making emerge in the context of using black box algorithms in medicine. These questions concern the role of responsibility, professional expertise, patient autonomy and trust.

This article is structured as follows: in section 2, we will clarify underlying notions in the debate about black box algorithms, including notions such as transparency, and methodological and epistemological opacity. Subsequently, we describe CR as a suitable framework for justifying the reliability of an algorithm as well as providing reasonable levels of confidence about the results of opaque algorithms. It is important to clarify that our claim is not that crediting reliability to an algorithm justifies its use in all cases and for all purposes. A reliable algorithm might still negatively influence in different ways expert's decisions[i], or forge a less resilient healthcare system by, for instance, outsourcing decision making to algorithms with the corresponding lack of proper training to healthcare personnel. Although these are important implications of reliable algorithms, they fall outside the scope of this paper. In section 3, we describe the ethical concerns that remain in the context of reliable black box algorithms. In the conclusion, we will show how our analysis contributes to a more nuanced and constructive understanding of limitations of opaque algorithms and its implications for clinical practice.

TRANSPARENCY AND OPACITY

If we are unable to entrench reliable knowledge from medical AI, what reasons do physicians have to follow their diagnosis and suggestions of treatment? This is a question about claims of epistemic trust over the AI system and its output.[25–27] Answers typically revolve around two core concepts, namely, *transparency* and *opacity*. The former refers to algorithmic procedures that make the inner workings of a black box algorithm interpretable to humans. To this end, an *interpretable predictor* is set out in the form of an exogenous algorithm capable of making visible the variables and relations acting within the black box algorithm and which are responsible for its outcome.[28] Opacity, on the other hand, focuses on the inherent impossibility of humans to survey an algorithm, both understood as a script as well as a computer process.[6 29]

The fundamental difference between transparency and opacity lies in that opacity is about claims on the non-surveyability of algorithms, whereas transparency contends that some relevant degree of surveyability is, indeed, possible.[ii] This contrast can be illustrated with a simple example. Consider any given algorithm A. To make A transparent is to have an interpretable predictor with procedures $P = \{p_1, p_2 \ldots p_n\}$, where any given p_i $(1 < i < n)$ describes a sequence of specific relations among variables and functions in A, and where p_i entails the results of A. Thus understood, if A is an algorithm for classifying different types of skin cancer, P would realistically include procedures that relate the size, the shape, and the colour of the mole with outputs such as 'melanomas; squamous cell carcinomas; basal cell carcinomas; nevi; seborrheic keratoses'. Thus understood, transparency is an epistemic manoeuvre intended to offer reasons to believe that certain algorithmic procedures render a reliable output. Furthermore, according to the partisan of transparency, such a belief also entails that the output of the algorithm is interpretable by humans.

Opacity is a different animal altogether. At its core, it is the claim that no human agent (or group of agents) is able to follow the computational procedure that enables the claim that P entails A.[6] To see this, consider halting a running algorithm at any given point. According to epistemic opacity, humans are neither able to account for the state of the algorithm (ie, its variables, relations, system status, etc) previous to the halt, nor to predict any of the future state of the algorithm after the halt. Furthermore, humans would not be able to account for the state of the algorithm and its variables at the time of the halt either. The reasons are rather simple: we are limited cognitive human agents, we can store up to a certain amount of information in our brains and we can reliably handle even less, our computations are slow and too prone to errors, and algorithms are extremely complex entities to be surveyed. Epistemic opacity, then, abandons the goal of transparency as a means to foster trust in algorithms and their results.

Now, designing and programming interpretable predictors that offer some form of insight into the inner workings of black box algorithms does not entail that the problems posed by opacity have been answered. To be more precise, transparency is a methodology that does not offer sufficient reasons to believe that we can reliably trust black box algorithms. At best, transparency *contributes* to building trust in the algorithms and their outcomes, but it would be a mistake to consider it as a solution to overcome opacity altogether. To see this, consider P again, the interpretable predictor that shows the inner workings of A, the black box algorithm. The partisan of transparency, S, claims that P consists of a sequence of procedures of which a given p_i entails A (or some of its outputs). But what reasons does S have to believe this? All that S holds is a very appealing visual output produced by P, like heatmaps or decision trees, and the—still unjustified—belief that such an output represents the inner workings of A. For all S knows, P is as opaque as A (eg, it can misleadingly create clusters which are biased, it can ignore relevant variables and functions that compromise the integrity of the results, etc). It follows that all we can say is that P *induces* on S the belief that S knows the output of A (ie, the idea that A is transparent), but at no point P is offering genuine reasons to believe that S has interpreted A. For this to happen, for S to be justified in believing that A is transparent, P must be sanctioned as transparent too. The problem has now been shifted to showing that P is transparent.

[i]This could be the case of COMPAS, a highly accurate AI system used for measuring the risk of recidivism among defendants. Studies show that COMPAS has negatively influenced the judge's decisions.[57 58]

[ii]Relevance here is understood in epistemic terms. That is, surveying the algorithm to the extent that the outcome can be interpretable by humans.

The fundamental problem with transparency is, to our mind, that it is itself based on opaque processes. Indeed, transparency displaces the question of opacity of *A* to the question of opacity of *P*, taking the latter as non-problematic. But *P* is, strictly speaking, still opaque, despite its use for making *A* more transparent. In the face of it, *S* is restricted in the kind of knowledge claims about *A*. As we shall discuss in section 2.3, we do not need to give up the requirement of epistemic trustworthiness for AI systems, even for cases where transparency fails to deliver trust. We claim that CR credits trustworthiness to black box algorithms through processes exogenous to the algorithm itself, and which do not require the sort of surveyability of *A* needed by transparency. But before discussing this, we need to review the different forms of opacity found in algorithmic processes.

Epistemological and methodological opacity

Recently, Burrell proposed a very useful distinction among three types of opacity: opacity as intentional corporate or state secrecy, opacity as technical illiteracy, and opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them successfully.[22] Although all three forms of opacity are an attempt to entrench what can and cannot be said about AI systems, only the third form is directly related to algorithmic methods and knowledge. For this reason, we will focus our interest in methodological and epistemological opacity as the basis for normative assessments.

Simply put, methodological opacity stems from the complexities inherent to the design and programming of algorithms. Burrell, for instance, argues that major challenges of scale and complexity are distinctive to machine learning algorithms.[22] These challenges include designing and programming large amounts of lines of code per module, interlinkages among modules and subroutines, and combining different styles of programming among the team members (a more technically informed discussion on methodological opacity can be found in[30] (p. 103ff)). Colburn and Shute have also called attention to the different levels in which an algorithm is construed, and how researchers abstract from details about the implementation and mathematical modelling. These authors argue that it is impossible to design an algorithm 'without abstraction tools that hide, but do not neglect, details that are essential in a lower-level processing context.'[31] (p. 176). This form of abstraction is called *information hiding*, and it constitutes a form of abstraction uniquely introduced by computer-based scientific practice.[31 32] Examples include abstracting from the details how the messages among computer processes are passed on, how the computer hardware represents the value of parameters, and how programming languages handle irrational numbers, among other exclusively computational-related issues.[31] Consider a simple algorithm that shows on the screen the results of a blood analysis. The way in which the system represents each one of the values in the analysis (eg, haemoglobin, urea, etc) is irrelevant for the design and programming of the system (eg, whether they are integers, pointers, arrays, etc). Similarly, the way in which such information is stored and retrieved by the system is hidden—but not neglected—to the designer, programmer and user of the system. Owing to information hiding, the practice of designing and programming a medical AI is, by and large, *methodologically* opaque.

Epistemological opacity is arguably a more pressing issue for medical AI. There is a growing concern that the black box nature of AI makes it impossible to ground the reliability of the algorithm and, consequently, on whether researchers, physicians and patients can trust the results of such systems. The implications of a fully epistemically opaque algorithm are that medical AI work as truly obscure entities of which very little can be epistemically warranted. In this context, prime epistemic functions such as predictions, reporting evidence, understanding and explaining results are, *ex hypothesis*, stripped of their scientific value. Thus understood, epistemic opacity poses a major challenge for black box algorithms, namely, that there are no reasons for trusting the results. On the face of it, some scholars have suggested that we should abandon the use of black box algorithms in favour of more transparent ones.[24] Some others take a more pessimistic turn and propose to give up any aspiration to explain the 'why' and 'how' of certain outcomes of opaque algorithms.[33] Such viewpoints strike as both untenable and undesirable. On the one hand, and as argued before, transparency will not provide solutions to opacity, and therefore having more transparent algorithms is not a guarantee for better explanations, predictions and overall justification of our trust in the results of an algorithm. On the other hand, giving up explanation altogether (or reducing explanation to a handful of alleged transparent algorithms) defeats much of the purpose of implementing AI in medical practice. That is, having automated systems capable of handling extremely large amounts of data in short periods of time, reliably informing us about diseases and drug doses, and effectively suggesting safe treatments for a large number of illnesses is, indeed, a major leap forward for medical science.

To Burrell's mind, there is a viable alternative to opacity. According to the author, problems relating opacity and black box algorithms can be tackled by 'some combination of regulations or audits (of the code itself and, more importantly, of the algorithm's functioning), the use of alternatives that are more transparent (ie, open source), education of the general public as well as the sensitisation of those bestowed with the power to write such consequential code'.[22] As Burrell eloquently tells us, one possibility is that machine learning models are simplified by means of 'feature extractions,' understood as approaches consisting in removing features from the model that do not matter for the classification outcome. In other words, Burrell is resorting to some form of algorithmic *transparency* as a means to make visible the core processes that lead to a given outcome. Again, this is a solution that does not comes cheaply.

Whereas we heartily share Burrell's concerns, her proposal to tackle epistemic opacity by means of some form of algorithmic transparency are misplaced. As argued before, 'audits' understood as surveying the code and the algorithm's functioning is inherently unviable in AI systems. Of course, such audits could be conceived at a higher level, that is, at the level of algorithmic functions informing us of the ways in which some results have been obtained. But this move does not solve the problem of epistemic opacity nor offer reasons that justify our belief that the results are to be trusted. It rather displaces the problem to a higher level of analysis.

Methodological and epistemic opacity, as discussed so far, give way to serious moral concerns in the context of medical practice. Of particular interest to us is that certain actions are morally unjustified given the lack the epistemic warrants required for the action to take place. A physician is not morally justified in giving a certain treatment to a patient unless the physician has reliable knowledge that the treatment is likely to benefit the patient. As we argue next, aside from a justified belief, other conditions need to be met. The problem that emerges in the context of medical AI, and more to our interests, in the context of implementing black box algorithms for medical practice, is that transparency falls short of offering the right epistemic reasons for trusting the outcome of the algorithms, and that epistemic opacity is, *ex*

hypothesis, preventing any meaningful surveillance of the algorithm. But physicians require their beliefs to be epistemically justified before acting. Solving epistemic opacity, then, is prior to, and a condition for, moral justification. In what follows, we show how prime concerns posed by epistemic opacity can be circumvented through CR.

WHO IS AFRAID OF BLACK BOX ALGORITHMS? COMPUTATIONAL RELIABILISM AND TRUST

As medical AI becomes methodologically more complex and epistemically more opaque, it is paramount to offer solutions that neither require the reduction of the complexity of algorithms nor to shift the question of epistemic opacity to the treatment of transparency and auditability. That solution, we submit, is *computational reliabilism* (CR).[29] Properly understood, CR offers epistemic justification for the belief that the algorithm is reliable and its results are trustworthy. This, without necessitating to rely on external algorithms (such as interpretable predictors) or relinquishing black box algorithms altogether. In fact, CR becomes the solution to epistemic opacity in the most unusual way: it makes no attempts to solve it. The strategy proposed is rather to admit our cognitive limitations in surveying algorithms and to circumvent epistemic concerns by offering reasons for trusting the algorithm and its results. In the following, we shortly present and discuss CR as elaborated by Durán and Formanek.[29]

CR is presented as a framework for the justification of algorithmic procedures and the results they render. In a nutshell, CR states that researchers are justified in believing the results of AI systems because there is a reliable process (ie, the algorithm) that yields, most of the time, trustworthy results. More formally, CR asserts that 'the probability that the next set of results of a reliable (AI system) is trustworthy is greater than the probability that the next set of results is trustworthy given that the first set was produced by an unreliable process by mere luck'[29] (p. 654). This formal definition can be illustrated with a simple example. Consider a black box AI system (let's call it *dose-AI*) used for calculating the doses of chemotherapy needed for different types of cancer. CR says that medical personnel is justified in believing that the results of *dose-AI* are trustworthy because such results have been produced, most of the time, by a reliable AI algorithm. In simpler words, medical personnel are justified in trusting that a given dose for chemotherapy is right because *dose-AI* is a reliable medical AI system.

The challenge now is to spell out what makes *dose-AI*—and any other medical AI system—a reliable algorithm in the sense just given. To this end, Durán and Formanek present and discuss four reliability indicators for CR, namely, verification and validation methods, robustness analysis, a history of (un)successful implementations, and expert knowledge.[iii] Briefly, verification and validation are methodologies that build and measure the developer's confidence in the computer system. Whereas verification is the assessment of the accuracy of the solution to a computational model by comparison with known solutions, validation is the assessment of accuracy of a computational system by comparison with experimental data.[34] Robustness analysis, on the other hand, allows researchers to learn about the results of a given model, and whether they are an artefact of it (eg, due to a poor idealisation) or whether they are related to core features of the model[35] (p. 156). A history of (un)successful implementations draws on different scientific and engineering methodologies and practices related to designing, coding and running algorithms[29] (p. 661). Finally, expert knowledge encompasses the experts' judgements, evaluations and sanctioning on which many automated systems nowadays depend. As the authors claim, all four reliability indicators amount to offering a justification in believing that the results of medical AI systems are epistemically trustworthy.

Whereas we heartily endorse these reliability indicators for CR, we also call attention to the fact that Durán and Formanek are holding their discussion in the context of computer simulations, arguably a different kind of algorithm than those used in medical AI. Noticing this has two small, but rather significant implications: first, some indicators need to be readjusted for AI systems; second, some other indicators need to be added if we intend to cover the practice of design and programming medical AI.

Of the four reliability indicators mentioned by Durán and Formanek, we are mostly interested in expert knowledge since the practice of medicine and healthcare are highly expert-dependent. In many cases, the expert or experts are the ultimate deciding authority on a given subject. In particular, highly sensitive medical decisions are typically made in the context of plurality of opinions of experts. Technology only offers technical support.

Medical AI brings together standard medical knowledge as well as a myriad of information, such as data obtained from different tests which cannot be formalised in terms of a medical theory. It should then be expected that expert knowledge remains tailored to the theory and practice of medicine. This, insofar as we still want the medical expert to be at the centre of medical decisions. There is, arguably, an intentionality of replacing, at least in some cases and to some extent, the human expert by medical AI. But for now, it is reasonable to claim that the knowledge and experience provided by the expert have found no equal in automated decision-making. In particular, this is true not only because operationalising expertise is a colossal challenge in medical AI, but also because the conditions of reliability for the use of a medical AI system in one institution vary from another. Indeed, the degree of reliability of Watson for Oncology at the Memorial Sloan Kettering Cancer Center is different from those used in the Rigshospitalet in Copenhagen, as has been extensively discussed.[36–38]

As for additional reliability indicators, we believe that transparency, understood as a process that informs the inner workings of black box algorithms is the type of right indicator that will contribute to the overall reliability of algorithms. As suggested before, transparency by itself is necessary, although not sufficient for entrenching the reliability of black box algorithms and the overall trustworthiness of their results. In this respect, transparency in conjunction with CR might be defended as entrenching medical AI systems as reliable algorithms. Admittedly, more needs to be said about how such combination takes place, as well as the role and relations between transparency and CR for medical AI. Unfortunately, this is not the place for such discussion. Instead, we now turn to the ethical challenges that emerge in the context of reliable medical AI.

ETHICAL CHALLENGES FOR CLINICAL PRACTICE

As outlined in the previous sections, black box algorithms are methodologically and epistemically opaque systems, which can be deemed as reliable processes that can produce trustworthy

[iii] We do not deviate from these authors' interpretation of the reliability indicators. For further details, see ref. 29 .

results. The aggregation of evidence is an important part of diagnosing, treating and prediction in medicine, regardless of the methods used. Once such results are available, they are used as input for clinical decision making. Clinical decision making is the process where physicians, commonly together with the patient, interpret these results and decide how these findings can be acted on these findings in the particular case.[39] In this section, we will outline some challenges of black box algorithms with regard to interpretation of clinical data, responsibility, expertise and patient autonomy.

Clinical data and interpretation

Algorithms in medicine should be understood to provide *input* for clinical decision-making but they cannot decide by themselves how it should be acted on the results. Analysis of the data is certainly important and can be done both by algorithms and by physicians. There are three interrelated aspects when moving from aggregating evidence to clinical decision making:

First, clinical data can be interpreted in several ways with different moral consequences; there may be several 'correct' ways of handling based on the data provided by the algorithm. Take for instance an algorithm that can detect whether a preterm baby will develop late-onset sepsis, a dangerous and potentially life-threatening infection. If the algorithm detects an increased risk for such an infection, some physicians would, in terms of safety first, administer antibiotics as soon as possible in order to prevent a potential infection. Other physicians may be hesitant to act on the prediction and wait until the first symptoms of the infection present itself, also guided by a notion of safety first, because antibiotics have side-effects that can only be considered proportionate once it is certain the baby develops this infection.[40] The interpretation of the data by these two groups of physicians differs. Even if they are guided by the same leading principle, they interpret the data in different ways, leading to different actions with different moral consequences. While the first group of physicians run the risk of overtreatment and thereby unnecessarily exposing some babies to side effects of antibiotics as they would eventually not develop the infection, the second group runs the risk of undertreating an infection that could have been prevented. Both ways of acting can be supported by clinical data.

Second, as the example above illustrates, the same leading principle (ie, safety first) can direct to different ways of acting, yet physicians—and patients as we will see below—can also have different leading values in clinical decision making. Different treatment options (including the option to not provide a treatment) have different morally relevant consequences, meaning that such a choice requires a trade-off between different values. This further complicates such decisions and illustrates the necessity of deliberation to understand different perspectives. Despite several attempts to operationalise values and trade-offs in algorithms, there is no convincing way that an algorithm is, by itself, capable of making such decisions.[41–43]

A third and related aspect is that there is contention about many aspects of medicine. This means that many diagnoses are unclear, or it is disputed whether a certain illness even exists, or what an acceptable risk is with regard to treatments. Diagnostic and treatment decisions are fundamentally evaluative judgements for which risks and uncertainties have to be weighed against a backdrop of medical knowledge, expert knowledge and intuitions. In other words, if black box algorithms diagnose an illness and predicts which type of treatment would be most effective, the question what an *acceptable* and *desirable* way of acting is needs to be deliberated further based on this information, for which professional expertise and patient values are important.

Professional responsibility and expertise

Physicians have expertise beyond factual knowledge about evidence and data. Good medical practice requires good judgement, which entails interpretations of facts, weighing the evidence, as well as other intellectual tasks.[44] It is important to realise that clinical findings and evidence need to be interpreted and contextualised, regardless of the methods used for analysis (ie, opaque or not), in order to determine how these should be acted on in clinical practice. Professional expertise, therefore, also requires the ability to deal with uncertainty, risk, and other variables, as well as being able to deal with ethical questions that obviously cannot be answered based on data alone. This means that even if recommendations provided by the medical AI system are trusted because the algorithm itself is reliable, these should not be followed blindly without further assessment. Instead, we must keep humans in the loop of decision making by algorithms.[40 42] Even if black box algorithms become more technologically advanced and able to somehow include contextual factors and patient preferences in the assessment, there could still be good reasons not to follow the algorithm's recommendation blindly (see footnote 1). It follows that it is unlikely and undesirable for algorithms to replace physicians altogether, as some scholars have argued in favour of.[45–47]

Some have uttered concerns about the use of black box algorithms in clinical practice, as it would result in responsibility gaps: physicians cannot be held responsible for results of algorithms they do not understand.[23] If the algorithm makes a mistake, for example, by making the wrong diagnosis or recommending the wrong treatment, who should be responsible for these mistakes? In traditional medicine, physicians are the ones responsible for such decisions, regardless of whether their way of coming to a diagnosis is understandable for the patient. If they make a mistake they are expected to explain why and how they came to a diagnosis or recommendation and are held accountable. The argument against black box algorithms is that *because* these algorithms are epistemically and methodologically opaque, physicians cannot explain their results and therefore they do not have a proper understanding of inner workings.[23 48] Consequently, so it is argued, the physician cannot be held morally responsible.[49] This claim is contestable, not only because physicians typically operate other technologies and machinery which they do not fully understand or cannot fully explain the inner working of (think of MRI scans, eg), yet they are sufficiently in control and understand enough of the workings to be considered responsible for operating these machines, including mistakes caused by these machines.[50] Similarly, for medical AI physicians can be responsible, in terms of accountability, for using these devices without fully knowing or understanding their inner workings. Furthermore, it has been pointed out that demanding explainability, including full technical transparency, when using AI may be overdemanding, given that we generally accept *ex-post* explanations—and deem these sufficient—of human actors in decision-making.[51] Finally, Zerilli *et al* argue against the assumptions that medical AI must obey higher standards of transparency than ordinarily would be imposed on human decision-makers, and that human decisions can be effectively inspected. By rejecting these assumptions, responsibility can be ascribed to physicians when, under conditions of reliability, they were not morally justified in their actions.[51]

Autonomy of patients

Aside from clinical evidence and professional expertise, patient autonomy is a central element of clinical decision making. Based on the exchange of information about the diagnosis, possible

treatments and the value and preferences of patients, it can be determined what sort of treatment is most desirable. Some scholars warn that black box algorithms can hamper patient autonomy in clinical decision making.[23 52] It is for example argued that AI may reintroduce a paternalistic model of decision making, by ranking treatments according to effectiveness to increase the lifespan of patients. The worry is that patients whose values do not align with the values that are built into the algorithm, for example, patients who would decide based on a minimisation of suffering, may not have the possibility to choose a treatment based on their own values, thereby threatening patient autonomy.[41] These worries are understandable and would indeed be worrisome if black box algorithms would automatise decision making, *without* humans in the loop. This would mean that not only (1) evidence is synthesised automatically and methodologically opaque, but also that based on this evidence, (2) only one possible treatment can be suggested and (3) that the patient does not get to decide whether she/he considers this treatment acceptable. Note though that these problems are not caused by the opaqueness of the underlying algorithm but by the lack of choice provided.

These worries about patient autonomy are helpful to determine how medical AI should be developed. For instance, the AI system should not simply suggest the treatment that seems most effective, but rather complement it with a ranking of possible suitable treatments. Similarly, the output should be supplemented with information about risks and side effects by physicians or the AI system. Furthermore, the patient autonomy stresses the importance of clinical decision-making: recommendations of an algorithm still have to be discussed with the patient, which shows that physicians need to be able to interpret and explain the implications of the prioritised treatments, rather than having to be able to explain the algorithm itself. It also shows the importance of discussing with the patient which treatment, if any at all, would be best for the patient's needs/set of value. Even if the process in which the algorithm prioritises treatments is opaque, still a meaningful conversation about patient preferences and possible treatments can and should be facilitated to provide the patient with the opportunity to choose according to their preferences and values. Further reflection on what patients and healthcare professionals need in order to have a conversation about treatment options when using opaque algorithms is required and what sort of values should be addressed by them, but black box algorithms are not necessarily hampering patient autonomy.

DISCUSSION

We have argued that holding epistemic justified beliefs in the algorithm and its results are necessary, but not sufficient conditions for acting on these results in medical practice. In other words, we have argued that under conditions of epistemic reliability, physicians are justified in trusting the results of the algorithm without being normatively justified in acting based on this knowledge. CR, to our mind, offers the right epistemic conditions for the reliability of black box algorithms and the trustworthiness of results in medical AI.

Our analysis contributes to the current literature on the ethics of AI in a number of ways. We draw attention to the real possibilities of having reliable knowledge about black box algorithms and results without being morally justified. We thereby illustrate that the epistemology of algorithms is prior to, and at the basis of studies on the ethics of algorithms. The lack of epistemological analysis in the literature on the ethics of AI pushes many authors to resort to some strategy that justifies their moral claims. This is done, most commonly, by assuming that epistemic opacity is a condition of possibility for black box algorithms, and then conclude without further argumentation that physicians are not justified in believing the results of such algorithms.[24] Ethical concerns, then, are built around the fact that physicians and engineers can neither explain the results nor the algorithm. This strategy falls short in explaining the moral consequences for cases where knowledge is guaranteed. Another strategy is to plea for refraining from using black box algorithms and instead advocate for *interpretable AI*.[24] Such a plea begs the question of whether building interpretable algorithms won't defeat the purpose of having AI altogether.

It should be noted that we have argued why black box algorithms are trustworthy: CR justifies trusting the outcome of opaque algorithms. Nevertheless, the public perception may differ, regardless of the reasons for trusting these algorithms. The general public may be sceptical or actually distrust such algorithms when applied in the medical context, which may be a problem for the public acceptance of medical AI. It has been argued that transparency may be required to foster public acceptance of AI.[53 54] While the acceptance of medical AI by physicians, patients and the general public is important for its implementation and use, it is questionable whether transparency is indeed as important for this acceptance as has been suggested. A few qualitative studies indicate that reliability in the medical AI system, appropriate training of physicians as well as keeping physicians in the loop, and the improvement of the diagnosing process are reasons for patient acceptance and trust in medical AI.[55 56]

Our analysis should not only be relevant for the academic debate on trustworthy AI and nuance the stance to opaque algorithms and its limitations, but also for clinical practice as well. Our analysis indicates that close collaboration and exchange between clinical and informatics experts is required when black box algorithms are being used in or developed for medical contexts. Given our plea to keep humans in the loop, it is important that clinicians are educated and informed about limitations and shortcoming of models that they rely on. This is not to say that clinicians should fully understand the algorithm used, but they should be able to work with it and rely on its workings. How such systems present their recommendations should be aligned with different possible interpretations, the presentation of the results should reflect different options and values of patients to enable good clinical decision making.

CONCLUSION

We have made an effort to show that black box algorithms in medicine can be credited reliable to the extent that physicians are justified in trusting their results. It follows that the implementation of medical AI in daily clinical practice can bring substantial benefits, even in cases where black box algorithms are the predominant tool. The reflection on the limitations of algorithms offers, indeed, insight into ways to improve their use. To our mind, being aware of the epistemic limitations of medical AI is a condition for entrenching responsible use and interaction with such systems. For these reasons, we believe that the debate needs to be widened by not solely focusing on the technological aspect of medical AI, but also on the interaction of humans with such technological systems. Indeed, while black box algorithms may provide challenges for their application in medicine, we should not dismiss their use altogether. Rather, the concerns and problems emerging in this context should guide the development of such technologies as well as the training of physicians

and medical informatics in order to equip these professionals to integrate opaque systems in good medical practice.

Contributors Both authors contributed equally. Both authors provided critical intellectual input and revisions. All authors read and approved the final manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement There are no data in this work.

ORCID iDs
Juan Manuel Durán http://orcid.org/0000-0001-6482-0399
Karin Rolanda Jongsma http://orcid.org/0000-0001-8135-6786

REFERENCES

1 Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. *NPJ Digit Med* 2018;1.
2 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44–56.
3 Esteva A, Robicquet A, Ramsundar B, *et al*. A guide to deep learning in healthcare. *Nat Med* 2019;25(1):24–9.
4 Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *JAMA* 2018;320(21):2199–200.
5 Pasquale F. *The black box Society*. Harvard University Press, 2015.
6 Humphreys P. The philosophical novelty of computer simulation methods. *Synthese* 2009;169(3):615–26.
7 Rajpurkar P, Irvin J, Zhu K. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv* 2017.
8 Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;375(13):1216–9.
9 Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. *NPJ Digit Med* 2018;1(1).
10 Berner ES, La Lande TJ. Overview of clinical decision support systems. In: Berner ES, ed. *Clinical decision support systems: theory and practice*. Cham: Springer, 2016: 1–17.
11 Athey S. Beyond prediction: using big data for policy problems. *Science* 2017;355(6324):483–5.
12 European Group on Ethics in Science and New Technologies. *Statement on artificial Intelligence, robotics and 'autonomous' systems*. Brussel: European Commission, 2018.
13 Esteva A, Kuprel B, Novoa RA, Novoa J, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115–8.
14 Kemper J, Kolkman D. Transparent to whom? no algorithmic accountability without a critical audience. *Inf Commun Soc* 2019;22(14):2081–96.
15 Martin K. Ethical implications and accountability of algorithms. *J Bus Ethics* 2019;160(4):835–50.
16 Ananny M, Crawford K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc* 2018;20(3):973–89.
17 Mittelstadt BD, Allo P, Taddeo M, *et al*. The ethics of algorithms: mapping the debate. *Big Data Soc* 2016;3(2):205395171667967–21.
18 Barocas S, Selbst A. Big data's disparate impact. *Calif Law Rev* 2016;104(1):671–729.
19 Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science* 2017;356(6334):183–6.
20 O'Neil C. *Weapons of math destruction: how big data increases inequality and threatens democracy*. New York: Crown Publishing Group, 2016.
21 van Amsterdam WAC, Verhoeff JJC, de Jong PA, *et al*. Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning. *NPJ Digit Med* 2019;2(1).
22 Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data Soc* 2016;3(1):1–12.
23 Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics* 2020;46(3):205–11.
24 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;1(5):206–15.
25 Humphreys PW. *Extending ourselves: computational science, empiricism, and scientific method*. Oxford University Press, 2004.
26 Newman J. Epistemic opacity, confirmation holism and technical debt: Computer simulation in the light of empirical software engineering. In: F GMT, ed. *History and Philosophy of Computing" Third International Conference, HaPoC 2015*. Springer, 2015: 256–72.
27 Symons J, Alvarado R. Epistemic Entitlements and the practice of computer simulation. *Minds Mach* 2019;29(1):37–60.
28 Guidotti R, Monreale A, Ruggieri S, *et al*. A survey of methods for explaining black box models. *ACM Comput Surv* 2019;51(5):1–42.
29 Durán JM, Formanek N. Grounds for trust: essential Epistemic opacity and computational Reliabilism. *Minds Mach* 2018;28(4):645–66.
30 Durán JM. *Computer simulations in science and engineering. Concepts - Practices - Perspectives*. Springer, 2018.
31 Colburn T, Shute G. Abstraction in computer science. *Minds Mach* 2007;17(2):169–84.
32 Colburn TR, The Hegeler Institute. Software, abstraction, and ontology. *Monist* 1999;82(1):3–19.
33 Dwork C, Hardt M, Pitassi T. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12*, 2012:214.
34 Oberkampf WL, Roy CJ. *Verification and validation in scientific computing*. Cambridge: Cambridge University Press, 2010.
35 Weisberg M. *Simulation and similarity*. Oxford University Press: Oxford, 2013.
36 Choi YI, Chung J-W, Kim KO, *et al*. Concordance rate between clinicians and Watson for oncology among patients with advanced gastric cancer: early, real-world experience in Korea. *Can J Gastroenterol Hepatol* 2019;2019:1–6.
37 Vulsteke C, del P, Ortega Arevalo M. Artificial intelligence for the oncologist: hype, hubris, or reality? *Belgian J Med Oncol* 2018;12(7):330–3.
38 Hamilton JG, Genoff Garzon M, Westerman JS, *et al*. "A Tool, Not a Crutch": Patient Perspectives About IBM Watson for Oncology Trained by Memorial Sloan Kettering. *J Oncol Pract* 2019;15(4):e277–88.
39 Sandman L, Munthe C. Shared decision making, paternalism and patient choice. *Health Care Anal* 2010;18(1):60–84.
40 Big data for small babies project. Available: https://www.finaps.nl/casestudies/predictive-analytics-solution/ [Accessed 29 Dec 2019].
41 McDougall RJ. Computer knows best? the need for value-flexibility in medical AI. *J Med Ethics* 2019;45(3):156–60.
42 Hodgkin PK. The computer may be assessing you now, but who decided its values? *BMJ* 2016;355.
43 van de Poel I, Poel vande I. Embedding values in artificial intelligence (AI) systems. *Minds Mach* 2020;30(3):385–409.
44 Davis M. A plea for judgment. *Sci Eng Ethics* 2012;18(4):789–808.
45 Goldhahn J, Rampton V, Spinas GA. Could artificial intelligence make doctors obsolete? *BMJ* 2018;363.
46 Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. *JAMA* 2016;315(6):551–2.
47 Coiera E. The fate of medicine in the time of AI. *Lancet* 2018;392(10162):2331–2.
48 Adadi A, Berrada M. Peeking inside the black-box: a survey on Explainable artificial intelligence (XaI). *IEEE Access* 2018;6:52138–60.
49 Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics Inf Technol* 2004;6(3):175–83.
50 Wolkenstein A, Jox RJ, Friedrich O. Brain-Computer interfaces: lessons to be learned from the ethics of algorithms. *Camb Q Healthc Ethics* 2018;27(4):635–46.
51 Zerilli J, Knott A, Maclaurin J, *et al*. Transparency in algorithmic and human decision-making: is there a double standard? *Philos Technol* 2019;32(4):661–83.
52 Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018;15(11):e1002689.
53 de Fine Licht K, de Fine Licht J. Artificial intelligence, transparency, and public decision-making. *AI Soc* 2020;35(4):917–26.
54 Hatherley JJ. Limits of trust in medical AI. *J Med Ethics* 2020;46(7):478–81.
55 Fink C, Uhlmann L, Hofmann M, *et al*. Patient acceptance and trust in automated computer-assisted diagnosis of melanoma with dermatofluoroscopy. *J Dtsch Dermatol Ges* 2018;16(7):854–9.
56 Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, *et al*. Artificial intelligence in skin cancer diagnostics: the patients' perspective. *Front Med* 2020;7(233).
57 Brennan T, Dieterich W, Ehret B. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Crim Justice Behav* 2009;36(1):21–40.
58 et alAngwin J, Larson J, Mattu SL. Machine bias. pro Publica, 2016. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Accessed 16 Feb 2021].