Philosophy & Technology (2021) 34:349–371
https://doi.org/10.1007/s13347-019-00391-6

RESEARCH ARTICLE

Artificial Intelligence and Patient-Centered Decision-Making

Jens Christian Bjerring[1] · Jacob Busch[1]

Received: 20 December 2018 / Accepted: 15 December 2019 / Published online: 8 January 2020

Abstract

Advanced AI systems are rapidly making their way into medical research and practice, and, arguably, it is only a matter of time before they will surpass human practitioners in terms of accuracy, reliability, and knowledge. If this is true, practitioners will have a prima facie epistemic and professional obligation to align their medical verdicts with those of advanced AI systems. However, in light of their complexity, these AI systems will often function as black boxes: the details of their contents, calculations, and procedures cannot be meaningfully understood by human practitioners. When AI systems reach this level of complexity, we can also speak of black-box medicine. In this paper, we want to argue that black-box medicine conflicts with core ideals of patient-centered medicine. In particular, we claim, black-box medicine is not conducive for supporting informed decision-making based on shared information, shared deliberation, and shared mind between practitioner and patient.

Keywords Black-box medicine · Patient-centered medicine · Evidence-based medicine · Medical decision-making · Artificial intelligence and medicine

✉ Jens Christian Bjerring
filjcb@cas.au.dk

Jacob Busch
filjab@cas.au.dk

[1] Department of Philosophy, Aarhus University, Jens Chr. Skous Vej 7, DK-8000 Aarhus C, Denmark

1 Introduction

AI systems are rapidly making their way into medical research and practice.[1] Various machine learning algorithms help practitioners estimate probabilities of specific diseases and suggest possible courses of treatment, and various natural language processing algorithms help extract information from medical journals or personal self-tracking devices. More than 100 companies are by now involved in developing AI systems for use in the healthcare sector, and AI applications can by 2026 potentially yield a $150 billion in annual savings for the US healthcare economy (Purdy and Daugherty 2016).[2] Indeed, as Esteva et al. (2019) claims, "[h]ealthcare and medicine stand to benefit immensely from deep learning because of the sheer volume of data being generated (150 exabytes or 10^{18} bytes in United States alone, growing 48% annually) as well as the increasing proliferation of medical devices and digital record systems" (Esteva et al. 2019, p. 24).

To be sure, much of the technology is still in its infancy, but already there are indications that at least some AI systems are as reliable and accurate as medical experts when it comes to diagnosing diseases. Deep learning networks, for instance, have successfully been applied in diagnostic imaging to make probabilistic estimates about breast and skin cancer that match those of experienced professionals (Jiang et al. 2017). Indeed, when it comes to classifying diseases using medical imaging, the recent comprehensive meta-analysis in Liu et al. (2019) yields strong support for the claim that deep learning networks can match the diagnostic performances of healthcare professionals—although it is recognized that little literature exists that actually compares the performances of deep learning networks and professionals on the same data sets. In a qualitative survey from 2017, several references are provided to studies, which, according to the author, give us reasons to believe that "AI has now been shown to be as effective as humans in the diagnosis of various medical conditions, and in some cases, more effective" (Loh 2018, p. 59). More speculatively, in Obermeyer and Emanuel (2016), it is predicted that machine learning algorithms in the near future "will displace much of the work of radiologists and anatomical pathologists" (Obermeyer and Emanuel 2016, p. 1218). In point of fact, the US Food and Drug Administration recently approved the first AI device that provides screening decisions for diabetic retinopathy without assisted interpretation by a clinician (US Food and Drug Administration 2018).

Of course, we should be careful not to exaggerate the epistemic wonders of current AI systems. But in light of the impressive work that AI systems are already performing in medicine—we will point out some more results in section 2 to further motivate this claim—we can at least with some warrant adopt the assumption that AI systems will eventually outperform human practitioners in terms of speed, accuracy, and reliability when it comes to predicting and diagnosing central disease types such as cancer, cardiovascular diseases, and diseases in the nervous system (Jiang et al. 2017, p. 231). Moreover, since current AI systems can already scan through thousands of documents

[1] Later, in section 2, we will be more precise about the types of AI systems that we have in mind, but for now we can keep the term "AI systems" vague and apply it at a level of generality that mirrors the use of "autonomous/intelligent systems" employed by IEEE.

[2] For an overview of the range of applications of AI in healthcare, refer to (He et al. 2019), (Jiang et al. 2017), and Kallis et al. (2018).

and samples per second, we can also with some warrant adopt the assumption that AI systems eventually will be more knowledgeable than human practitioners. For in contrast to human practitioners—and even groups of human practitioners—it is not unreasonable to hold that AI systems eventually will be able to make verdicts that are informed by all the available medical evidence in a given domain.

Given these assumptions about (future) AI systems, it is pertient to ask how the use of AI systems in medical decision making will affect commonly held conceptions of what constitutes good practice in healthcare. On the one hand, insofar as AI systems will eventually outperform the best practitioners in specific medical domains, practitioners will have *epistemic obligation* to rely on these systems in medical decision-making. After all, if a practitioner knows of an epistemic source that is more knowledgeable, more accurate, and more reliable in decision-making, she should treat it as an expert and align her verdicts with those of the source. Of course, we may imagine ways in which such an obligation can be overridden by other epistemic or non-epistemic factors, but as a working assumption, it seems highly reasonable to proceed on the thought that practitioners have an obligation, everything being equal, to follow epistemically superior AI systems in decision-making.[3] Yet, on the other hand, if it turns out, as we shall argue in this paper, that the reliance on AI systems can ultimately conflict with values and ideals that we otherwise deem central to medical practice, then there might be a reason for practitioners, patients, and policy makers alike to dispense with this epistemic obligation and critically engage in a discussion about the extent to which we should allow AI-based decision-making in medicine.[4]

More specifically, we will argue that reliance on AI systems in medical decision-making conflicts with core ideals of *patient-centered medicine*. Although we shall be more precise in section 3, the basic idea behind this paradigm is that medical practice should serve to support informed and autonomous patient decision-making by establishing "a state of shared information, shared deliberation, and shared mind" between practitioner and patient (Epstein et al. 2010, p. 1491). But this idea, we argue in section 4, is not tenable in light of the *black-box* nature of the machine learning algorithms that are predicted to play a central role in AI-assisted medical decision-making; as Topol puts it, "[a]lmost every type of clinician, ranging from specialty doctor to paramedic, will be using AI technology, and in particular deep learning, in the future" (Topol 2019, p. 44). Following Price II (2018), we can adopt the term "black-box medicine" to give an initial characterization—the precise characterization will come in section 2—of medical practices in which black-box algorithms are in the decision-theoretic driving seat:

[3] What may such overriding reasons look like? To give an example, suppose you—along the lines of Montgomery (2006), for instance—believe that there is a practical and irreducible know-how component to medical judgment and decision-making, which, importantly, cannot meaningfully be encoded in AI systems. If so, we can imagine how an appeal to such medical know-how may help us explain how a practitioner can be excused from acting in accordance with the recommendations of an epistemically superior AI system. Alternatively, as recently argued by Ploug and Holm (2019), it might be that patients have a medical *right* to withdraw from AI-assisted diagnostics and treatment. In order to respect such rights, it may be argued, practitioners can be excused from acting in accordance with the recommendations of AI systems. While there are lots to say about these issues, it is well beyond the scope of a single paper. For present purposes, we shall proceed on the assumption that practitioners are under an obligation to follow (future) black-box AI systems in decision-making.

[4] For some recent discussions of some of the broader ethical and legal issues surrounding the use of AI in healthcare, see Ploug and Holm (2019) and Schönberger (2019).

> The explosive proliferation of health data has combined with the rapid development of machine-learning algorithms to enable a new form of medicine: black-box medicine. In this phenomenon, algorithms troll through tremendous databases of health data to find patterns that can be used to guide care, whether by predicting unknown patient risks, selecting the right drug, suggesting a new use of an old drug, or triaging patients to preserve health resources. These decisions differ in kind from previous data-based decisions because black-box medicine is, by its nature, opaque; that is, the bases for black-box decisions are unknown and unknowable. (Price II 2018, p. 295.)

Put in these terms, our central argument will be that black-box medicine conflicts with central ideals of patient-centered medicine. As a result, the increased reliance on AI in medical decision-making presents a challenge for proponents of patient-centered medicine: if a practitioner honors her epistemic obligation, she will align her medical verdict with that of the superior AI system, but in doing so, she will end up violating central tenets of patient-centered medicine.[5]

While you may agree with our central conclusion, you might worry that it is old wine in new bottles.[6] Since opaque decision-making is already common in good old-fashioned medicine, goes the worry, black-box medicine does not introduce a fundamentally new epistemic challenge to patient-centered medicine. In section 5, we address this worry by arguing that there are many interesting cases in which opaque AI-assisted decision-making comes apart from ordinary opaque medical decision-making. Based on these differences, we claim, there is reason to think that black-box medicine poses an interestingly new challenge to current ideals in medicine.

It is not our aim in this paper to resolve the conflict between black-box medicine and patient-centered medicine. Rather, our main aim is to unfold the argument for why the conflict exists. Granting that our argument is successful, some might attempt to preserve the central ideals of patient-centered medicine by either restricting the use of AI systems in medicine to transparent or non-opaque systems, or by putting faith in the prospects of "explainable" or "transparent AI" and the hope of eventually creating transparent AI systems that perform as well—or at least nearly as well—as typical black-box systems such as deep neural networks.[7] Alternatively, some might attempt to relax or even give up some of the ideals of patient-centered medicine to benefit from the superior accuracy and reliability of black-box AI systems. While each of these options deserves a full exploration, we cannot hope to accomplish it here. Also, there are familiar concerns about the wide-scale use of artificial intelligence in society—legal and ethical concerns about privacy and biases and value-ladenness in algorithmic processing, to name just a few—that may well have an influence on how we should adjudicate the conflict between black-box medicine and patient-centered medicine. Again, while these issues deserve a full treatment, we shall set them aside here and instead focus on the specific conflict that we claim arises between black-box medicine and patient-centered medicine.

[5] In fact, as we shall point out in section 6, black-box medicine also presents a challenge to central elements of evidence-based medicine.

[6] We are grateful to an anonymous referee for stressing this worry.

[7] For more on explainable AI, see Burrell (2016), Doran et al. (2017), Holzinger et al. (2017), and Wachter et al. (2017). For general discussions on the value of transparency in decision-making, see Forssbæck and Oxelheim (2014), Heald (2006), Miller (2018), and Prat (2006).

2 Black-Box Medicine

Generally, AI systems in healthcare can be divided into systems that operate on structured data and systems that operate on unstructured data. Structured data include imaging and genetic data, and typical AI systems that operate on such data include machine learning systems and deep learning networks. Neural networks have been put to work in analyzing clinical images of skin lesions, and applications of deep convolutional neural networks have been proven effective in "demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists" (Esteva et al. 2017, p. 115). Likewise, developers of machine learning systems claim that the accuracy of these systems is higher than that of trained professionals in areas such as magnetic resonance (MR) imaging (MRI) interpretation.[8] For instance, in cases of distinguishing radiation necrosis from recurrent brain tumors, support vector machine classifiers have been shown to identify "12 of 15 studies correctly, while neuroradiologist 1 diagnosed 7 of 15 and neuroradiologist 2 diagnosed 8 of 15 studies correctly, respectively", where neuroradiologists 1 and 2 are "expert neuroradiologists who had access to the same MR imaging sequences (gadolinium T1WI, T2WI, and FLAIR) as the classifier" (Tiwari et al. 2016, p. 2231). Finally, in a recent study published in *Nature*, a deep learning architecture demonstrates "performance in making a referral recommendation that reaches or exceeds that of experts on a range of sight-threatening retinal diseases after training on only 14,884 scans" (De Fauw et al. 2018, p. 1342). In particular, the architecture has a 5.5% error rate compared to the 6.7% and 6.8% error rates of the two best retina specialists in the study.[9]

AI systems that operate on unstructured data are less commonly used for diagnostic purposes but rather as tools for mining data, for processing texts from unstructured sources such as laboratory reports and medical literature, and for assisting clinical decision-making. As technological developments accelerate, natural language processing algorithms can be of huge potential value not only to doctors but also to other AI systems. For instance, natural language algorithms might be employed to translate information from various physical examination reports in, say, medical English into a structured format that can be utilized by other AI systems.[10] Given the speed by which

[8] A central reason for why automated image recognition has seen such great progress in recent years is in large part due to the high *quality* of imaging data. For recent information about the progress being made to improve the quality of imaging data sets even further, see Harvey and Glocker (2019) and van Ooijen (2019).

[9] Note that the reported result does not show that deep learning systems *generally* outperform clinicians and experts. As reported by De Fauw et al. (2018), differences in performance between deep learning systems and clinicians were considerably reduced when clinicians had access to all the information—such as patient history and clinical notes—that they ordinarily make use of in addition to an OCT scan; for conclusions that point in a similar direction, see Faes et al. (2019). As a reviewer for this journal interestingly noted, this observation might point toward a correlation between the amount of information that a clinician has and the degree to which he or she is epistemically obliged to rely on AI systems in decision-making. In particular, since junior clinicians might have access to less information than an experienced clinician, the former might be under a greater epistemic obligation to take the output of an AI system at face value than the latter. While these discussions are obviously interesting for current practices in AI-assisted medical decision-making, we assume, as mentioned, that AI systems (will) have access to more information than even expert human practitioners. As such, we need not worry too much about results showing that there is not a great difference in performance between humans and AI systems when they have access to roughly the same amount of medical information.

[10] In Holzinger et al. (2019), for instance, it is stressed how AI performance can be optimized by integrating multiple independent data sets (e.g., imaging, omics profiles, and clinical data).

AI systems can mine existing data, it is not hard to imagine the potential explosion of structured data that will be accessible to AI systems; IBM's Watson, for example, has been claimed to be able to process 500,000 research papers in about 15 s (Captain 2017). So if the training data is solid and up to date, and if the potential supervised training is performed by experts, it is not unreasonable to expect that AI systems will eventually have at their disposal most of the medical evidence that is relevant in a specific context.

To repeat, we should take preliminary reports on the reliability of AI systems in medicine with a grain of salt. As of yet, it is hard to reproduce the results of various machine learning studies, which makes it difficult to convince the medical community of their reliability and accuracy (Faes et al. 2019; Olorisade et al. 2017). As of yet, there are no standardized performance metrics for machine learning systems, which makes it hard to compare these systems with each other and with human practitioners (Japkowicz and Shah 2011). And, as of yet, there is no requirement that AI developers publish information about situations in which their systems fail. For the purposes of this paper, however, we can assume that such "childhood diseases" have been eradicated and instead focus solely on AI systems that—by assumption—outperform human practitioners in terms of accuracy, reliability, and knowledge in a given medical domain.

Granting these assumptions about AI systems, it seems, as mentioned, clear that practitioners (will) have an *epistemic* obligation to align their medical verdicts with those of AI systems. Essentially, the relationship between practitioner and AI system is epistemically analogous to the relationship between practitioner and expert. If a general practitioner is unsure about how to interpret, say, a MR image, it seems perfectly reasonable to say that he should adopt the opinion that an expert in MRI interpretation would form after inspecting the image. In part he should do so because the expert is more knowledgeable about the medical domain, more reliable in her diagnostic work, and better at making accurate predictions based on the image. But these platitudinous claims also seem true when we focus on the relationship between practitioner or expert and AI system. For on the assumption that the AI system outperforms the expert on all parameters on which the expert outperforms the general practitioner, the AI system will play the expert role in the relationship.

As much as this is clear, however, it is also clear that the inclusion of AI systems in medical decision-making can incur an epistemic *loss* in medical understanding and explanation. As mentioned, machine learning techniques and deep learning networks are expected to play a central role in medical decision-making, and they are often, as seen above, highlighted as examples of AI systems that can epistemically outperform human practitioners. But deep learning networks are also often cited as prime examples of black-box or opaque AI systems:

> "This is especially true of some top performing algorithms, like the deep neural networks used in image recognition software. These models may reliably discriminate between malignant and benign tumours, but they offer no explanation for their judgments." (Watson et al. 2019, p. 2 out of 4)

> "[...] the black box nature of the current deep learning technique should be considered. Even when the deep learning based method shows excellent results, in many occasions, it is difficult or mostly impossible to explain the technical and logical bases of the system." (Lee et al. 2017, p. 580)

> "Despite this accuracy, deep learning systems can be black boxes. Although their designers understand the architecture of these systems and the process by which they generate the models they use for classification, the models themselves can be inscrutable to humans." (London 2019, p. 17)

So, roughly put, if we rely on deep learning networks for medical decision-making, we risk losing medical understanding because of their black-box nature: we cannot understand nor explain how and why they make the medical recommendations that they do based on the inputs that they receive. Let us put some flesh on this claim.[11]

Simplistically, a (vanilla type) deep learning network consists of an input layer, an output layer, and a sequence of multiple hidden layers in between. Each layer consists of a number of neurons or units that are characterized functionally: they receive as inputs the unit activations from the previous layer and perform a simple computation on the input—for instance, by taking the weighted sum of the input values—the result of which is passed on to an activation function, typically a certain kind of nonlinear function, which produces the unit's output. A unit's output is then used as inputs for units in subsequent layers. So a deep neural network is essentially a collection of units that jointly implement a highly complex (nonlinear) function that maps a given input to the network to a particular output. The network learns this functional mapping through training on data. As data accumulates, the network utilizes various algorithms to adjust the weights of the individual units by automatically adjusting the weights of the connections between units in different layers with a view toward minimizing the differences between inputs and desired outputs. As a result of the way the network adjusts the weights, "internal hidden units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units" (Holzinger et al. 2017, p. 5). During training, that is, the network automatically extracts its own features by aggregating and recombining sets of features from previous layers, moving sequentially layer by layer from various low-level features to more high-level and abstract representations of the data.

Even only equipped with this general characterization, it is clear why we are tempted to describe the process from input to output in a deep learning network as opaque. For if the hidden layers in a network contain millions of weights and thousands of distinct features, and if the network is trained on data sets that far surpass what individual, and even groups of individual experts can be expected to comprehend, it is no wonder that we struggle to explain why the network yields the predictions or recommendations that it does. More precisely, we can say, we struggle to explain why a deep learning network

[11] We will offer a somewhat detailed explanation of the sense in which deep learning networks count as black-box AI systems, in part because it is useful to have a clear understanding of what makes an AI system a black-box system, and in part because we will appeal to deep neural networks in our discussions in sections 4 and 5.

yields a particular prediction because the correlations within the data, which the network utilizes for decision-making, are not based on well-understood principles, rules, and criteria. This contrasts with more traditional model building where we often can predict the behavior of a model because of the way we have *constructed* its elements. That is, we articulate which features of the data should be represented as features in the model, and we attempt to implement into the model adequate responses—for instance, through classical if-then-else inference rules—to a range of possible inputs. Largely, the adequate responses that we want our model to give are informed by antecedent knowledge of certain principles and rules that we believe sensibly can help us explain the data. As such, the model's verdicts reflect to a certain extent the explanatory expectations that we have, and in applying the model, we understand the reasoning behind its predictions.

To illustrate, consider a basic expert system in AI for a given task domain. Such a system consists of a knowledge base, an inference engine, and a user interface. The user interface takes a query and passes it to the inference engine. Once it receives the result from the inference engine, it produces an answer to the query. The knowledge base is a repository of facts that stores expert knowledge about the task domain. The inference engine consists of a list of reasoning methods, principles, or rules that allow the system to draw inferences from the knowledge base. In crafting a particular model of an expert system, a knowledge engineer may extract knowledge from a group of human experts by interviewing and observing how they express their knowledge and how they reason with it. The engineer translates the knowledge into a computer language and designs an inference engine that codifies the types of inferences that the experts make. In simple cases, the inference engine might consist of a long list of "if, then" rules, which map different elements of the knowledge base, but they may of course also be much more complex. The important point is not the details but rather the process of constructing the model: the whole system encodes information that we understand and principles and rules that we know. Based on knowledge of the system, we can (ideally) not only predict which answers the system will give but also (ideally) explain why it gives a particular answer through a reconstruction of the sequence of rule applications that the system utilizes. Given that the rules codify inferential patterns that experts actually employ, such a reconstructive explanation will be informative—at least for experts in the domain. In this sense, expert systems are organized around well-understood principles or rules that allow us to explain why the system yields the predictions that it does.

Deep learning networks—and many other machine learning systems—do not have these features. Rather, when it comes to deep learning networks, the guiding thought is that it is "far easier to train a system by showing it examples of desired input-output behavior than to program it manually by anticipating the desired response for all possible inputs" (Jordan and Mitchell 2015, p. 255). Effectively, this means that the behavior of a deep learning network cannot be anticipated before the network is up and running. The behavior of the network is built up, so to speak, on the fly when the network begins to adjust its parameters to the training data. As a result,

> "the operations and the behavior of a [deep learning network] cannot in principle be anticipated during its set-up, since they are only determined by, say, the 'history' or 'experience' of the [deep learning network] made through its training.

> [...] The programmer has to check if the [deep learning network] works as it is supposed to by using parts of the training data not for training, but for testing the trained model. Thereby, the error rate of the [deep learning network] can be calculated." (Schubbach 2019, p. 10)

So the lack of an antecedently fixed set of domain principles and rules makes models for deep learning networks significantly different from more classical models in AI. Since we do not construct a deep learning network by encoding principles that we take to reflect antecedent knowledge of the relevant data—be they logical, semantic, mechanistic, or causal—we can only attempt to extrapolate an explanation of the network by looking at its *actual* behavior.

However, it can be close to practically impossible to extrapolate an explanation of a deep learning network by looking at its actual behavior. While it is easy enough to understand the mathematical functions that characterize each unit in a layer, these functions do not—in abstraction from the specific weights of the specific network—help us explain why a particular network predicts as it does. The "reason is that the algorithm used to calculate the input of different [deep learning systems] is every time the same, it is only parametrized by different numbers, i.e. the weights" (Schubbach 2019, p. 15).[12] So to explain why a particular deep learning network behaves the way it does, we must appeal to the specific weights of the units in the network. But clearly, when the number of weights or actual parameters reach into the millions, there is no guarantee that we, as finite beings, can extrapolate a sensible explanation from the model. For we would be hard struck, to put it mildly, to comprehend an absurdly complex mathematical function that involves such a large number of parameters. Moreover, even if the number of relevant parameters could somehow be reduced by (automatically) clustering various features of the network's data, there is no guarantee that we would be able to understand what these aggregated features represent—we will give an example in section 5.

So we can seemingly explain the behavior of a deep learning network neither through a predefined framework of well-understood principles and rules nor through observation of the network's actual behavior. Plausibly, these features of deep learning networks help us appreciate why they are often characterized as black-box systems: we can observe what a deep learning network does in terms of its input-output behavior, but we cannot explain why it correlates a particular input with a particular output.

On the assumption that deep learning networks will be widely employed in medical decision making, it is thus clear how they can incur a loss in medical understanding and explanation. Typically, we gain medical understanding by producing medical explanations, and medical explanations typically appeal to some causal or mechanistic structure.[13] In particular contexts of diagnosis and treatment, as we can put it, practitioners typically rely on their being some causal or mechanistic explanation of why certain symptoms are correlated with certain diagnostic outputs and treatment suggestions. It may happen, of course, that individual practitioners are not able to

[12] Schubback references Smolensky (1988) for this observation.

[13] Most of the types of medical explanations discussed in Marcum (2008, chap. 8) include such causal or mechanistic components. Also, mechanistic and causal considerations play prominent roles in many contemporary views on abductive inferences; see, for instance, Lipton (2003).

provide the relevant explanation themselves, but at least "we expect that experts can marshal well-developed causal knowledge to explain their actions or recommendations" (London 2019, p. 15).[14] But when it comes to deep learning networks, we do not have these assurances. Since not even expert data scientists, as motivated above, may be able to explain why a particular output of a deep learning network is correlated with a particular input, we obviously have no guarantee that there exists a causal or mechanistic explanation of why a successful network produces the recommendations it does.

With deep learning networks serving as prime examples of black-box systems, let us then, abstractly speaking, characterize a black-box AI system as a system in which we can control the inputs and observe the corresponding outputs, but in which we have no explanation of why the input is correlated with the output. In turn, let us characterize "black-box medicine" as the kind of medical practice in which black-box AI systems play an essential role in decision-making. As characterized, we can think of black-box medicine as a subspecies of AI-informed medicine, where "AI-informed medicine" can be understood as a label that covers medical practices in which AI systems—whether opaque or transparent—play an essential role in decision-making.[15]

Thus, so far: as in the orginal), based on the assumed superior reliability, accuracy, and knowledge of black-box AI systems, black-box medicine promotes higher accuracy and reliability in medical decision-making, but it does so at the cost of incurring a loss in medical understanding and explanation. And this loss in medical understanding, as we shall argue next, has troublesome consequences for a central paradigm in contemporary medical practice: the so-called *patient-centered* paradigm.

3 Patient-Centered Medicine

In healthcare, we have long moved beyond the conception of patients as mere bundles of diseases:

> "Recent decades have seen what some colleagues have described as a paradigm-shift from a singular focus on disease towards a focus on patient and person-oriented care. Until the Second World War, medicine was mainly focused on the eradication of (acute) diseases. The definition of health as formulated by the World Health Organisation (WHO) in 1946, emphasized that health was described not only by the absence of disease [...], so contributing to the shift towards "patient-centered care" that today increasingly forms the basis of care in many medical disciplines." (De Maeseneer et al. 2012, p. 602)

[14] Note, though, that London (2019) is critical of this view of experts in medicine. For further discussion of these issues, see section 5.

[15] So, to be clear, we are not thinking of black-box medicine as a practice in which practitioners are removed from medical decision-making nor as one in which practitioners are bound to follow the recommendations of black-box systems. Rather, we think of black-box AI systems as decision *aids* that can serve practitioners in both diagnostic and treatment contexts.

While it can be hard to define precisely this "patient-centered" paradigm in healthcare, we can distill two core ideals that can help us characterize it.[16]

First, healthcare should treat patients as *people* whose values, beliefs, and psychosocial context all play important roles in establishing and maintaining their health. Patients are more than their somatic symptoms, and while a patient might die of, say, lung cancer, the more fundamental reason for his death might better be explained by citing the non-somatic reasons for his excessive smoking and drinking behaviors: reasons such as depression, unemployment, low level of education, and the like (McGinnis and Foege 1993).

Second, healthcare should treat patients as equal partners in medical decision-making: their wants should be heard, their wishes respected, and their knowledge considered. Viewed through this lens, *communication*, *explanation*, and *exchange of information* become the central elements in the practitioner-patient relationship. Importantly, for proponents of patient-centered medicine, the exchange of information between practitioner and patient "means going well beyond providing just facts and figures. Using a patient-centered approach, the clinician frames and tailors information in response to an understanding of a patient's concerns, beliefs, and expectations" (Epstein et al. 2010, p. 1491). That is, practitioners and patients should meet on a level playing field—in a state of "shared mind"—where both parties explicitly understand what the symptoms are and how they relate to various matters of diagnostic and treatment relevance. Based on this stock of shared information, the patient can, on informed grounds, reach a decision about how to continue the healthcare process.[17] In doing so, the central thought is that patients will have access to enough information to reach medical decisions in "autonomous and rational ways" (Hall et al. 2012, p. 535).

While this characterization is undoubtedly general and quite vague, it is clear enough to highlight the central differences between patient-centered medicine and more traditional "disease-based" conceptions of medicine. Patients are no longer seen as passive recipients of orders from the paternalistic and all-knowing doctor but rather understood as active partners who—in an informed manner—can and should exercise their own autonomy to help choose their own ends of action. In patient-centered medicine, the practitioner is more akin to a guide than an authority, and it is recognized that the ultimate goal of treatment is not necessarily identified with the eradication of a disease but rather with the creation of an informed decision space in which a patient's specific values and preferences are represented and respected.

Patient-centered medicine can be motivated empirically. There are studies that show that patient-centered medicine has resulted in better health outcomes, higher patient satisfaction, fewer rehospitalizations, and reduced medical costs.[18] Ethically, as well, patient-centered medicine has advantages:

[16] Note: the description of patient-centered medicine below captures the *ideals* of patient-centered medicine and not necessarily the way it is practiced in real healthcare.

[17] Throughout, for the sake of argument, we will assume that patients place central value on the ideal of shared decision-making. But we should acknowledge that real healthcare situations are rarely so straightforward. As pointed out to us by an anonymous referee, it has been documented by Vogel et al. (2008) that patients vary in the extent to which they genuinely want to be involved in the decision-making processes surrounding cancer.

[18] For an overview over these findings, see Delaney (2018).

> From the perspective of medical ethics, patient-centered care fulfills health care professionals' obligation to place the interests of the patient above all else and to respect patients' personal autonomy. Autonomy is often enhanced by caring partnerships between physicians and patients that support patients' ability and willingness to consider all reasonable options and to participate in their own care. (Epstein et al. 2010, p. 1491)

Insofar as one of the central demands of medical ethics is to give maximal attention to the interests and autonomy of patients, the very core of patient-centered medicine seems ethically promising.

4 Why Black-Box Medicine Conflicts with Patient-Centered Medicine

While "[u]ltimately, patient-centered interactions strive to achieve a state of shared information, shared deliberation, and shared mind", black-box medicine, we want to argue now, ultimately conflicts with these ideals (Epstein et al. 2010, p. 1491).

Consider the first core tenet of patient-centered medicine: practitioners should strive to meet patients as unique, autonomous persons with specific values, preferences, and life goals. On the face of it, we might think, black-box medicine makes it impossible to uphold this view of patients. Black-box AI systems operate on data, and we might worry that the complex set of factors that constitute a person—in the sense relevant to patient-centered medicine—cannot be reduced to a bundle of useful data points that we can feed to an algorithm. Had we considered (future) medical practices in which black-box systems *replace* human practitioners, and in which patients interact directly with black-box systems, there would indeed be deep issues to consider about how to sensibly encode preferences and values into artificial systems.[19] But for our purposes, we can largely sidestep such complications. As mentioned, we think of black-box systems as *decision aids* that practitioners can utilize when interacting with patients. As such, the whole discussion concerning preferences and values can be conducted between practitioners and patients. Consider, for instance, a black-box AI system that gives medical recommendations based on a primary goal of eradicating a particular disease. If the system gives a treatment recommendation that conflicts with central values of the patient, practitioner and patient can decide to look for alternative treatments that square better with those values. Clearly, opting for an alternative treatment in such a case might inflict on the practitioner's epistemic obligation to rely on the expert system, but we might think that this obligation, as mentioned, can be overridden by other factors. So a central reliance on black-box systems in medical decision-making does not in any obvious way conflict with the idea of viewing patients as persons.

When we focus on the second core tenet of patient-centered medicine, however, things begin to look much more problematic. In striving to establish a state of shared mind between patient and practitioner, *exchange of information*, *explanation*, and *understanding* become key components. In standard patient-centered interactions, the practitioner is in charge of facilitating this exchange of information—about how, say,

[19] For discussions of some of these issues, see Di Nucci (2019), Floridi (2011), and McDougall (2019).

the presence of specific symptoms relate to a specific diagnosis or to specific treatment options. But when we include black-box AI systems into the mix, it becomes potentially impossible to uphold this practice.

The core reason is simple: since black-box AI systems do not reveal to practitioners how or why they reach the recommendations that they do, then neither can practitioners who rely on these black-box systems in decision-making—assuming that they honor their epistemic obligation—explain to patients how and why they give the recommendations that they do. Yet, for patients to make decisions in autonomous and rational ways, it is a requirement that they have "the capacity to make sense of the [medical] information presented and can process it rationally to reach a decision that furthers their health care goals" (Bernat and Peterson 2006, p. 88). But when the practitioner cannot make sense of the relevant medical information buried in the deep learning network, then neither can he present the information in a way that enables a patient to comprehend and process it rationally. Borrowing a term from Mittelstadt et al. (2016), the evidence from black-box systems is "inscrutable" both for practitioners and patients who seek to make use of the "expert" system in decision-making. In this way, black-box decision making

> contrasts with traditional decision-making, where human decision-makers can in principle articulate their rationale when queried, limited only by their desire and capacity to give an explanation, and the questioner's capacity to understand it. (Mittelstadt et al. 2016, p. 7)

Hence the very idea of creating a context in which informed or rational medical decision-making is possible is jeopardized in black-box medicine.

To make our argument a bit more vivid, consider first a case in which a deep learning network is trained to deliver (probabilistic) risk estimates of developing a particular type of highly aggressive breast cancer.[20] Assume this network has access to electronic data from millions of patients, including both biomedical and demographic data, and access to millions of relevant research articles. Assume that the data has been properly collected, selected, and represented for the network and that the network has extracted its own data features involving hundreds of different patient variables. In terms of performance, assume that the network has demonstrated a very high degree of diagnostic accuracy—an accuracy that significantly surpasses what any human expert achieves with respect to the type of breast cancer—and that it has come up with surprising correlations linking various clusters of patient variables and symptoms to various risk estimates of developing breast cancer. Suppose these correlations are surprising to the medical community in the sense that there exists no explanation of why the clusters of patient variables and symptoms correlate with the specific risk estimates. In fact, let us suppose that the sheer amount of data and the large number of patient variables in the network make it close to practically impossible for the scientific community to even speculate about an underlying causal or biological explanation of

[20] While our example is purely hypothetical, there are several studies that investigate the prospects of using machine learning systems for breast cancer stratification and treatment options (see, for instance, Ferroni et al. (2019) and Xiao et al. (2018)).

the network's behavior, and to validate the network's predictions in a controlled experimental setting.[21] Enters the patient Mary. In addition to all the electronic data that is already available about Mary, the practitioner examines and asks her a number of questions about her current condition—we can leave it vague exactly what those questions are but imagine that they include a range of physiological and psychosocial elements. The practitioner feeds the additional information to the network, which gives Mary a particular risk of developing the type of breast cancer. For concreteness, suppose the calculated risk is high: 92% likelihood. Honoring his epistemic obligation to rely on experts in medical decision-making, the practitioner aligns his medical verdict with that of the network and reports the sad news to Mary.

In this case, it is hard to see how the practitioner can honor his epistemic obligation while simultaneously living up to the ideals of patient-centered medicine. Given the myriad of evidence, weights, and features that influence the network's estimate, the practitioner cannot reconstruct the reasoning behind the network's estimate nor can he interestingly question it. So aside from citing the network's superior reliability and accuracy, the practitioner cannot explain and justify to Mary why the network yields the prediction that it does.[22] When these epistemic deficits are combined, it becomes clear that the practitioner is unable to create a context in which *informed* deliberation and decision-making are possible. For instance, the practitioner will not be able to answer some very natural "why" questions such as "why do *I* have such a big risk of developing breast cancer?" that we may imagine Mary would have. So when the black-box system is in the diagnostic driving seat, the central goal of promoting informed decision-making through a state of shared information and deliberation appears unattainable.

While the case above illustrates how black-box medicine can conflict with patient-centered ideals in a diagnostic context, we can imagine similar cases in a treatment context. Consider a deep learning system like the above one, but suppose this time that the network also delivers a number of treatment options, ranked by likelihood of success, for the type of breast cancer in question. For concreteness, suppose the network yields a high likelihood for treatment X that involves surgical removal of both breasts. We can imagine that the network also outputs a range of side effects associated with the various treatment options and that the practitioner has access to all this information. Due to the complexity of the associated AI system, we can, as above, suppose that the doctor cannot reconstruct the reasons for why treatment X came out as the treatment most likely to be successful; the type of breast cancer is not well-understood, and the fact that the network assigns high probability to X but only low probabilities to the chemotherapeutic treatment Y or the hormonal treatment Z is unknown to the practitioner. Enters Mary who is informed of the system's rankings by the doctor.

[21] Note: even if we assume that a post hoc interpretation of such a network was possible—for instance, through clustering "significant" units in the network to yield information that those specific 1237 patient variables with those specific interconnected range values result in that specific risk estimate—it is far from clear how such an interpretation could be of practical use to the scientific community. For similar critical pointers, see London (2019).

[22] To be sure, the practitioner may be able to give an *abstract* explanation of how the deep learning network operates. Plausibly, the sort of explanation that the practitioner can offer Mary will be of the following form: "Somehow, based on a complex analysis of vast amount of data about people who share a high number of characteristics with you, the system determines with high reliability and accuracy that you are in significant danger of developing breast cancer; yet, I do not understand how and why it does it." But, as indicated above, it is doubtful whether such an explanation is even minimally informative for normal people.

In this case, it is clear that the practitioner can tell Mary about the different treatment options, and it is clear that he can discuss the pros and cons of the associated side effects with her. What is not so clear is whether this—in a treatment context—suffices for meeting the ideals of patient-centered medicine. Since Mary is about to make a life-altering decision that may considerably clash with her preferences and values, she seems entitled to get a reasonably informative explanation of *why* treatment X is likely to yield significantly better results than treatments Y or Z. Yet, as above, the practitioner cannot provide such an explanation. But then again it is hard to see how *informed* decision-making and deliberation are possible. Suppose Mary has a strong preference for, say, treatment Z. With a view toward making an autonomous and rational decision that furthers her healthcare and life goals, understanding why treatment Z receives such a low probability compared to treatment X will matter greatly to her. Yet, since the practitioner does not have access to this information, he cannot share it with Mary. So despite being aware of the potential side effects of the various treatment options, Mary is arguably still lacking a crucial element of understanding that prevents her from making a decision in an informed and rational manner.

So it seems that black-box medicine can conflict with the ideals of patient-centered medicine in both diagnostic and treatment contexts. The cases above are of course "extreme" in the sense that the medical community cannot even engage in qualified guesswork about the operations of the relevant deep learning networks. It goes without saying, though, that many potentially applicable AI systems in medicine will not have such radically opaque natures. Yet, for the purpose of making our central points, we can appeal to the extreme cases to get a clear feeling of the tension between black-box medicine and patient-centered medicine. More generally, we can hold that the more medicine turns into black-box medicine, the less we can retain of the ideals in patient-centered medicine.

5 Opacity in Medical Decision-Making

So far we have motivated the difficulties involved in squaring the central ideals of patient-centered medicine with the practices of black-box medicine. Yet, while one might agree with our argument, one might worry that the black-box problems we have described are not fundamentally different from currently existing problems of opacity in medical decision-making. For

> "[a]lthough medicine is one of the oldest productive sciences, its knowledge of underlying causal systems is in its infancy; the pathophysiology of disease is often uncertain, and the mechanisms through which interventions work is either not known or not well understood. As a result, decisions that are atheoretic, associationist, and opaque are commonplace in medicine. [...]
>
> As counterintuitive and unappealing as it may be, the opacity, independence from an explicit domain model, and lack of causal insight associated with some of the

> most powerful machine learning approaches are not radically different from routine aspects of medical decision-making. Our causal knowledge is often fragmentary, and uncertainty is the rule rather than the exception." (London 2019, pp. 17–18)

Consider a case in which a practitioner uses a new method of blood testing to diagnose a patient with a particular disease. While the practitioner knows that the new method is more reliable than older ones, he does not really understand how the blood testing works nor why it is more reliable. Yet, clearly, the practitioner is under an epistemic obligation to rely on the blood testing in his medical decision-making. So why, goes the worry, is this familiar kind of opacity in medical decision-making any different from the case of black-box medicine?[23]

It is clear that individual practitioners may often be even highly uncertain about how specific technological aids really function. But it remains true that there often are humans in the loop who *do* possess the relevant knowledge. These might be medical specialists, researchers, or other types of experts who could, if needed, provide some useful explanation of how specific technological aids such as blood testing methods work. Whether these explanations appeal to software, mechanistic, or causal considerations is not particularly important. Rather, what matters is that uncertainty at the level of individual practitioners typically does not translate into medical uncertainty *simpliciter*.

On the face of it, this contrasts with the use of black-box systems in medical decision-making. When it comes to such systems, as we have argued, the uncertainty is of a principled kind. That is, when we deal with black-box AI systems, there exists no expert who can provide practitioners with useful causal or mechanistic explanations of the systems' internal decision procedures. So in black-box medicine, uncertainty at the level of individual practitioners typically does translate into medical uncertainty *simpliciter*.

But one might deny that these differences are substantive. For might there not be cases in ordinary, non-AI-based medicine where there exists no expert explanation of how and why a given medical procedure works? Consider a practitioner who prescribes a drug to a patient, and suppose that there exists no causal or mechanistic explanation of why the drug works for the relevant symptoms. Insofar as the drug outperforms other available methods, the practitioner is under an obligation to prescribe the drug. Yet, one might suggest, since there are no humans in the decision-theoretic loop who can explain how and why the drug works, there need not be any substantive differences between opacity in black-box medicine and ordinary medicine. Roughly put, decision-making in both types of medicine can result from purely correlational evidence.

But even in this case, it seems, there are important differences. For although there exists no causal or mechanistic explanation of how and why a particular prescribed drug works, there is likely still *some* useful information that a practitioner can pass on to patients to encourage rational and informed deliberation.[24] In typical cases, that is, the practitioner will at least have access to various basic facts about the drug in question: facts about general characteristics of the test participants (age, sex, ethnicity), about the experiments

[23] Thanks to an anonymous reviewer for raising this issue. For sentiments similar to those aired by London (2019) in the quotes, see also Schönberger (2019) and Zerilli et al. (2018). For arguments that indicate that typical medical explanations might not be as opaque as the quotes above suggest, see Lipton (2017).

[24] As reported by Walker et al. (2019), such basic information is present even in medical studies that rely purely on correlational evidence.

and analyses that validated the effect of the drug, and about various statistical measures that inform the effect result. Based on such information, it seems plausible that the practitioner can answer—or at least give qualified guesses about answers to—some of the "why" questions that patients might have.

When it comes to black-box medicine, however, no such basic information needs to be available. As illustrated above, if the number of patient variables that help determine a particular probabilistic output reaches into the hundreds or thousands, there need not be any informative cluster of patient characteristics or traits that a practitioner can utilize to address pertinent "why" questions. Indeed, even if a high proportion of the relevant patient variables could be extracted in some fashion, there is no guarantee that they would make any sense to us. A deep learning network, for instance, might have extracted some complicated environmental characteristics—involving a mixed bag of features such as, say, educational level, distance to work, coffee consumption, hair color, and holiday trips to Italy—that are intuitively incomprehensible to us. So even in cases where practitioners rely on non-AI decision aids whose epistemic superiority cannot be explained nor justified by mechanistic or causal considerations, they do not (typically) inherit the radical opacity associated with black-box systems. That is, in the case of decisions based on black-box systems, we can literally fail to have a minimally sensible basic interpretation and explanation of the information that the algorithm employs for producing its recommendations. Notwithstanding the absence of a relevant causal or mechanistic explanation, this is not the case for ordinary medical decision-making.

We can also point out some important differences between ordinary opacity and AI opacity in medical decision-making by focusing on the role of *trust* in patient care. It is well-known that "the provision of explanations can affect levels of trust and acceptance of algorithmic decisions" (Binns et al. 2018, p. 377). Intuitively, the guiding thought is, if we can explain the reasons behind a certain AI prediction, then it is more likely that people will *trust* the AI system and act on the prediction. Witness Ribeiro et al. (2016):

> "Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one." (Ribeiro et al. 2016, abstract)

Insofar as explanatory transparency correlates with trust, we might then reasonably think that recommendations based on black-box systems fare poorly in promoting trust and hence in positively affecting patients' decision-making.[25] More specifically, if explanatory transparency correlates with trust, and if—as argued above—ordinary medical decision-making does not inherit the radical opacity associated with black-

[25] The lack of transparency in black-box systems will not only present a challenge for eliciting trust in patients but also for avoiding automation biases in practitioners. Roughly put, the less practitioners have access to the internal operations of black-box systems, the less they are in a position to determine whether their trust in a given AI output is medically justified or whether it stems from an automation bias: a propensity to over-rely on outputs from automated decision-making systems. For more on the issues surrounding AI and automation biases, see Challen et al. (2019) and Goddard et al. (2011).

box decision-making, then it will, everything being equal, be easier for a practitioner to promote a patient's trust in medical recommendations when those are not based on black-box systems.

In the era of black-box medicine, individual practitioners can thus find themselves in an uncomfortable decision-theoretic position. Suppose a practitioner wants to honor his epistemic obligation and align his medical recommendation with that of a black-box system. While realizing that the black-box recommendation is the most likely to cure the patient's disease, the practitioner might worry that the patient will not trust the system and hence not act on its recommendation. As a result, the practitioner might contemplate not basing his recommendation on the black-box system at all. In this case, the practitioner will be in an epistemic position where he knowingly feels compelled to recommend an evidentially suboptimal solution to increase trust, or where he knowingly feels compelled to compromise on trust to increase accuracy and reliability. Rarely, it seems, does ordinary opaque medical decision-making cause such conflicts in practice.

Moreover, black-box medicine has the potential to affect the practitioner-patient relationship in interesting new ways. Consider a case in which a practitioner suspects that her patient holds certain negative views about a treatment that otherwise would be very useful for him. While respecting the patient's values and preferences, suppose it becomes clear for the practitioner during conversations with the patient that the negative views are based on a gross misunderstanding of the treatment. In this case, the practitioner can try to inform the patient to help him realize that he ought to accept the treatment in question. In typical cases, the practitioner can tell the patient why she recommends the treatment by reconstructing her scientific reasons for doing so. Clearly, though, the patient will only be swayed by these reasons if he trusts the practitioner. In cases where the practitioner cannot reconstruct the reasons for her treatment recommendation—perhaps because there exists no causal or mechanistic explanation to guide her—this trust will be even more important. For in the latter case, perhaps the best the practitioner can do to persuade the patient to accept the treatment is to make an appeal to authority: "In my view as a medical expert, I strongly encourage you to accept the treatment." Acknowledging her uncertainty about the workings of the treatment, the patient's trust in the practitioner as a *person* becomes vital.

When it comes to recommendations based on black-box AI systems, however, practitioner and patient are in the same epistemic situation vis a vis the black-box AI systems: both have excellent reasons to trust the recommendations of these systems—because of their known superior knowledge, reliability, and accuracy—but neither can explain the grounds for its superior performance.[26] Plausibly, then, if a patient knows that a practitioner honors her epistemic obligation to align her medical verdict with that of the black-box system, the patient is disposed to view the black-box system rather

[26] In a loose sense, it might be instructive to think of black-box medicine as reintroducing a kind of *epistemic paternalism* into medical practice. Of course, the epistemic paternalism in question is special. It applies both to patients and practitioners and involves no deliberate withholding of information on the part of the practitioner. As argued, since black-box medicine does not suggest an approach to medicine where patient values and autonomy are ignored, we should not understand black-box medicine as promoting a return to an all-out paternalism in medicine. Yet, insofar as epistemic paternalism can be characterized, paraphrasing Goldman (1991), in terms of the withholding of information that it is in the subject's best interest to have—for instance, to enable informed decision-making—then black-box medicine does count as a sort of interesting, new type of epistemic paternalism.

than the practitioner as the expert. In the absence of available explanations of why a certain medical recommendation was made, the patient's trust in the black-box system becomes vital.

So, while traditional medical decision-making might struggle to explain how (potentially blind) trust in practitioners can be promoted, black-box medicine must struggle to explain how blind trust in black-box systems can be promoted. And it is by no means obvious that these two challenges are similar.[27] In contrast to traditional human decision-making, artificial decision-making must countenance the risk that black-box systems might occasionally be perceived as acting "foolishly."[28] Deep learning networks, for instance, can sometimes give excellent results on training data but give radically wrong results when applied to data "from the wild", and sometimes they may rest their predictions on what we would regard as spurious correlations that do not track anything real in the world. Insofar as such behavior plausibly counts as foolish and distinctively "nonhuman"—humans often make mistakes, no doubt, but they are rarely easily fooled into making radical changes in their predictions and estimates—decision-making based on black-box systems provokes challenges that are not faced by traditional (opaque) medical decision-making.

Obviously, there is a lot more to say about the putative differences between ordinary opaque decision-making and opaque black-box decision-making. But for now we trust that the arguments and examples above do enough to motivate the claim that black-box medicine constitutes an interestingly new epistemic challenge to ideals and practices in current healthcare.

6 Concluding Remarks

If practitioners honor their epistemic obligations to provide the most accurate and reliable treatments for their patients, we have argued that black-box medicine conflicts with core ideals of patient-centered medicine. In particular, as we saw, black-box medicine is not conducive for supporting informed decision-making based on shared information, shared deliberation, and shared mind between practitioner and patient.

Throughout we have focused on patient-centered medicine. But black-box medicine also plausibly casts doubt on central ideals of *evidence-based medicine*. Initially, it might seem as if black-box medicine could count as the ultimate realization of evidence-based medicine. In a nutshell, proponents of evidence-based medicine hold that decision-making should be grounded in the best available scientific and clinical evidence. Of course, practitioners have always sought to make decisions based on evidence, but often—or so the evidence-based medical critique goes—they relied on evidence that was distorted by unfounded medical dogmas, biased expert opinions, and

[27] Clearly, a lot hinges on how we spell out what it means to trust someone or something. If the relevant criteria for trust are criteria such as reliability and accuracy, then it may well turn out that people ought to put as much trust in black-box systems as in human experts. But if the relevant trust criteria include more fuzzy ones such as benevolence and honesty, then it may well turn out that people ought to trust human experts more than black-box systems; for instance, as pointed out by Ploug and Holm (2019), it may be that patients simply fear AI technology and, as a result, are disposed to distrust black-box systems more than human experts.

[28] For more on the ways in which AI systems can act "foolishly," see Nguyen et al. (2015), Ribeiro et al. (2016), and Su et al. (2019).

idiosyncratic assessments of patients. According to evidence-based medicine, the gold standard of evidence is the kind of evidence that results from randomized controlled clinical trials and various meta-analyses of medical data. By basing medical decision-making on these sources of evidence, the hope is that we can minimize the influences of personal biases and opinions. Insofar as black-box systems can make medical recommendations that take *all* this evidence into consideration—a feat that might not be possible for individual practitioners—it is thus tempting to hold that black-box medicine can count as the ideal manifestation of evidence-based medicine. If this was true, and if it is true, as we have argued, that black-box medicine conflicts with patient-centered medicine, then we would also have an argument showing that evidence-based medicine conflicts with patient-centered medicine.

While such a conclusion would be significant, it is not very plausible on more well-articulated conceptions of evidence-based medicine. It is true that evidence-based medicine—at least in theory—typically ranks evidence that does *not* stem from randomized controlled trials and meta-analyses *lower* in the hierarchy of tools that aid medical decision-making (Seshia and Young 2013). But that does not mean that evidence from interactions with patients does not figure in the hierarchy. According to Straus et al. (2019), "evidence-based medicine requires the integration of the best research evidence with our clinical expertise and our patient's unique values", where they by "patient values" mean "the unique preferences, concerns, and expectations that each patient brings to a clinical encounter and that must be integrated into shared clinical decisions if they are to serve the patient" (Straus et al. 2019, p. 18). Depending on how much weight we put on the idea of shared clinical decision-making, it is clear that the central conflict between black-box medicine and patient-centered medicine can translate into a conflict between black-box medicine and evidence-based medicine. Insofar as practitioners cannot reconstruct the reasons behind a black-box AI recommendation, they are not able to create a context in which informed shared deliberation and decision-making are possible. So proponents of evidence-based medicine should also be aware of the opacity problems that follow in the wake of black-box medicine.

Throughout the paper we have been rather one-dimensional in our understanding of black-box medicine as medical practices in which black-box systems play essential roles in decision making. Yet, this description hides a number of potentially complicating factors that might affect the scope of our conclusions.[29] For instance, do problems of opacity arise with the same force in all types of healthcare? In the context of patient-centered medicine, for instance, one might think that problems of opacity are most acute in primary and secondary healthcare, where the bulk of a patient's informed decision-making arguably takes place but less acute in tertiary and quaternary healthcare. Likewise, we have not considered cases in which black-box systems act as *autonomous* systems that deliver medical recommendations directly to patients without any human involvement. While we might, on the face of it, expect that opacity problems will increase in strength when the degree of human involvement decreases, the finer details might well depend on specific AI applications—there are important differences between, say, autonomous symptoms checkers and autonomous mental chatbots. While all these complicating factors are relevant to our overall argument, we must postpone a detailed discussion to future work.

[29] Thanks to an anonymous referee for encouraging us to acknowledge these complicating factors.

While our conclusions pertain to black-box medicine, note that there is no obvious reason to suspect that AI-*informed* decision-making, more generally, should be in conflict with core ideals of patient-centered medicine. If practitioners, for instance, only utilize transparent AI systems for medical decision-making—according to some relevant notion of transparency—there are no immediate problems related to opacity. Likewise, as mentioned, we have not ruled out that progress in explainable AI may one day light up the black-box in ways that make it clear to human decision-makers what explains various deep learning predictions. We have made our arguments on the assumption that black-box AI systems exist, but if there are ways to solve the apparent tension between "the best-performing methods (e.g., deep learning) [being] the least transparent, and the ones providing a clear explanation (e.g., decision trees) [being] less accurate", our conclusions should of course be adjusted (Holzinger et al. 2017, p. 2). Despite this, however, the high practical stakes that are involved in a widespread implementation of black-box systems in medicine make it timely to discuss these matters already now. For black-box medicine not only has ramifications for practitioners and patients but also for policy makers in charge of allocating funds and determining responsibility in the healthcare system.[30]

References

Bernat, J. L., & Peterson, L. M. (2006). Patient-centered informed consent in surgical practice. *Archives of Surgery, 141*(1), 86–92.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N. (2018). 'It's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 377). ACM.

Burrell, J. (2016). How the machine "thinks": understanding opacity in machine learning algorithms. *Big Data and Society, 3*(1), 1–12.

Captain, S. (2017). Can IBM's Watson do it all. *Fast Company.* Retrieved from https://www.fastcompany.com/3065339/can-ibms-watson-do-it-all (accessed online 29/10/2019).

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Qual Saf, 28*(3), 231–237.

Calo, R. (2015). Robotics and the lessons of cyberlaw. *California Law Review, 103*(3), 513–563.

Danaher, J. (2016). Robots, law and the retribution-gap. *Ethics and Information Technology, 18*(4), 299–309.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine, 24*(9), 1342–1350.

Delaney, L. J. (2018). Patient-centred care as an approach to improving health care in Australia. *Collegian, 25*(1), 119–123.

De Maeseneer, J., van Weel, C., Daeren, L., Leyns, C., Decat, P., Boeckxstaens, P., Avonts, D., & Willems, S. (2012). From "patient" to "person" to "people": the need for integrated, people-centered healthcare. *The International Journal of Person Centered Medicine, 2*(3), 601–614.

Di Nucci, N. (2019). Should we be afraid of medical AI? *Journal of Medical Ethics, 45*(8), 556–558.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

Epstein, R. M., Fiscella, K., Lesser, C. S., & Stange, K. C. (2010). Why the nation needs a policy push on patient-centered health care. *Health affairs, 29*(8), 1489–1495.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine, 25*(1), 24–29.

[30] For some of the issues concerning responsibility in AI-informed decision making, see Calo (2015), Floridi et al. (2018), Danaher (2016), Nyholm (2018), and Price II (2017).

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118.

Faes, L., Liu, X., Kale, A., Bruynseels, A., Shamdas, M., Moraes, G., Fu, D.J., Wagner, S.K., Kern, C., Ledsam, J.R. and Schmid, M.K. (2019). Deep learning under scrutiny: performance against health care professionals in detecting diseases from medical imaging-systematic review and meta-Analysis (preprint).

Ferroni, P., Zanzotto, F., Riondino, S., Scarpato, N., Guadagni, F., & Roselli, M. (2019). Breast cancer prognosis using a machine learning approach. *Cancers, 11*(3), 328.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707.

Floridi, L. (2011). The informational nature of personal identity. *Minds & Machines, 21*, 549–566.

Forssbæck, J., & Oxelheim, L. (2014). The multifaceted concept of transparency. In J. Forssbæck & L. Oxelheim (Eds.), *The Oxford handbook of economic and institutional transparency* (pp. 3–31). New York: Oxford University Press.

Goddard, K., Roudsari, A., & Wyatt, J. C. (2011). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127.

Goldman, A. (1991). Epistemic paternalism: communication control in law and society. *Journal of Philosophy, 88*(3), 113–131.

Hall, D. E., Prochazka, A. V., & Fink, A. S. (2012). Informed consent for clinical treatment. *CMAJ, 184*(5), 533–540.

Harvey, H., & Glocker, B. (2019). A standardized approach for preparing imaging data for machine learning tasks in radiology. Artificial Intelligence in Medical Imaging (pp. 61–72). Springer, Cham.

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine, 25*(1), 30–36.

Heald, D. (2006). Transparency as an instrumental value. In C. Hood & D. Heald (Eds.), *Transparency: the key to better governance?* (pp. 59–73). Oxford: Oxford University Press.

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*.

Holzinger, A., Haibe-Kains, B., & Jurisica, I. (2019). Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging*. https://doi.org/10.1007/s00259-019-04382-9.

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge: Cambridge University Press.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol, 2*(4), 230–243.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science, 349*(6245), 255–260.

Kallis B., Collier M., Fu R. (2018). 10 promising AI applications in health care. *Harvard Business Review*, https://hbr.org/2018/05/10-promising-ai-applications-in-health-care (accessed online 11/12/2018).

Lee, J. G., Jun, S., Cho, Y. W., Lee, H., Kim, G. B., Seo, J. B., & Kim, N. (2017). Deep learning in medical imaging: general overview. *Korean Journal of Radiology, 18*(4), 570–584.

Lipton, P. (2003). *Inference to the best explanation*. Abingdon: Routledge.

Lipton, Z. C. (2017). The doctor just won't accept that!. *arXiv preprint arXiv:1711.08037*.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health, 1*(6), e271–e297.

Loh, E. (2018). Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health. *BMJ Leader, 2*, 59–63.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report, 49*(1), 15–21.

Marcum, J. A. (2008). *An introductory philosophy of medicine: Humanizing modern medicine* (Vol. 99). Springer Science & Business Media.

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics, 45*(8), 156–160.

McGinnis, J. M., & Foege, W. H. (1993). Actual causes of death in the United States. *JAMA, 270*(18), 2207–2212.

Miller, T. (2018). Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence*, https://arxiv.org/pdf/1706.07269.pdf (accessed online 11/12/2018).

Mittelstadt B. D., Allo P., Taddeo M., Wachter S., Floridi L. (2016). The ethics of algorithms: mapping the debate. *Big Data & Society*, pp. 1–21.

Montgomery, K. (2006). *How doctors think: Clinical judgment and the practice of medicine*. Oxford: Oxford University Press.
Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 427–436).
Nyholm, S. (2018). Attributing agency to automated systems: reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics, 24*(4), 1201–1219.
Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine, 375*(13), 1216–1219.
Olorisade, B. K., Brereton, P., & Andras, P. (2017). Reproducibility in machine learning-based studies: an example of text mining.
Ploug, T., & Holm, S. (2019). The right to refuse diagnostics and treatment planning by artificial intelligence. *Medicine, Health Care, and Philosophy.* https://doi.org/10.1007/s11019-019-09912-8.
Prat, A. (2006). The more closely we are watched, the better we behave? In C. Hood & D. Heald (Eds.), *Transparency: the key to better governance?* (pp. 91–103). Oxford: Oxford University Press.
Price II, W. N. (2017). Artificial intelligence in healthcare: applications and legal implications. *The SciTech Lawyer, 14*(1), 10–13.
Price II, W. N. (2018). Medical malpractice and black-box medicine. In I. Cohen, H. Lynch, E. Vayena, & U. Gasser (Eds.), *Big Data, Health Law, and Bioethics* (pp. 295–306). Cambridge: Cambridge University Press.
Purdy, M., & Daugherty, P. (2016). Why artificial intelligence is the future of growth. *Remarks at AI Now: The Social and Economic Implications of Artificial Intelligence Technologies in the Near Term*, 1–72.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM.
Schönberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology, 27*(2), 171–203.
Schubbach, A. (2019). Judging machines: philosophical aspects of deep learning. *Synthese*, pp. 1–21.
Seshia, S. S., & Young, G. B. (2013). The evidence-based medicine paradigm: where are we 20 years later? Part 1. *Canadian Journal of Neurological Sciences, 40*(4), 465–474.
Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1–74.
Straus, S., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2019). *Evidence-based medicine: How to practice and teach EBM* (5rd ed.). Edinburgh; New York: Elsevier.
Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*.
Tiwari, P., Prasanna, P., Wolansky, L., Pinho, M., Cohen, M., Nayate, A. P., et al. (2016). Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric MRI: a feasibility study. *American Journal of Neuroradiology, 37*(12), 2231–2236.
Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine, 25*(1), 44.
US Food and Drug Administration. (2018). FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. *News Release, April* (retrieved online Accessed August 7, 2018).
van Ooijen, P. M. (2019). Quality and curation of medical images and data. In *Artificial Intelligence in Medical Imaging* (pp. 247–255). Cham: Springer.
Vogel, B. A., Helmes, A. W., & Hasenburg, A. (2008). Concordance between patients' desired and actual decision-making roles in breast cancer care. *Psycho-Oncology, 17*(2), 182–189.
Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics, 2*(6).
Walker, M.J., Bourke, J. and Hutchison, K. (2019). Evidence for personalised medicine: mechanisms, correlation, and new kinds of black box. *Theoretical medicine and bioethics, 40*(2), pp. 103–121.
Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: beyond the black box. *BMJ, 364*, l886.
Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine, 153*, 1–9.
Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology, pp.*, 1–23.