Future Healthcare Journal 11 (2024) 100171

Contents lists available at ScienceDirect

Future Healthcare Journal

journal homepage: https://www.sciencedirect.com/journal/future-healthcare-journal

Explaining decisions without explainability? Artificial intelligence and medicolegal accountability

Melissa D. McCradden[a,b,c,*], Ian Stedman[d]

[a] Australian Institute for Machine Learning, University of Adelaide, Australia
[b] Women's and Children's Hospital, Adelaide, Australia
[c] SickKids Research Institute, Toronto, Canada
[d] School of Public Policy and Administration at York University, Toronto, Ontario, Canada

Introduction

Explanations are fundamental to our culture. From medical ethics to cookie policies, the range of situations where we seek explanations – and in fact, in some cases, have a *right* to them – is broad. It makes sense that, for artificial intelligence (AI) in medicine, explanations are widely desired in situations where a prediction made that we wish to use.

The issue isn't whether or not we need explanations, but what the term 'explanation' entails. And right now in health AI we have a problem, because the way in which the term explanation is being understood is out of sync with the actual computational evidence.[1-3] Further, contrary to what one might think, we have increasing evidence that shows explanations might worsen decision-making in some situations.[4-6]

In this article, we will cover the state of the science regarding explainability techniques that are applied to health AI tools, the evidence emerging about its effects on decision-making, and the current medicolegal landscape that may apply. We argue that, strictly speaking, 'explanations' as currently understood are not sufficient and may not be necessary for good clinical decision-making. A good clinical decision is not only one that advances the goals of care, but it also has to be legally defensible. Clinicians must calibrate their judgement against a whole constellation of other factors, even if they are using an AI tool that is well validated and highly accurate. We offer two case examples where we demonstrate how ethical and medicolegally accountable decisions can be made without reliance on explainability.

Explainabilty: what's in an explanation?

There are two types of explainability approaches: ***inherent*** explainability (also know as interpretability), which refers to understanding how the model as a whole functions, and ***post hoc*** explainability (also known as instance-level), which refers to attempts to understand the means by which a specific prediction was generated by the model. Some models are directly interpretable, meaning that the operations from inputs to outputs are easy to follow and clear (eg decision trees); others are more opaque, meaning that the process from inputs to outputs is difficult or impossible to follow precisely, even for developers (eg deep learning). We focus specifically on post hoc explainability in this piece, as this is a more controversial issue.

A common use of post hoc explanations is heat mapping for image-based AI tools. In these cases, the model purportedly highlights areas of the image proportionate to their influence on the model's prediction. If the model predicts that a patient has a pneumothorax, the explanation should highlight exactly that area on the image. Only this is not quite what happens in every case; models can highlight both relevant and non-relevant areas, may highlight areas that are important but not to the specific task of interest to the clinician, and produce the same explanation even when the clinical facts can differ dramatically.[2,3,7,8] We often find explanations compelling simply because we assume that if the model is highlighting the same area that we would believe is relevant, we believe the model is generating the decision the same way that we, as humans, would.[9] However, this is not true.[2,3,9]

The same criticism is offered of other explainability methods (eg Shapley values, locally interpretable model-agnostic explanations, and other feature-based explainability techniques).[2] The bottom line is that none of these explanations are proven techniques for providing *specific, individual accounts* of how a prediction was generated for a specific patient.

Clinician, explain thyself

A major challenge in health AI is that clinicians typically believe that an 'explanation,' as commonly understood,[1] is what they are getting when they see something like a heatmap or a prediction accompanied by the 'reasons' why the patient received this output.[4] This belief may be what is contributing to a phenomenon called 'automation bias' – the tendency to over-rely on machine-based decisions and disregard human ones. Automation bias has a long history of study in fields like aviation,[10] and its clear relevance to health AI has lead to the expansion of research in human–computer interaction.

This article reflects the opinions of the author(s) and should not be taken to represent the policy of the Royal College of Physicians unless specifically stated.

* Corresponding author at: Australian Institute for Machine Learning, Lot Fourteen, Adelaide, SA 5000, Australia.
E-mail address: melissa.mccradden@adelaide.edu.au (M.D. McCradden).

https://doi.org/10.1016/j.fhj.2024.100171
Received 26 July 2024; Accepted 6 August 2024

Explanations themselves might independently influence decision-making – for better, or for worse. For example, Tschandl and colleagues showed that when AI outputs were incorrect, clinician decisions were worse than when made without the AI tool across a range of levels of experience.[6] Jacobs and colleagues tested five conditions and found that incorrect AI recommendations worsened clinician judgement overall, but this was most pronounced with feature-based explanations (where the model identifies the specific aspects of the patient in question which supposedly related to the output).[4] Individual factors may play a strong role in how clinician judgement is influenced; Gaube and colleagues found that, across a range of experience levels, some clinicians accepted every incorrect recommendation from AI, while others rejected every incorrect recommendation from AI.[5]

So, if explainability isn't a reliable way of accounting for individual-level predictions, and can worsen our judgement when the output is incorrect (and every AI tool will have some proportion of incorrect outputs – no model is perfect!), should we use it at all?

The ethical imperative for explainability?

The call for the need for explanations has spread across many disciplines, most notably ethics and regulatory environments. Some ethics guidance documents go so far as to consider explanations to be essential to 'ethical AI'.[11] Many in the literature and in common parlance consider the ability to explain a model's prediction as necessary.[12] We disagree, for the reasons listed above. While we encourage further developments in the field of explainable AI, we suggest that explainability alone cannot and should not serve as an *essential component* of ethical decision-making.

Keeping in mind the common notion of explanations, it is important to note that this is not a standard to which other areas of medicine are held.[13] A clinician does not need to know the specific mechanism behind a particular drug's action in order to responsibly prescribe it to a patient.[13] Rather, they need to know that it *works* – they draw from clinical evidence of efficacy and safety, understanding of the patient population it was evaluated within, and the conditions under which it was evaluated (eg stage in disease trajectory, relevant comorbidities, side effects, etc).[14] We have proposed that the intervention ensemble for clinical AI tools can provide an analogous foundation for responsible use.[15]

A final concern is that explainability centres the tool and not the patient. Medical decisions have historically been grounded in the interests of the patient, rather than a deterministic set of actions based on any individual technology. Piling more weight onto the value ascribed to the AI tool's output further shifts the emphasis away from the patient – their wishes, their culture, their context.

While the field progresses, AI tools may have variable evidentiary backing, which is beyond the control of the clinician. To provide guidance, we turn to the medicolegal standard of reasonableness (See Fig. 1).

Making 'reasonable' clinical decisions

While previous scholars have emphasised medicolegal accountabilities as a binary issue (eg clinician is right/wrong, AI is right/wrong),[16] we approach our analysis from the position that these issues will not always be clearly binary, nor do we consider AI tools as the sole source of information on which a clinician would rely for clinician decision-making.

Somewhat distinct from being 'good' or 'ethical', clinical decisions must meet a 'reasonableness' standard under the law. Decisions can be reasonable even when they lead to a bad outcome. Clinicians using AI should consider what would be accepted as reasonable under the law, particularly when navigating uncertainty about the potential outcomes for their patient (Box 1).

Box 1. Case examples

Case 1: Predicting discharge readiness

Consider a clinician using an AI tool that analyses biological signals to predict the likelihood of a patient's discharge readiness (ie whether they are likely to be readmitted within 48 h). Patient A lives in the city centre, has private health insurance and is relatively well-resourced. Patient B lives in a rural area, is underinsured and, based on their medical history, less likely to seek healthcare when needed.

The risk for patient A is lower if they are discharged and the prediction is not correct – because what the clinician knows about patient A suggests that they can easily come back to the well-resourced, low wait-time hospital they were discharged from. For patient B, however, the information that should be considered when making a decision about discharge should lead an attentive clinician to conclude that the risk will be higher if the AI prediction is incorrect. Even if the discharge readiness prediction is the same for both patients, a court might determine that anything that is reasonably knowable by the clinician at the time of the decision ought to be taken into consideration when making the decision. If that clinician has learned anything about the impact that social determinants can have on health outcomes, then the court might determine that they ought to have at least attempted to control for social determinants in their decision-making process. In other words, the discharge risk might be relative to what we ought to reasonably know about the patient and not simply based on an algorithmic prediction. What guides the clinician in these cases is the knowledge about the patient's unique context and a commitment to their best interests above all else.

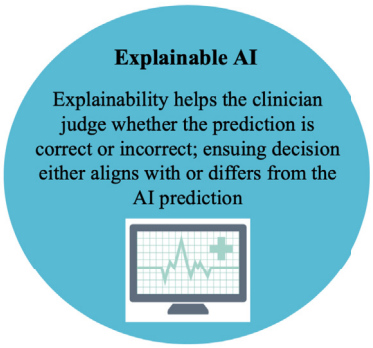
Case 2: ChatGPT for diagnostics

Consider the example of using ChatGPT for diagnostic purposes, where a physician is exploring a differential diagnosis for their patient who is experiencing multiple symptoms which could be driven by a number of potential causes. The physician might use ChatGPT to input the symptoms and see what comes up.

OpenAI clearly includes a disclaimer that ChatGPT should not be used for medical advice. They also include the disclaimer that ChatGPT can 'hallucinate' (ie output false and untrue information). ChatGPT is not a regulator-approved medical device. Physicians are thus assuming the role of a learned intermediary and assuming any liability resulting from the use of these systems.[17]

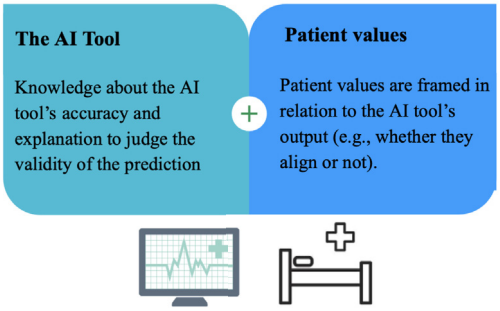
Physicians have used search engines to assist them for decades – the use of a different form of technology is not new. But how those outputs are used may be unique to AI. The same way that it would be considered unreasonable to click on a random link in a list of search engine results and decide that is the diagnosis, so too is it unreasonable to take whatever predictions are made by ChatGPT as definite. The possible risks at this stage can include confirmation bias, where an output sets a physician down a particular diagnostic trajectory to the exclusion of other possibilities. This risk is particularly concerning given the evidence of surreptitious and subtle racial bias that readily misleads physicians against evidence-based practice.[18] Should a delay occur in the correct diagnosis, the patient be subjected to an unnecessary test or other harms occur, the courts might consider whether it was reasonable to rely on ChatGPT's output. Even though ChatGPT and similar tools can identify under-recognised, rare conditions, so too can they identify entirely erroneous conditions which can result in serious harms.[19]

What can guide the physician in this case is to consider what a similarly situated physician would do in the absence of ChatGPT – would the proposed decision be considered a reasonable one? What other information and evidence can, together, form the justificatory foundation for the proposed decision? Finally, what are the potential consequences of the proposed decision versus other possible decisions, and how can harm be minimised to the patient, all things considered? (See Fig. 1).



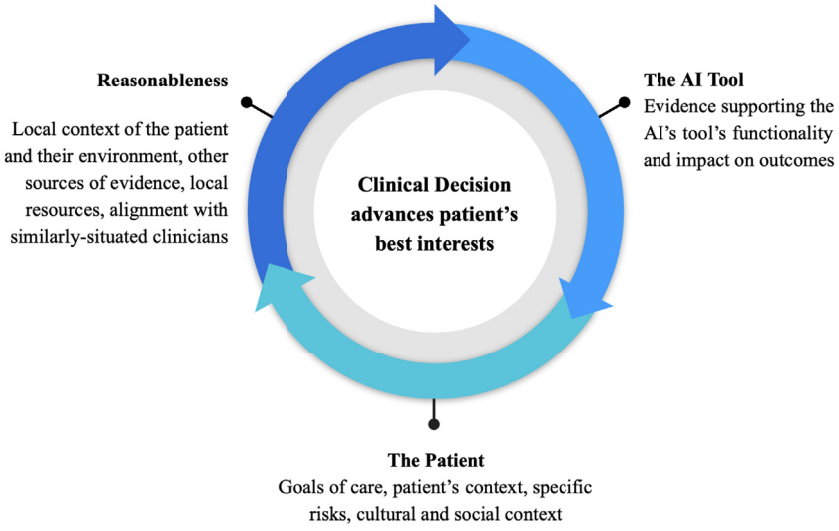
Current dominant paradigm

In much of the published literature, a good decision is presumed to be a factor of whether or not the prediction is valid. Clinicians use explanations to determine whether to follow the prediction or not. The clinician decision then rests on the correct interpretation of the AI output. Patients are de-centred, as is relevant context surrounding their care, and emphasis is placed on the AI output.



The 'AI + Patient Values' paradigm

Attempting to advance the dominant paradigm to include patient values, many have articulated to importance of retaining the patient's wishes as a guidepost for decisions. Still, this framing positions patient values as either aligned to or differing from the AI prediction, which retains the centring of the AI tool as the primary factor in decision-making.



Ethical practice and the reasonableness standard

Instead of focussing on the AI prediction alone, clinicians should consider the specific evidence behind the tool's testing against real-world outcomes to calibrate their judgment. To contextualize the AI prediction, they should consider the goals of care, social context, and specific risks and benefits of the decision for the individual patient. They should consider what additional information can be used to support their decision, and whether they think similarly-situated clinicians would make the same decision. Finally, they should document clearly the reasons behind their clinical recommendation, which may include the AI prediction as one component of the larger picture.

Fig. 1. Clinical decision-making paradigms.

Medical malpractice happens when there is a duty of care that is not met, which then results in harm to the patient. It is the 'standard of care' that must be met when a duty of care exists. The standard of care is determined by establishing what a similarly situated physician with access to similar resources would have done in the same circumstances.

AI tools complicate this picture for a few reasons: AI tools developed 'in-house' are hyper-localised; the evidence base for AI's efficacy is variable (some tools work well, other tools don't); sometimes there is no way to know whether a particular prediction was 'right' or not. Until caselaw establishes a standard of care for the use of a particular AI,[20,21] courts are unlikely to simply ask 'is the AI tool accurate and did you follow it?'. Reasonable clinical decision-making is much more nuanced.

Historically, reasonable judgements have been made on the basis of the totality of evidence available to the clinician, contextualised in light of the patient's specific situation. It is highly unlikely that an AI prediction would be the sole source of information by which a clinician makes a decision, particularly as their performance is never 100% perfect. It will, for the foreseeable future, always be necessary to

triangulate sources of evidence to point to a reasonable decision. In this sense, physicians should consider what, specifically, the AI tool's output contributes to the overall clinical picture (Fig. 1).

Finally, clinicians should clearly and thoroughly document the reasons behind their decisions. The reasoning process is critical to tracing the clinician's judgement, particularly where there is the potential for harm to the patient.

Conclusion

When it comes to AI for patient care, it is still early days. Due to the uncertainty around explainability contrasted with the generally well-established reasonableness standard, explainability is not a suitable foundation for good decision-making (Fig. 1). Instead, we advocate for physicians to utilise the totality of evidence available to them to factor in how the information supplied by AI fits within this larger picture. Moreover, it is often said that the law is the minimum standard to which we should strive; care should always be guided by a desire to act in a patient's best interest.

Ethics Information

This is a conceptual/analytical commentary piece and so ethics approval is not required.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Melissa McCradden reports financial support was provided by The Hospital Research Foundation Group. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Melissa D. McCradden: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. **Ian Stedman:** Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing.

References

1. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. *Machine learning for healthcare conference*; 2019:359–380.
2. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745–e750.
3. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *Adv Neural Inf Process Syst*. 2018:31.
4. Jacobs M, Pradier MF, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl Psychiatry*. 2021;11:108.
5. Gaube S, Suresh H, Raue M, *et al*. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med*. 2021;4(1):31.
6. Tschandl P, Rinner C, Apalla Z, *et al*. Human–computer collaboration for skin cancer recognition. *Nat. Med.*. 2020 Aug;26(8):1229–1234.
7. Gu J, Tresp V. *Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Saliency methods for explaining adversarial attacks; 2019.
8. Rajpurkar P, Irvin J, Zhu K, et *al*. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225. 2017 Nov 14.
9. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1:206–215.
10. Cummings ML. Automation bias in intelligent time critical decision support systems. *Decision making in Aviation*; 2017:289–294.
11. Ethics and Governance of Artificial Intelligence for health: WHO Guidance. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO.
12. Amann J, Vetter D, Blomberg SN, *et al*. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Health*. 2022;1(2):e0000016.
13. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*. 2019;49(1):15–21.
14. Kimmelman J, London AJ. The structure of clinical translation: efficiency, information, and ethics. *Hastings Center Report*. 2015;45(2):27–39.
15. McCradden MD, Joshi S, Anderson JA, London AJ. A normative framework for artificial intelligence as a sociotechnical system in healthcare. *Patterns*. 2023;4(11).
16. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA*. 2019;322(18):1765–1766.
17. Stedman, I., Brudno, M. Trust, Tort law and the integration of black box artificial intelligence into clinical care. Health Law in Canada. 2021;2(2).
18. Omiye JA, Lester JC, Spichak S, *et al*. Large language models propagate race-based medicine. *NPJ Digit Med*. 2023;6:195.
19. Longwell JB, Hirsch I, Binder F, *et al*. Performance of large language models on medical oncology examination questions. *JAMA Network Open*. 2024;7(6):e2417641.
20. Ross, C. & Herman, B. "Denied by AI" Investigate Series by STAT News. March, July, November, December 2023. Available from: https://www.statnews.com/denied-by-ai-unitedhealth-investigative-series/.
21. Froomkin AM, Kerr I, Pineau J. When AIs outperform doctors: confronting the challenges of a tort-induced over-reliance on machine learning. *Ariz L Rev*. 2019;61:33.