Cambridge Quarterly of Healthcare Ethics (2023), 1–7
doi:10.1017/S0963180123000294

ARTICLE

On the Justified Use of AI Decision Support in Evidence-Based Medicine: Validity, Explainability, and Responsibility

Sune Holm

Department of Food and Resource Economics, University of Copenhagen, 1958 Frederiksberg C, Denmark
Email: suneh@ifro.ku.dk

Abstract
When is it justified to use opaque artificial intelligence (AI) output in medical decision-making? Consideration of this question is of central importance for the responsible use of opaque machine learning (ML) models, which have been shown to produce accurate and reliable diagnoses, prognoses, and treatment suggestions in medicine. In this article, I discuss the merits of two answers to the question. According to the Explanation View, clinicians must have access to an explanation of why an output was produced. According to the Validation View, it is sufficient that the AI system has been validated using established standards for safety and reliability. I defend the Explanation View against two lines of criticism, and I argue that within the framework of evidence-based medicine mere validation seems insufficient for the use of AI output. I end by characterizing the epistemic responsibility of clinicians and point out how a mere AI output cannot in itself ground a practical conclusion about what to do.

Keywords: Ethics; AI; medical decision-making; explainability; responsibility; evidence-based medicine

Background

Should artificial intelligence (AI)-generated predictions be deployed in medical decision-making? Should clinicians align their medical verdicts with the output of black-box AI? Is it justified to use opaque AI output in medical decision-making? Consideration of these questions is of central importance for the responsible use of machine learning (ML) models, which have been shown to produce accurate and reliable diagnoses, prognoses, and treatment suggestions in medicine.[1]

The impressive technical achievements of ML models have been proclaimed to show that medical decision-making is on the verge of a revolution. Articles and editorials in medical journals promise a future where medical decision-making is increasingly deferred to AI devices. Some even go as far as suggesting that medical reasoning can be reduced to the application of a computer program: "if all past, present, and future predictors and processes that contribute to future events were known and quantifiable, algorithms could be constructed that produce perfect risk estimates for individuals—that is, they would predict with perfect accuracy whether an event would occur or not in every individual."[2]

The most accurate ML models such as deep neural networks are opaque or black boxes in the sense that they are so complex that it is impossible for humans to comprehend why an output is produced from an input. This opacity is a concern for the use of AI systems in medical decision-making, and it has been remarked that there is "a growing chorus of clinicians, lawmakers, and researchers calling for explainable AI models."[3] Moreover, it has become clear that it is not straightforward to ensure the deployment of AI systems in clinical settings and that clinicians hesitate to use accurate and reliable AI support in decision-making. [4] Thus, the expected improvements in patient care are not achieved.

Research exploring why clinicians are reluctant to deploy AI systems suggests that clinicians find that in order to justify AI-supported decisions, they must have access to an explanation of why an output was produced from an input.[5] Ensuring explainability is expected to make clinicians more willing to trust a black-box AI output and thus use AI more frequently in decision-making. I will call the view that explanations are required for the justified use of AI systems the Explainability View.

As a result of the need for explanations, there has been a surge in developing methods that can generate *post hoc* explanations of why a black-box AI produced an individual output. Examples of such explanations are heat maps that provide clinicians with images where, for example, the red colors indicate areas that the AI model considers high importance and other colors, for example, blue, indicate areas with lower importance.[6] Because they allegedly can provide explanations of individual AI output, explainable AI (XAI) methods are expected to make these outputs more trustworthy to clinicians and thereby make clinicians more willing to use them to make decisions.

The use of the notion of trust in the debate about opaque AI is confusing. According to a standard distinction in philosophy, there are two forms of trust: trust as *mere reliability* and *genuine trust*.[7] According to this distinction, an opaque AI cannot be genuinely trustworthy and trusted because it cannot entertain the right motives. However, it can be trustworthy in the sense of merely reliable if its accuracy and reliability have been carefully validated. The central question often articulated as a problem of trust is really a question about whether clinicians are justified in using the output of an opaque but reliable and accurate AI when making medical decisions. On the assumption that the performance of the AI has been validated, clinicians would, as a minimum, seem to be justified and perhaps even obligated to use such output as premises when reasoning about what to do.[8]

While recognizing the value of *post hoc* explanations for the development of fair and reliable AI systems, some scholars caution that XAI methods should not be deployed to explain and justify the use of individual output.[9] The proper way to justify black-box AI output is to ensure the "thorough and rigorous validation of these systems across as many diverse and distinct populations as possible, showing that patient and health-care outcomes are improved and that marginalised groups are not disproportionately affected by any given system."[10] Clinicians should rely on the output of validated AI systems and not require explanations. I call this the Validation View of AI justification.

In support of the Validation View, its proponents argue that many drugs and devices used routinely in medicine are in fact black boxes.[11] And in the case of drugs, the mechanism that generates the desired effect might not be scientifically accounted for despite decades of research. Still, such drugs are considered acceptable for use because they have been validated using "gold standard" randomized controlled trials. If the output of black-box AI systems has been validated using such standards, then they should be considered justified for clinical use, too.

The debate about the justified use of opaque AI output in medical decision-making can thus be divided into two views. On the Validation View, validation is *sufficient* for justified use. On the Explanation View, explanations are *necessary* for justified use. Here is a statement of the two views:

> *Explanation View* If a clinician is justified in using a black-box AI output in medical decision-making, then the clinician must be offered a post hoc explanation enabling her to understand why the output was produced.[12]
>
> *Validation View* If a black-box AI device has been validated according to the standards of evaluating the safety and reliability of medical drugs and devices, then a clinician is justified in using its output in decision-making.

Against the Explanation View

Critics of the Explanation View point out that *post hoc* explanations suffer from at least two problems. One problem is *the interpretability gap*. Consider a heat map of lung images produced by XAI methods.[13]

https://doi.org/10.1017/S0963180123000294 Published online by Cambridge University Press

The heat map will indicate the relative importance of areas of the image for its output. The interpretability gap arises because it is not clear from the heat map *what* it is about the hot areas that the AI considers important for its output. Thus, it will be up to the clinician to "fill out the blanks" in the explanation. However, as is well documented, humans tend to confirm their own perception, and thus it is very likely that the clinician will assume that what mattered to the model is the same features of the highlighted area that the clinician herself finds important.[14]

In response to this problem, heat maps may be supplemented with natural language text stating what features it takes to be significant. This will enhance the clinician's ability to interpret the heat map. Thus, multimodal models comprising, for example, both visual cues and natural language will mitigate the interpretation gap.[15] In this context, it is important to highlight that how fine-grained the explanations should be will depend on the purpose of the audience.[16] There is not a one-size-fits-all level of granularity that an explanation must achieve to explain an output.

The second problem is that post hoc explanations merely generate "rationales of black-box predictions, which are not necessarily the actual reasons behind those predictions or related causally to them."[17] Accordingly, *post hoc* explanations are unlikely to contribute to our understanding of the inner workings of the model. Instead, we are left with the false impression that we understand it better."[18] Rudin seconds this criticism claiming that *post hoc* explanations "must be wrong" because they misrepresent the way in which the opaque model arrives at its output.[19]

Proponents of the Explanation View may respond to the criticism in different ways. Against the claim that *post hoc* explanations cannot be genuine explanations, it can be argued by analogy that if *idealized* scientific models such as the ideal gas law can provide genuine explanations that enable people to better understand complex natural phenomena, then XAI methods can provide genuine explanations too.[20]

Like scientific models, the simplified models deployed to explain the black box will not be completely faithful to the target they are modeling, but that does not exclude them from providing genuine explanations in the sense of enabling users to make causal inferences relevant for some purpose.[21] The real challenge for XAI is to produce explanation models that meet the standards of validated idealized scientific models. However, there is much work going on to develop XAI methods, so it would be premature to assume that they will not become available.

The Insufficiency of Validation

I have presented how proponents of the Explanation View may respond to two central objections to requiring explanations of AI output to be used when making medical decisions. In this section, I consider the case for the Explanation View within the framework of evidence-based medicine (EBM) and shared decision-making (SDM).

Some scholars find it "hard to imagine a person who would feel comfortable in blindly agreeing with a system's decision in … highly consequential and ethical situations without a deep understanding of the decision making rationale of the system."[22] This worry is succinctly expressed by Lipton when he imagines a model developer exclaiming that "we can train a model, and it can even give us the right answer. But we can't just tell the doctor 'My neural network says this patient has cancer!' The doctor just won't accept that! They want to know why the neural network says what it says. They want an explanation."[23] Moreover, as already noted, much of the discussion of trustworthy AI, in fact, centers on the observation that clinicians hesitate to accept AI output when explanations are unavailable.[24]

To require explanations is to claim that the fact that predictions are produced by an accurate and reliable source may not be sufficient for being justified in using them when making medical decisions. Call this the *Insufficiency Claim.* Proponents of the Explanation View are committed to the Insufficiency Claim. However, we should ask whether the Insufficiency Claim is justified.

The central aim of EBM is "to ensure that decision making in health care incorporates the best available evidence."[25] Importantly, the incorporation of the best evidence is supposed to be *judicious* "taking into account both clinical expertise and the needs and wishes of individual patients."[26] As outlined, validated black-box AI provides decision-makers with predictions that may be useful for

https://doi.org/10.1017/S0963180123000294 Published online by Cambridge University Press

making informed decisions. And while EBM is in large part associated with a focus on "cold hard facts," properly understood it is "a systematic approach to clinical problem solving which allows the integration of the best available research evidence with clinical expertise and patient values"[27] Importantly, clinical expertise involves both interpretation and appropriate application of the evidence in the circumstances.[28] A central clinical skill is thus to be able to "interpret the evidence and apply it appropriately to the circumstances—doing the right things."[29] Finally, EBM involves communication to patients of "the information they need to make an informed choice."[30]

Respect for patient values and autonomy is typically understood in terms of the ideal of SDM. In SDM, "clinicians and patients share the best available evidence when faced with the task of making decisions, and (…) patients are supported to consider options, to achieve informed preferences."[31] Thus, SDM emphasizes patient autonomy or self-determination, and "clinicians need to support patients to achieve this goal, wherever feasible." Still, as Sandman and Munthe point out, aiming for patient autonomy need not result in "abandoning the patient or giving up the possibility to influence how the patient is benefited."[32] It is thus too simplistic to think of SDM as an exchange between clinician and patient where clinicians simply provide the patients with "facts about the diagnosis and about the prognoses without treatment and with alternative treatments" and leave patients to make decisions on their own.[33] To support autonomous and informed decision-making, patient and clinician must engage in shared *deliberation* on the basis of shared information.[34] Thus, if "the patient is not provided with sufficient information concerning the confidence of a given diagnosis or the rationale of a treatment prediction, she might not be well equipped to give her consent to treatment decisions."[35]

To illustrate how SDM requires more from clinicians and AI systems than merely being offered the best available evidence, Bjerring and Busch ask us to imagine that an opaque AI is deployed to rank treatment options for a patient with breast cancer.[36] It recommends that both breasts be surgically removed and informs about what side effects can be expected. When deliberating with the patient about what treatment to choose, the clinician will arguably not be in a position to provide the sort of information required to honor the principles of SDM and, in turn, EBM. A central piece of information that is missing concerns *why* the AI system finds a treatment involving the removal of both breasts to be more likely to have good results than alternatives. Simply being told that the AI has been validated to be very accurate and reliable does not seem to suffice for justifying deferring to its output when making this decision.

This consideration highlights that it is the ideal of SDM that introduces explainability as a requirement for justification. The clinician may be said to be justified in thinking that there is good reason to think that the AI system gets it right because it has been adequately validated. However, a significant class of medical decisions—those governed by the principles of SDM—require explainability too. Hence, my claim here is not that a clinician may not, in some cases, be justified in deferring to an AI system's output simply due to its validated accuracy and reliability, when making a medical decision. The point is that insofar as SDM principles are in play, more will seem to be required.

I have argued that on recent accounts of EBM, it seems plausible to require that for clinicians to be justified in using black-box AI output in their deliberations with patients, such output should be explained, and that it is not in principle impossible for XAI methods to provide such explanations.

Implementation and Responsibility

Coming back to the initial issue concerning how to address the low uptake and unwillingness to use AI output, an important insight from the current debate is that when managers want to implement AI devices in the clinic, they must recognize that it is the medical decision-makers that are responsible for curating and communicating evidence relevant for shared decision-making. Hence medical decision-makers must be able to justify and deliberate about the relevance and quality of the evidence that they present in favor of a diagnosis or treatment suggestion regarding an individual in a particular context. Responsible decision-makers must be able to understand the evidence that they rely on in their reasoning.[37]

Thinking about the responsible implementation of opaque AI output forces reflection on the notions of clinical reasoning and judgment. The introduction of accurate and reliable AI tends to emphasize that

https://doi.org/10.1017/S0963180123000294 Published online by Cambridge University Press

such tools produce valuable information about the condition of the patient. In that way, clinical reasoning and judgment are more complex tasks than translating an AI output into a decision. Medical decision-making is characterized by uncertainty along several dimensions,[38] and AI output does not in itself mitigate all these uncertainties. It must still be negotiated by the clinical reasoner.[39]

Given that clinical reasoning involves a range of tasks that cannot be deferred to AI systems, but must be undertaken by clinicians in collaboration with patients, it seems important to recognize clinicians as *epistemologically responsible.*[40] They are the ones who will have to fit the output of AI systems into a more comprehensive picture of the patient that incorporates and balances other information than the statistical evidence used by the AI system to arrive at its output.

Importantly, given the epistemic responsibility of the clinicians curating the available information and best evidence about the patient, it only seems appropriate that they have the means to take on this responsibility. Being epistemically responsible, they must also be able to justify their reasoning and judgments. Requiring explanations of opaque AI output seems to support this.

These considerations lead to a final important observation. AI is often presented as “better than experts” at some tasks such as diagnosing skin cancer. However, given the argument of this article, this way of presenting AI output is misleading in a clinical context. Making decisions about diagnosis, prognosis, and treatment is not a merely statistical procedure. The practical reasoning behind a medical decision will typically include premises referring to scientific evidence and statistical information. However, being *practical* reasoning, it also involves evaluative premises.

Practical reasoning, as I understand it here, is reasoning about what to *do.*[41] This is a normative question posed from the perspective of the individual patient. Which of the available options is the best treatment for patient Hannah? As such, it involves reference to evaluative premises, that is, premises based on Hannah’s personal values. For example, Hannah might find a certain kind of treatment very bad because it will make her unable to pursue an activity, which she finds very important to her quality of life. Or she might have certain religious beliefs that makes her exclude a certain type of treatment. These value considerations form part of the practical reasoning that a clinician will perform, in addition to available scientific evidence and statistical information about what to expect given a certain treatment. Hence, there is no way for the statistical information provided by an AI output to directly determine the right decision. And the further information to be added is (also) the responsibility of the clinician. Hence, the information afforded by the AI output cannot in itself determine medical decisions about individual patients regardless of how accurate and reliable it is.

Funding Statement. This work received funding under Project number ERAPERMED2021-324 and Innovation Fund Denmark.

Competing interest. The author declares no competing interests.

Notes

1. See Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *NPJ Digital Medicine* 2020; available at https://doi.org/10.1038/s41746-020-00324-0 for a recent survey of AI devices approved for marketing by the FDA.
2. Sniderman AD, D’Agostino RB, Pencina MJ. The role of physicians in the era of predictive analytics. *JAMA* 2015;314:25–26.
3. Ghassemi M., Oakden-Rayner L, Beam A. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 2021;**3**:e745–50.
4. For the lack of uptake, see, for example, Lindsell CJ, Stead WW, Johnson KB. Action-informed artificial intelligence-matching the algorithm to the problem. *JAMA* 2020;**323**:2141–42. For the need to better integrate AI in clinical workflow see also Shah NH, Milstein A, Bagley, SC. Making machine learning models clinically useful. *JAMA* 2019;**322**:1351–2. For a suggestion about how to improve

https://doi.org/10.1017/S0963180123000294 Published online by Cambridge University Press

clinical relevance and uptake see Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine 2019;25:1337–40. For the evaluation and validation of ML models for use in clinical care, see, for example, McCradden MD, Stephenson EA, Anderson JA. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nature Medicine* 2020;26:1325–6.

[5] Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: Contextualizing explainable machine learning for clinical end use. *arXiv* 2019; available at http://arxiv.org/abs/1905.05134 (preprint).

[6] Ghassemi M, Oakden-Rayner L, Beam A. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 2021;**3**:e745–50.

[7] Hawley K. *How to Be Trustworthy*. New York, NY: Oxford University Press; 2019.

[8] Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. *Philosophy* & Technology 2020;**34**:349–71.

[9] See note 6, Ghassemi et al. 2021.

[10] See note 6, Ghassemi et al. 2021.

[11] See note 6, Ghassemi et al. 2021. See also London AJ. Artificial intelligence and black-box medical decisions: Accuracy versus Explainability. *Hastings Center Report* 2019;**49**:15–21.

[12] As Miller shows, such why questions are typically asked in a *contrastive* sense. What people want to know is why the AI produced this output rather than another. Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 2019;**267**:1–38.

[13] For a description of machine learning approaches to generating heat maps, see Montavon G, Samek W, and Müller K-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 2018;**73**:1–15; Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine* 2018;15: e1002686; Zednik C. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy* & Technology 2019;**34**:265–88.

[14] For example, Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems* 2018;**31**:9505–15; Gu J, Tresp V. Saliency methods for explaining adversarial attacks. *arXiv* 2019; available at http://arxiv.org/abs/1908.08413 (preprint). For further discussion of current XAI methods, see Grote T, Berens P. Uncertainty, evidence, and the integration of machine learning into medical practice. *The Journal of Medicine and Philosophy* 2023. doi:10.1093/jmp/jhac034.

[15] See note 14, Grote, Berens 2023.

[16] Nyrup R, Robinson D. Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics Inf Technol* 2022. doi:10.1007/s10676-022-09632-3.

[17] Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. *Science* 2021;**373**:284–6.

[18] See note 6, Ghassemi et al. 2021, the authors write that "on an individual level, the explanations we can produce for the behaviour of complex AI systems are often confusing or even misleading."

[19] Rudin C. Stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 2019;**1**:206–15.

[20] Fleisher W. Understanding, idealization, and explainable AI. *Episteme* 2022;**19**:534–60.

[21] See note 16, Nyrup, Robinson 2022.

[22] Doran D, Schulz S, Besold TR. What does explainable ai really mean? A new conceptualization of perspectives. *ArXiv* 2017; available at https://doi.org/10.48550/arXiv.1710.00794.

[23] Lipton ZC. The doctor just won't accept that! *arXiv* 2017; available at https://doi.org/10.48550/arXiv.1711.08037.

[24] Binns R, van Kleek, M, Veale M. Lyngs U, Zhao J. and Shadbolt N. '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* 2018; available at https://doi.org/10.1145/3173574.3173951. Ferrario A, Loi M. How explainability contributes to trust in AI. In: *2022 ACM*

Conference on Fairness, Accountability, and Transparency 2022; available at https://doi.org/10.1145/3531146.3533202.

25. Bate L, Hutchinson A, Underhill J, Maskrey N. How clinical decisions are made. *British Journal of Clinical Pharmacology* 2012;**74**:614–20.
26. See note 25, Bate et al. 2012.
27. Haynes RB, Sackett DL, Richardson WS, Rosenberg W, Langley GR. Evidence-based medicine: how to practice and teach EBM. *Canadian Medical Association Journal* 1997;**157**(6):788.
28. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *BMJ Evidence-Based Medicine* 2002;**7**:36–8.
29. See note 28, Haynes et al. 2002.
30. See note 28, Haynes et al. 2002.
31. Elwyn G, Coulter A, Laitner S, Walker E, Watson P, Thomson R. Implementing shared decision making in the NHS. *BMJ* 2010;**341**:c5146.
32. Sandman L, Munthe C. Shared decision-making and patient autonomy. *Theoretical Medicine and Bioethics* 2009;**30**:289–310.
33. Brock DW. The Ideal of shared decision making between physicians and patients. *Kennedy Institute of Ethics Journal* 1991;**1**:28–47.
34. Epstein RM, Fiscella K, Lesser CS, Stange KC. Why the nation needs a policy push on patient-centered health care. *Health Affairs* 2010;**29**:1489–95.
35. See note 14, Grote, Berens 2023.
36. See note 8, Bjerring, Busch 2020.
37. See also Hatherley J, Sparrow R, Howard M. The virtues of interpretable medical AI. *Cambridge Quarterly of Healthcare Ethics* 2023. doi:10.1017/S0963180122000664.
38. Djulbegovic B, Hozo I, Greenland S. Uncertainty in clinical medicine. In: Gifford F, Gabbay DM, Thagard P, Woods J, eds. *Philosophy of Medicine*. Amsterdam: Elsevier; 2011:299–356.
39. Chin-Yee B, Upshur R. Clinical judgement in the era of big data and predictive analytics. *Journal of Evaluation in Clinical Practice* 2018;**24**:638–45.
40. van Baalen S, Boon M. From EBM to epistemological responsibility. *Journal of Evaluation in Clinical Practice* 2015;**21**:433–9.
41. Wallace RJ. Practical reason *The Stanford Encyclopedia of Philosophy* 2020; available at https://plato.stanford.edu/archives/spr2020/entries/practical-reason/.

Cite this article: Holm S (2023). On the Justified Use of AI Decision Support in Evidence-Based Medicine: Validity, Explainability, and Responsibility. *Cambridge Quarterly of Healthcare Ethics*: 1–7, doi:10.1017/S0963180123000294