Cambridge Quarterly of Healthcare Ethics (2026), 35: 3, 167–176
doi:10.1017/S0963180122000305

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

The Virtues of Interpretable Medical Artificial Intelligence

Joshua Hatherley* , Robert Sparrow and Mark Howard

School of Philosophical, Historical, and International Studies, Monash University, Clayton, Victoria 3168, Australia
*Corresponding author. Email: joshua.hatherley@monash.edu

Abstract

Artificial intelligence (AI) systems have demonstrated impressive performance across a variety of clinical tasks. However, notoriously, sometimes these systems are "black boxes." The initial response in the literature was a demand for "explainable AI." However, recently, several authors have suggested that making AI more explainable or "interpretable" is likely to be at the cost of the accuracy of these systems and that prioritizing interpretability in medical AI may constitute a "lethal prejudice." In this article, we defend the value of interpretability in the context of the use of AI in medicine. Clinicians may prefer interpretable systems over more accurate black boxes, which in turn is sufficient to give designers of AI reason to prefer more interpretable systems in order to ensure that AI is adopted and its benefits realized. Moreover, clinicians may be justified in this preference. Achieving the downstream benefits from AI is critically dependent on how the outputs of these systems are interpreted by physicians and patients. A preference for the use of highly accurate black box AI systems, over less accurate but more interpretable systems, may itself constitute a form of lethal prejudice that may diminish the benefits of AI to—and perhaps even harm—patients.

Keywords: ethics; medicine; healthcare; artificial intelligence; black box; explainable AI; deep learning

Introduction

Deep-learning artificial intelligence (AI) systems have demonstrated impressive performance across a variety of clinical tasks, including diagnosis, risk prediction, triage, mortality prediction, and treatment planning.[1,2] A problem, however, is that the inner workings of these systems have often proven thoroughly resistant to understanding, explanation, or justification, not only to end-users (e.g., doctors, clinicians, and nurses) but also to the designers of these systems themselves. Such AI systems are commonly described as "opaque," "inscrutable," or "black boxes." The initial response to this problem in the literature was a demand for "explainable AI." However, recently, several authors have suggested that making AI more explainable or "interpretable" is likely to be achieved at the cost of the accuracy of these systems and that a preference for explainable systems over more accurate AI is ethically indefensible in the context of medicine.[3,4]

In this article, we defend the value of interpretability in the context of the use of AI in medicine. We point out that clinicians may prefer interpretable systems over more accurate black boxes, which in turn is sufficient to give designers of AI reason to prefer more interpretable systems in order to ensure that AI is adopted and its benefits realized. Moreover, clinicians may themselves be justified in this preference. Medical AI should be analyzed as a sociotechnical system, the performance of which is as much a function of how people respond to AI as it is of the outputs of the AI. Securing the downstream therapeutic benefits from diagnostic and prognostic systems is critically dependent on how the outputs of these systems are interpreted by physicians and received by patients. Prioritizing accuracy over interpretability overlooks the various human factors that could interfere with downstream benefits to patients. We argue that, in some cases, a less accurate but more interpretable AI may have better effects

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press

on patient health outcomes than a "black box" model with superior accuracy, and suggest that a preference for the use of a highly accurate black box AI systems, over less accurate but more interpretable systems, may itself constitute a form of lethal prejudice that may diminish the benefits of AI to patients—and perhaps even harm them.

The Black Box Problem in Medical AI

Recent advances in artificial AI and machine learning (ML) have significant potential to improve the current practice of medicine through, for instance, enhancing physician judgment, reducing medical error, improving the accessibility of medical care, and improving patient health outcomes.[5,6,7] Many advanced ML AI systems have demonstrated impressive performance in a wide variety of clinical tasks, including diagnosis,[8] risk prediction,[9] mortality prediction,[10] and treatment planning.[11] In emergency medicine, medical AI systems are being investigated to assist in the performance of diagnostic tasks, outcome prediction, and clinical monitoring.[12] A problem, however, is that ML algorithms are notoriously opaque, "in the sense that if one is a recipient of the output of the algorithm […], rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs."[13] This can occur for a number of reasons, including a lack of relevant technical knowledge on the part of the user, corporate or government concealment of key elements of an AI system, or at the deepest level, a cognitive mismatch between the demands of human reasoning and the technical approaches to mathematical optimization in high-dimensionality that are characteristic of ML.[14] Joseph Wadden suggests that the black box problem "occurs whenever the reasons why an AI decision-maker has arrived at its decision are not currently understandable to the patient or those involved in the patient's care because the system itself is not understandable to either of these agents."[15]

A variety of related concerns have been raised over the prospect of black box clinical decision support systems being operationalized in clinical medicine. Some authors worry that human physicians may act on the outputs of black box medical AI without a clear understanding of the reasons behind them,[16] or that opacity may conceal erroneous inferences or algorithmic biases that could jeopardize patient health and safety.[17,18,19] Others are concerned that opacity could interfere with the allocation of moral responsibility or legal liability in the instance that patient harm results from accepting and acting upon the outputs of a black box medical AI system,[20,21,22] or that the use of black-box medical AI systems may undermine the accountability that healthcare practitioners accept for AI-related medical error.[23] Still, others are concerned that black box medical AI systems cannot, will not, and perhaps ought not to be trusted by doctors or patients.[24,25,26] These concerns are especially acute in the context of emergency medicine, where decisions need to be made quickly and coordinated across teams of multispecialist practitioners.

Responding to these concerns, some authors have argued that medical AI systems will need to be "interpretable," "explainable," or "transparent" in order to be responsibly utilized in safety-critical medical settings and overcome these various challenges.[27,28,29]

The Case for Accuracy

Recently, however, some authors have argued that opacity in medical AI is not nearly as problematic as critics have suggested, and that the prioritization of interpretable over black box medical AI systems may have several ethically unacceptable implications. Critics have advanced two distinct arguments against the prioritization of interpretability in medical AI.

First, some authors have highlighted parallels between the opacity of ML models and the opacity of a variety of commonplace medical interventions that are readily accepted by both doctors and patients. As Eric Topol has noted, "[w]e already accept black boxes in medicine. For example, electroconvulsive therapy is highly effective for severe depression, but we have no idea how it works. Likewise, there are many drugs that seem to work even though no one can explain how."[30] The drugs that Topol is referring to here include aspirin, which, as Alex John London notes, "modern clinicians prescribed […] as an

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press

analgesic for nearly a century without understanding the mechanism through which it works," along with lithium, which "has been used as a mood stabilizer for half a century, yet why it works remains uncertain."[31] Other authors have also highlighted acetaminophen and penicillin, which "were in widespread use for decades before their mechanism of action was understood," along with selective serotonin reuptake inhibitors, whose underlying causal mechanism is still unclear.[32] Still, others have highlighted that the opacity of black-box AI systems is largely identical to the opacity of other human minds, and in some respects even one's own mind. For instance, John Zerilli and coauthors observe that "human agents are […] frequently *mistaken* about their real (internal) motivations and processing logic, a fact that is often obscured by the ability of human decision-makers to invent post hoc rationalizations."[33] According to some, these similarities imply that clinicians ought not to be anymore concerned about opacity in AI than they are about the opacity of their colleagues' recommendations, or indeed the opacity of their own internal reasoning processes.[34]

This first argument is a powerful line of criticism of accounts that hold that we should entirely abjure the use of opaque AI systems. However, it leaves open the possibility that, as we shall argue below, interpretable systems have distinct advantages that justify our preferring them.

The second argument assumes—as does much of the AI and ML literature—that there is an inherent trade-off between accuracy and interpretability (or explainability) in AI systems. In their 2016 announcement of the "Explainable AI (XAI)" project, for instance, the United States Defence Advanced Research Projects Agency claims that "[t]here is an inherent tension between ML performance (predictive accuracy) and explainability; often the highest performing methods (e.g., deep learning) are the least explainable, and the most explainable (e.g., decision trees) are less accurate."[35] Indeed, attempts to enhance our understanding of AI systems through the pursuit of intrinsic or ex ante interpretability (for instance, by restricting the size of the model, implementing "interpretability constraints," or using simpler, rule-based classifiers over more complex deep neural networks) are often observed to result in compromises to the accuracy of a model.[36,37,38] In particular, the development of a high-performing AI system entails an unavoidable degree of complexity that often interferes with how intuitive and understandable the operations of these systems are in practice.[39]

Consequently, some authors suggest that prioritizing interpretability over accuracy in medical AI has the ethically troubling consequence of compromising the accuracy of these systems, and subsequently, the downstream benefits of these systems for patient health outcomes.[40,41] Alex London has suggested that "[a]ny preference for less accurate models—whether computational systems or human decision-makers—carries risks to patient health and welfare. Without concrete assurance that these risks are offset by the expectation of additional benefits to patients, a blanket preference for simpler models is simply a lethal prejudice."[42] According to London, when we are patients, it is more important to us that something works than that our physician knows precisely how or why it works.[43] Indeed, this claim appears to have been corroborated by a recent citizen jury study, which found that participants were less likely to value interpretability over accuracy in healthcare settings compared to non-healthcare settings.[44] London thus concludes that the trade-off between accuracy and interpretability in medical AI ought therefore to be resolved in favor of accuracy.

The Limits of Post Hoc Explanation

One popular response to these concerns is to hope that improvements in post hoc explanation methods could enhance the interpretability of medical AI systems without compromising their accuracy.[45] Rather than pursuing ex ante or intrinsic interpretability, post hoc explanation methods attempt to extract explanations of various sorts from black-box medical AI systems on the basis of their previous decision records.[46,47,48] In many cases, this can be achieved without altering the original, black-box model, either by affixing a secondary explanator to the original model or by replicating its statistical function and overall performance through interpretable methods.[49]

The range of post hoc explanation methods is expansive, and it is beyond the scope of this article to review them all here. However, some key examples of post hoc explanation methods include sensitivity

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press

analysis, prototype selection, and saliency masks.[50] *Sensitivity analysis* involves "evaluating the uncertainty in the outcome of a black box with respect to different sources of uncertainty in its inputs."[51] For instance, a model may return an output with a confidence interval of 0.3, indicating that it has produced this output with low confidence, with the aim of reducing the strength of a user's credence. *Prototype selection* involves returning, in conjunction with the output, an example case that is as similar as possible to the case that has been entered into the system, with the aim of illuminating some of the criteria according to which output was generated. For instance, suppose a medical AI system, such as IDx-DR,[52] was to diagnose a patient with diabetic retinopathy from an image of their retina. A prototype selection explanator might produce, in conjunction with the model's classification, a second example image that is most similar to the original case, in an attempt to illustrate important elements in determining its output. Lastly, *saliency masks* highlight certain words, phrases, or areas of the image that were most influential in determining a particular output.

Post hoc explanation methods *have* demonstrated some potential to minimize some of the concerns of the critics of opacity discussed in section "The Black Box Problem in Medical AI," while also sidestepping the objections of critics of interpretability discussed in section "The Case for Accuracy." However, post hoc explanation methods also suffer from a number of significant limitations, which preclude them from entirely resolving this debate.

First, the addition of post hoc explanation methods to "black box" ML systems adds another layer of uncertainty to the evaluation of their outputs and inner workings. Post hoc explanations can only offer an approximation of the computations of a black-box model, meaning that it may be unclear how the explanator works, how faithful it is to the model, and why its outputs or explanations ought to be accepted.[53,54,55]

Second, and relatedly, such explanations often only succeed in extracting information that is highly incomplete.[56] For example, consider an explanator that highlights the features of a computed breast tomography scan that were most influential in classifying the patient as high-risk. Even if the features highlighted were intuitively relevant, this "explanation" offers a physician little reason to accept the model's output, particularly if the physician disagrees with it.

Third, the aims of post hoc explanation methods are often under-specified, particularly once the problem of agent-relativity in explanations is considered. Explanations often need to be tailored to a particular audience in order to be of any use. As Carl Zednik has expressed, "although the opacity of ML-programmed computing systems is traditionally said to give rise to the Black Box Problem, it may in fact be more appropriate to speak of many Black Box Problems—one for every stakeholder."[57] An explanation that assumes a background in computer science, for instance, may be useful for the manufacturers and auditors of medical AI systems, but is likely to deliver next to no insight for a medical professional that lacks this technical background. Conversely, a simple explanation tailored to patients, who typically lack both medical and computer science backgrounds, is likely to provide little utility to a medical practitioner. Some post hoc explanations may prove largely redundant or useless, while others may influence the decisions of end-users in ways that could reduce the clinical utility of these systems.

Finally, the focus on explanation has led to the neglect of *justification* in explainable medical AI.[58,59] Explanations are descriptive, in that they give an account of why a reasoner arrived at a particular judgment, but justifications give a normative account of why that judgment is a *good* judgment. There is a significant overlap between explanations and justifications, but they are far from identical. Yet within the explainability literature in AI, explanations and justifications are rarely distinguished, and when they are, it is the former that is prioritized over the latter.[60]

Consequently, despite high hopes that explainability could overcome the challenges of opacity and the accuracy-interpretability trade-off in medical AI, post hoc explanation methods are not currently capable of meeting this challenge.

Three Problems with the Prioritization of Accuracy over Interpretability

In this section, we highlight three problems underlying the case for accuracy, concerning (1) the clinical objectives of medical AI systems and the need for accuracy maximization; (2) the gap between technical

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press

accuracy in medical AI systems and their downstream effects upon patient health outcomes; and (3) the reality of the accuracy-interpretability trade-off. Both together and separately, these problems suggest that interpretability is more valuable than critics appreciate.

First, the accuracy of a medical AI system is not always the principal concern of human medical practitioners and may, in some cases, be secondary to the clinician's own ability to understand and interpret the outputs of the system, along with certain elements of the system's functioning, or even the system as a whole. Indeed, the priorities of clinicians are largely dependent upon their conception of the particular aims of any given medical AI system. In a recent qualitative study, for instance, Carrie Cai and coauthors found that the importance of accuracy to medical practitioners varies according to the practitioners' own conception of a medical AI system's clinical objectives.[61] "To some participants, the AI's objective was to be as accurate as possible, independent of its end-user. […] To others, however, the AI's role was to merely draw their attention to suspicious regions, given that the pathologist will be the one to make sense of those regions anyway: "It just gives you a big picture of this is the area it thinks is suspicious. You can just look at it and it doesn't have to be very accurate".[62] In these latter cases, understanding a model's reasons or justifications for drawing the clinician's attention to a particular area of an image may rank higher on the clinicians' list of priorities than the overall accuracy of the system, in order that they may reliably determine why a model has drawn the clinician's attention to a particular treatment option, piece of information, or area of a clinical image. This is not to deny the importance of accuracy in, say, a diagnostic AI system for the improvement of patient health outcomes, but rather to suggest that, in some cases, and for some users, the accuracy of an AI system may not be as critical as London and other critics have supposed, and may rank lower on the clinician's list of priorities than the interpretability of the system. Depending upon the specific performance disparities between black box and interpretable AI systems, there may be cases where clinicians prefer less accurate systems that they can understand over black-box systems with superior accuracy. If users prefer interpretable models over "black box" systems, then the potential downstream benefits of "black box" AI systems for patients could be undermined in practice if, for instance, clinicians reject them or avoid using them. Implementing "black box" systems over interpretable systems without respect for the preferences of the users of these systems may result in suboptimal outcomes that could have otherwise been avoided through the use of less accurate but more interpretable AI systems. Even if clinicians' preference for interpretability *is* a prejudice, if it is sufficiently widespread and influential, it may be sufficient to justify designers of AI to prioritize interpretability in order to increase the likelihood that AI systems will be adopted and their benefits realized.

Second, *contra* London, clinicians may themselves be justified in this preference. The case for accuracy appears to erroneously assume a necessary causal link between technical accuracy and improved downstream patient health outcomes. While diagnostic and predictive accuracies are certainly important for the improvement of patient health outcomes, they are far from sufficient. Medical AI systems need to be understood as intervening in care contexts that consist of an existing network of sociotechnical relations, rather than as mere technical "additions" to existing clinical decision-making procedures.[63,64,65] How these AI systems will become embedded into these contexts, and alter existing relations between actors, is crucially important to the extent to which they will produce downstream health benefits. As Sara Gerke and coauthors argue, the performance of medical ML systems will be influenced by a variety of broader human factors beyond the system itself, including the way that clinicians respond to the outputs of the systems, "the reimbursement decisions of insurers, the effects of court decisions on liability, any behavioral biases in the process, data quality of any third-party providers, any (possibly proprietary) ML algorithms developed by third parties, and many others."[66] Thus, as Thomas Grote observes in his recent discussion of clinical equipoise in randomized clinical trials of diagnostic medical AI systems, "even if the AI system were outperforming clinical experts in terms of diagnostic accuracy during the validation phase, its clinical benefit would still remain genuinely uncertain. *The main reason is that we cannot causally infer from an increase in diagnostic accuracy to an improvement of patient outcome*" (emphasis added).[67] There is a gap, in other words, between the accuracy of medical AI systems and their effectiveness in clinical practice, insofar as improvements in the accuracy of a technical system do not automatically translate into improvements in downstream health

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press

outcomes. Indeed, this observation is borne out by the current lack of evidence of downstream patient benefits generated from even the most technically accurate of medical AI systems. Superior accuracy is, in short, insufficient to demonstrate superior outcomes.

One reason for this gap comes from the fact that human users do not respond to the outputs of algorithmic systems in the same way that we respond to our own judgments and intuitions, nor even to the recommendations of other human beings. Indeed, according to one recent systematic review in human–computer interaction studies, "the inability to effectively combine human and nonhuman (i.e., algorithmic, statistical, and machine) decision making remains one of the most prominent and perplexing hurdles for the behavioral decision making community."[68] Prevalent human biases affect the interpretation of algorithmic recommendations, classifications, and predictions. As Sara Gerke and coauthors observe, "[h]uman judgement […] introduces well-known biases into an AI environment, including, for example, inability to reason with probabilities provided by AI systems, over extrapolation from small samples, identification of false patterns from noise, and undue risk aversion."[69] In safety-critical settings and high-stakes decision-making contexts such as medicine, these sorts of biases could pose significant risks to patient health and well-being.

Moreover, some of these biases are more likely to occur in cases where the medical AI system is opaque, rather than interpretable. Algorithmic aversion, for instance, is a phenomenon in which the users of an algorithmic system consistently reject the outputs of an algorithmic system, even when the user has observed the system performance to a high standard consistently over time, and when following the recommendations of the system would produce better outcomes overall.[70,71] Algorithmic aversion is most commonly observed in cases where the users of the system have expertise in the domain for which the system is designed (e.g., dermatologists in the diagnosis of malignant skin lesions);[72] in cases where the user has seen the system make (even minor) mistakes;[73] but most importantly for our purposes, *in cases where the algorithmic system is perceived to be opaque by its user*.[74] "Thus," claim Michael Yeomans and coauthors, "it is not enough for algorithms to be more accurate, they also need to be understood."[75]

Finally, in passing, it is worth noting that some authorities have begun to contest the reality of the accuracy-interpretability trade-off in AI and ML. In particular, Cynthia Rudin has recently argued that the accuracy-interpretability trade-off is a myth, and that simpler, more interpretable classifiers can perform to the same general standard as deep neural networks after preprocessing, particularly in cases where data are structured and contain naturally meaningful features, as is common in medicine.[76,77] Indeed, Rudin argues that interpretable AI can, in some cases, demonstrate *higher* accuracy than comparatively black box AI systems. "Generally, in the practice of data science," she claims, "the small difference in performance between ML algorithms can be overwhelmed by the ability to interpret results and process the data better at the next iteration. In those cases, the accuracy/interpretability trade-off is reversed—more interpretability leads to better overall accuracy, not worse."[78] In a later article coauthored with Joanna Radin,[79] Rudin highlights a number of studies that corroborate the comparable performance of interpretable and black box AI systems across a variety of safety-critical domains, including healthcare.[80,81,82] Rudin and Radin also observe that even in computer vision and image-recognition tasks, in which deep neural networks are generally considered the state of the art, a number of studies have succeeded in implementing interpretability constraints to deep learning models without significant compromises in accuracy.[83,84,85] Rudin concludes that the uncritical acceptance of the accuracy-interpretability trade-off in AI often leads researchers to forego any attempt to investigate or develop interpretable models, or even develop the skills required to develop these models in the first place.[86] She suggests that black box AI systems ought not to be used in high-stakes decision-making contexts or safety-critical domains unless it is demonstrated that no interpretable model can reach the same level of accuracy. "It is possible," claim Rudin and Radin, "that an interpretable model can always be constructed—we just have not been trying. Perhaps if we did, we would never use black boxes for these high-stakes decisions at all."[87] While, as we have argued here, it will, at least in some circumstances, be defensible to prioritize interpretability at the cost of accuracy, if Rudin is correct, the price of the pursuit of interpretability may not be as high as critics—and our argument to this point—have presumed.

Superior accuracy is, therefore, not enough to justify the use of black-box medical AI systems over less accurate but more interpretable systems in clinical medicine. In many cases, it will be genuinely

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press

uncertain *a priori* whether a more accurate black-box medical AI system will deliver greater downstream benefits to patient health and well-being compared to a less accurate but more interpretable AI system. Indeed, under some conditions, less accurate but more interpretable medical AI systems may produce better downstream patient health outcomes than more accurate but nevertheless opaque systems.

Conclusion

The prioritization of accuracy over interpretability in medical AI, therefore, carries its own lethal prejudices. While proponents of accuracy over interpretability in medical AI are correct to emphasize that the use of less accurate models carries risks to patient health and welfare, their arguments overlook the comparable risks that the use of more accurate but less interpretable models could present for patient health and well-being. This is not to suggest that the use of opaque ML AI systems in clinical medicine is unacceptable or ought to be rejected. We agree with proponents of accuracy that "black box" AI systems could deliver substantial benefits to medicine, and that the risks may eventually be reduced enough to justify their use. However, a blanket prioritization of accuracy in the technical trade-off between accuracy and interpretability itself looks to be unjustified. Opacity in medical AI systems may constitute a significant obstacle to the achievement of improved downstream patient health outcomes, despite how technically accurate these systems may be. More attention needs to be directed toward how medical AI systems will become embedded in the sociotechnical decision-making contexts for which they are being designed. The downstream effects of medical AI systems on patient health outcomes will be mediated by the decisions and behavior of human clinicians, who will need to interpret the outputs of these systems and incorporate them into their own clinical decision-making procedures. The case for prioritizing accuracy over interpretability pays insufficient attention to the reality of this situation insofar as it overlooks the negative effects that opacity could have upon the hermeneutic task of physicians in interpreting and acting upon the outputs of black box medical AI systems and subsequent downstream patient health outcomes.

Acknowledgments. Earlier versions of this article were presented to audiences at Macquarie University's philosophy seminar series, the University of Wollongong's seminar series hosted by the Australian Centre for Health Engagement, Evidence, and Values (ACHEEV), the University of Sydney's Conversation series hosted by Sydney Health Ethics, and the Ethics in AI Research Roundtable sponsored by Facebook Research, and the Centre for Civil Society and Governance of the University of Hong Kong. The authors would like to thank the audiences of these seminars for comments and discussion that improved the article.

Funding Statement. This research was supported by an unrestricted gift from Facebook Research via the Ethics in AI in the Asia-Pacific Grant Program. The work was conducted without any oversight from Facebook. The views expressed herein are those of the authors and are not necessarily those of Facebook or Facebook Research.

Conflicts of Interest. R.S. is an Associate Investigator in the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005) and contributed to this article in that role. J.H. was supported by an Australian Government Research Training Program scholarship.

Notes

1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, *et al.* A guide to deep learning in healthcare. *Nature Medicine* 2019;**25**(1):24–9.
2. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine* 2019;**25**(1):44–56.
3. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept 'black box' medicine? *Annals of Internal Medicine* 2020;**172**(1):59–60.
4. London AJ. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report* 2019;**49**(1):15–21.

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press

5. See 1, Esteva et al. 2019.
6. See 2, Topol 2019.
7. Topol EJ. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books; 2019.
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;**542**(7639):115–8.
9. Cheng Y, Wang F, Zhang P, Hu J. Risk prediction with electronic health records: A deep learning approach. In: *Proceedings of the 2016 SIAM International Conference on Data Mining*. Philadelphia, PA: Society for Industrial Applied Mathematics; 2016:432–40.
10. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. *BMC Medical Informatics and Decision Making* 2018;**18**(4):55–64.
11. Valdes G, Simone CB, Chen J, Lin A, Yom S, Pattison A, *et al.* Clinical decision support of radiotherapy treatment planning: A data-driven machine learning strategy for patient-specific dosimetric decision making. *Radiotherapy and Oncology* 2017;**125**(3):392–7.
12. Stewart J, Sprivulis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. *Emergency Medicine Australasia* 2018;**30**(6):870–4.
13. Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society* 2016;**3**(1):1.
14. See 13, Burrell 2016.
15. Wadden JJ. Defining the undefinable: The black box problem in healthcare artificial intelligence. *Journal of Medical Ethics* 2021;**4**:107529. doi:10.1136/medethics-2021-107529.
16. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. *Journal of the American Medical Association* 2018;**320**(21):2199–200.
17. Caruana R, Lou Y, Microsoft JG, Koch P, Sturm M, Elhadad N. Intelligible models for health care: Predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery; 2015:1721–30.
18. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Quality and Safety* 2019;**28**(3):231–7.
19. Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics* 2021: forthcoming. doi:10.1136/medethics-2020-107102.
20. See 19, Yoon et al. 2021.
21. Price WN. Medical malpractice and black-box medicine. In: Cohen IG, Lynch HF, Vayena E, Gasser U, eds. *Big Data, Health Law, and Bioethics*. Cambridge: Cambridge University; 2018:295–306.
22. Neri E, Coppola F, Miele V, Bibbolino C, Grassi R. Artificial intelligence: Who is responsible for the diagnosis? *La Radiologia Medica* 2020;**125**(6):517–21.
23. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. *Journal of Global Health* 2018;8(2):1–**8**.
24. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine* 2018;**15**(11):4–7.
25. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. *JAMA* **322**:497–8.
26. Hatherley JJ. Limits of trust in medical AI. *Journal of Medical Ethics* 2020;**46**(7):478–81.
27. There is a flourishing literature on the best way to characterize what is lacking in "black box AI" and thus the value that designers of AI should be promoting if they wish to avoid this problem. The "explainable AI" research program promotes "explainability" (see Defense Advanced Research Projects Agency. *Explainable Artificial Intelligence (XAI).* Arlington, VA: Defense Advanced Research Projects Agency; 2016). Other critics prefer "interpretability" (see Chen C, Li O, Tao C, Barnett AJ, Su J, Rudin C. This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems* 2019; **32**:1–12; Oliveira DM, Ribeiro AH. Contextualized interpretable machine learning for medical diagnosis. *Communications of the*

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press

ACM 2020;**63**(11):56–8). "Transparency" also has its advocates (see Rudin C, Ustunb B. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*. 2018;**48**(5):449–66). For the purposes of the current discussion, we use "interpretability" and "explainability" interchangeably.

28. See 16, Shortliffe, Sepúlveda 2018.
29. See 24, Vayena et al. 2018.
30. See 7, Topol 2019, at 29.
31. See 4, London 2019, at 17.
32. See 3, Wang et al. 2020, at 59.
33. Zerilli J, Knott A, Maclaurin J, Gavaghan C. Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy and Technology* 2019;**32**(4):666.
34. See 33, Zerilli et al. 2019.
35. See 27, Defense Advanced Research Projects Agency 2016, at 7.
36. See 4, London 2019.
37. See 17, Caruana et al. 2015.
38. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G. Explainable artificial intelligence (XAI). *Science Robotics* 2019;**4**(37):eaay7120.
39. See 13, Burrell 2016.
40. See 3, Wang et al. 2020.
41. See 4, London 2019.
42. See 4, London 2019, at 18.
43. See 4, London 2019.
44. van der Veer SN, Riste L, Cheraghi-Sohi S, Phipps D, Tully M, Bozentko K, et al. Trading off accuracy and explainability in AI decision-making: Findings from 2 citizens' juries. *Journal of the American Informatics Association* 2021;**28**(10):2128–38.
45. See 27, Defense Advanced Research Projects Agency 2016.
46. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery; 2016:1135–44.
47. Zihni E, Madai VI, Livne M, Galinovic I, Khalil A, Fiebach J, *et al.* Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PLoS One* 2020;15(4):1–**15**.
48. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology* 2018;**31**(2):1–52.
49. Guidotti R, Monreale A, Ruggieri S, Turini F, Pedreschi D, Giannotti F. A survey of methods for explaining black box models. *ACM Computing Surveys* 2018;**51**(5):1–42.
50. See 49, Guidotti et al. 2018.
51. See 49, Guidotti et al. 2018, at 16.
52. van der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Opthamologica* 2018;**96**(1):63–8.
53. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 2019;**1**(5):206–15.
54. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Health* 2021;**3**(11):e745–50.
55. Babic BB, Gerke S, Evgeniou T, Glenn Cohen I. Beware explanations from AI in health care. *Science* 2021;**373**(6552):284–6.
56. See 53, Rudin 2019.
57. Zednik C. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy and Technology* 2021;**34**:285.
58. Grote T, Berens P. How competitors become collaborators—Bridging the gap(s) between machine learning algorithms and clinicians. *Bioethics* 2022;**36**(2):134–42.

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press

59. Sparrow R, Hatherley J. The promise and perils of AI in medicine. *International Journal of Chinese and Comparative Philosophy of Medicine* 2019;**17**(2):79–109.
60. Selbst AD, Barocas S. The intuitive appeal of explainable machines. *Fordham Law Review* 2018;**87**(3):1085–139.
61. Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. 'Hello AI': Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. *Proceedings of the ACM on Human–Computer Interaction* 2019;**3**(CSCW):1–24.
62. See 60, Cai et al. 2019, at 16.
63. Mumford E. The story of socio-technical design: Reflections on its successes, failures and potential. *Information Systems Journal* 2006;**16**(4):317–42.
64. Berg M. Patient care information systems and health care work: A sociotechnical approach. *International Journal of Medical Informatics* 1999;**55**(2):87–101.
65. Baxter G, Sommerville I. Socio-technical systems: From design methods to systems engineering. *Interacting with Computers* 2011;**23**(1):4–17.
66. Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digital Medicine* 2020;**3**(1):53.
67. Grote T. Randomised controlled trials in medical AI: Ethical considerations. *Journal of Medical Ethics* 2021: forthcoming, 3. doi:10.1136/medethics-2020-107166.
68. Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 2020;**33**(2):220.
69. See 65, Gerke et al. 2020, at 2.
70. See 67, Burton et al. 2020.
71. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 2015;**144**(1):114–26.
72. See 67, Burton et al. 2020.
73. See 70, Dietvorst et al. 2015.
74. Yeomans M, Shah A, Mullainathan S, Kleinberg J. Making sense of recommendations. *Journal of Behavioral Decision Making* 2019;**32**(4):403–14.
75. See 73, Yeomans et al. 2019, at 2.
76. See 53, Rudin 2019.
77. Rudin C, Radin J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review* 2019;1(2):**1**–9.
78. See 53, Rudin 2019, at 207.
79. See 76, Rudin, Radin 2019.
80. See 17, Caruana et al. 2016.
81. Razavian N, Blecker S, Schmidt AM, Smith-Mclallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data* 2015;**3**(4):277–87.
82. See 27, Rudin, Ustunb 2018.
83. See 27, Chen et al. 2019.
84. Ming Y, Qu H, Xu P, Ren L. Interpretable and steerable sequence learning via prototypes. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery; 2019:903–13.
85. Li Y, Murias M, Major S, Dawson G, Dzirasa K, Carin L, *et al.* Targeting EEG/LFP synchrony with neural nets. *Advances in Neural Information Processing Systems* 2017;**30**:4621–31.
86. See 56, Rudin 2019.
87. See 76, Rudin, Radin 2019, at 7.

Cite this article: Hatherley J, Sparrow R and Howard M (2026). The Virtues of Interpretable Medical Artificial Intelligence. *Cambridge Quarterly of Healthcare Ethics* 35: 167–176, doi:10.1017/S0963180122000305

https://doi.org/10.1017/S0963180122000305 Published online by Cambridge University Press