Discussion Paper:

The Limits of Explainability in Health AI - Why Current Concepts of AI Explainability Cannot Accommodate Patient Interests

Thomas Ploug[1,2] & Søren Holm[3,4]

[1] Centre for AI Ethics, Law, and Policy, Department of Communication and Psychology, Aalborg University

[2] School of Health and Welfare, Halmstad University

[3] Centre for Social Ethics and Policy, Department of Law, School of Social Sciences, University of Manchester

[4] Centre for Medical Ethics, HELSAM, Faculty of Medicine, University of Oslo

Abstract

In this paper we explicate the general concept of 'an explanation' and show that because there are many kinds of explanation there must be many kinds of 'explainability'. Subsequently we analyse the types of explanations we can give of Artificial Intelligence (AI) systems and their output using current explainability methods, and then discuss what types of explanation patients are likely to seek as part of the diagnostic process or as part of choice of therapy. We argue that the types of explanation that is provided by current AI explainability methods do not adequately answer many reasonable requests for explanation that patients can make when their diagnosis or treatment choice has involved the use of AI advice.

Keywords: Artificial Intelligence, Contestability, Explanation, Explainability

INTRODUCTION

Artificial intelligence (AI) systems for diagnosis and treatment choice are currently being introduced into routine use in a wide range of health care settings. (Rajpurkar et al., 2022) The visions for the future development and use of AI in health care are grand and it is claimed that AI will revolutionise health care and the role of health care professionals in the near future. (Topol, 2019) The use of AI systems is also a crucial component in the visions for precision and personalised medicine.(Bonkhoff & Grefkes, 2022; Nirvik & Kertai, 2022; Subramanian et al., 2020) Most of the AI systems being implemented and under development are based on a number of machine-learning architectures such as neural networks, support vector machines, random forest classifiers or similar.(Ferdous et al., 2020) These architectures are all algorithmic. When the AI model has been trained there is a determinate connection between the input and the output, and if the learning function has been turned off and the model 'locked' the same input will always produce the same output. Despite being algorithmic these AI systems are nevertheless 'black boxes' in the sense that it is almost always impossible

for human beings, even those who are domain experts to understand in detail how a certain input leads to a certain output. The systems are too complex and they may pick up on features in the input that have no readily available human interpretation, e.g. a certain pixel texture in a subsection of an X-ray. This is widely recognised as a problem. To simply quote the Little Britain catch phrase "Computer says No!" to a patient would obviously not be a good reason to deny that patient treatment if that is all that can be said. The most common response to this problem in the literature is to claim that AI systems in health care should be explainable, and there are very active research programs aimed at developing 'explainable AI' on the basis of the current AI architectures (see more below). A requirement for explainability is mentioned in many guidelines and policy documents, e.g. in the recent WHO guidance on AI in health.(World Health Organisation (WHO), 2021) A demand for explainability has also begun to appear in legislation and regulation especially in Europe. The European Union's General Data Protection Regulation Article 13(3)f requires the data controller to disclose "the existence of automated decision-making, including profiling referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, …".(General Data Protection Regulation, 2016) The requirement of provision of information about 'the logic involved' may charitably be interpreted to involve not only a general account, e.g. 'we process your information using a machine-learning support vector machine', but an individualised explanation of the processing of the data of the data subject. The EU Artificial Intelligence Act goes further and requires in Article 13 that "High-risk AI systems [which would include most or all AI systems in health care] shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately" whereas Recital 59 concerned with AI systems in law enforcement calls for such systems to be "explainable", and requires them to be defined as high-risk if they are not.(Artificial Intelligence Act (Regulation (EU) 2024/1689))

In the US the White House Office of Science and Technology Policy has issued a non-legally binding white paper entitled 'Blueprint for an AI Bill of Rights' that *inter alia* states that:

> "Automated systems should provide explanations that are technically valid, meaningful and useful to you and to any operators or others who need to understand the system, and calibrated to the level of risk based on the context."(White House Office of Science and Technology, 2022)

But is explainability the most important desideratum in a health AI system, alongside the (hopefully) *sine qua non* of precision and accuracy at a level equal or superior to that of a trained health care professional?

In this paper we will first explicate the general concept of 'an explanation' and show that because there are many kinds of explanation there must be many kinds of 'explainability' and a significant risk of equivocation when using these terms. We will analyse some of the kinds of explanation we are able to give of AI systems and their output using current explainability methods; and then discuss what types of explanation patients are likely to seek as part of the diagnostic process or as part of choice of therapy. Comparing the kind of explainability we can obtain in relation to AI systems with the explanations patients want and need will show that AI explainability falls short of meeting the reasonable requirements of patients. Based on previous work we will argue that some requests for explanation are better understood as examples of contestation of AI pronouncements on behalf of patients, and that answering such contestations may involve, but is not limited to explainability.

WHAT IS AN EXPLANATION?

When a philosopher hears the term 'explanation' the first example that springs to mind is probably Hempelian nomological-deductive or 'covering law' explanation in the philosophy of science where an observation is fully explainable if it can be deduced from a covering natural law. But explanation is a protean concept and there are many other types of explanation. From philosophy of science we also get the concept of a probabilistic explanation of an observation of a phenomenon, and from the philosophy of action a number of different explanatory schemata for explaining actions (e.g. in terms of goals, intentions, motives, compulsions etc.). In addition we have semantic explanations of the meanings of words and phrases, practical explanations of the functions or working of things, social explanations of the meaning or conventionality of social practices, justificatory explanations for apparent transgressions, existential explanations for the wonders and woes of being and many more. In each case there is an explanandum, something to be explained, and an explanans the thing (very broadly conceived) that explains the explanandum. Although all of these explanation schemata are therefore structurally similar, the character of the explananda and the appropriate explanantia, and therefore what counts as a good explanation, vary widely between cases. It is important to note that something can be a good explanation even if there are deeper and more basic explanations available. The explanation 'probably because you have 60 pack-years of smoking' is a perfectly good and adequate

explanation in response to the question 'doctor, why have I got lung cancer?' even though there is a more basic causal explanation in terms of carcinogens and mutational events in lung cells.

In a paper published in 1962 Passmore suggests that despite the wide variety of types of explanation there is nevertheless something that unites all proper explanations, and that is a lack of some specific knowledge and a sense of puzzlement or unfamiliarity in relation to that lack. To quote Passmore:

> "Everything depends, then, on what I know and what I want to know. If I am asked by an adult human being why Jones died, it will be no explanation to reply: 'All men are mortal', for so much, it can be presumed, he knows already; that is not the unfamiliar feature of the situation that is bothering him."(Passmore, 1962)

Passmore's analysis may be rephrased in terms of epistemic interests. Essentially his claim is that underlying all different types of explanations is a specific epistemic interest, i.e. an interest in acquiring some specific knowledge, and this interest determines what counts as a relevant explanation. The adult's interest in knowing why Jones died is an interest that renders general information on the mortality of all men irrelevant and picks out some set of specific information on Jones' death as relevant.

Because of the very large, possibly infinite number of different explanation types there is a risk of equivocation when talking about explanations, especially in relation to whether something is an explanation or whether an explanation has been provided. If a person asks for an explanation of a certain type and is given an explanation of another type it is at the same time true that an explanation has been given and that no explanation has been given to the person who asked for one since they are none the wiser in relation to their original query. To put it in terms of epistemic interests, the relevant person's epistemic interest has not been satisfied. This may happen even if the person offering the explanation does so in good faith intending to properly answer the request for an explanation.

In medicine this probably most often happens when a medical professional provides a technical explanation in relation to a question which was not technical or only partly technical. A patient having a contrast-enhanced ultrasound scan could potentially ask for an explanation along the lines of 'Why was contrast being used for my scan?'. In response a health care professional could offer an explanation something like 'the contrast more strongly reflects sound waves, and this can be seen on the scans as lighter areas. Since the contrast is in your blood the scan therefore reveals the bloodflow through your body and organs. Cancer and other kinds of diseased tissue usually have higher blood flow and therefore look lighter on the ultrasound image.' While such an explanation certainly answers the question posed by the patient, it is unlikely to satisfy the patient's potential, underlying interest in knowing 1) why it was used in his or her particular case, e.g. if there is a suspicion of a specific disease requiring contrast-enhanced ultrasound scans, 2) what are the pros of using contrast, e.g. that it increases the accuracy of the diagnostics, and 3) what are the cons of using contrast, e.g. what are the potential side-effects. The trained and experienced health care professional may obviously offer explanations that to a large extent accommodate the patients' interest. The example shows, however, that in medicine as elsewhere explanations can be provided that answers a specific request for an explanation but does not satisfy an underlying epistemic interest.

It follows from this analysis of the concept of explanation illustrating its protean nature that there will also be a large number of types of 'explainability' each referring to a specific type of explanation that is provided in response to a specific type of explainability question.

WHAT TYPE OF EXPLAINABILITY IS AI EXPLAINABILITY?

Because the current machine-learning AI structures are fully algorithmic and deterministic once learning mode is switched off and they are locked one type of explanation that could be provided would, if we take a neural network architecture as an example be a full mapping of input to neurons in the first layer with the weightings these neurons assigned to the input and transmitted to neurons in the second layer and up through the layers until we reached the output layer and the weightings of the neurons in that layer that generated the readable output, e.g. that there is a 77% chance that the patient has condition X, 10% chance that they have Y, and 13% chance that they have a condition that the network cannot classify. This would be a complete explanation, but it would not be helpful except in exceptionally simple cases, e.g. cases where the network had picked up on a patognomic feature of a condition (in which case there would be 100% certainty in the output), or where a very specific genetic variant made most of the difference. It would not be helpful because the weightings of the neurons would be uninterpretable, partly because of the amount of data, partly because many of them would have no readily available human interpretation.

The work in the research area of explainable AI is therefore aimed at generating simplified mappings of the 'reasoning processes'[1] employed by the AI

1 Reasoning processes in scare quotes because current(?) AI architectures do not have any mental representations of any reasoning processes.

system in calculating the output from the input. Most of the currently developed and widely implemented AI interpreters provide explanations of the output-input mapping at the system level, and not in relation to an individual case. Because the explanations are simplifications they may obscure important features of the actual workings of the system.(Ghassemi et al., 2021)

The probably most used method is the Local Interpretable Model – Agnostic Explanations – LIME method which can be applied to any type of machine-learning architecture. It works by performing multi-feature perturbations around a specific output and uses the result to fit a linear model incorporating feature importance and feature interactions. (PricewaterhouseCoopers (PwC), 2018; Ribeiro et al., 2016) A LIME model of a diagnostic system would for instance be able to provide an explanation of what features of the input, and what feature interactions are important when the system output is a diagnosis of a particular condition/disease.

More specific methods are available for particular architectures that can provide more accurate information about how the AI model uses input features to perform the classification task. In relation to Deep Neural Network (DNN) it is for instance possible to apply Relevance Propagation methods. A relevance propagation algorithm starts at the output and works backwards through the layers of the neural network, and assigns relevance scores to the inputs received from the preceding layer relative to a 'neutral' activation state of the network.(Shrikumar et al., 2017) These relevance scores provides relatively accurate explanations of the workings of the DNN, but may not always be readily explainable in human terms.

In relation to AI analysis of medical images a range of model specific methods have been used to generate 'heat maps' or saliency maps that visually indicate which areas of the image have been most important for the AI algorithm in classifying the image.(Rajpurkar et al., 2017; Selvaraju et al., 2017) This is supposed to aid the human interpreter, but may in certain circumstances be misleading because the heat map does not indicate which feature in the highlighted area that was important. (Ghassemi et al., 2021)

It follows from the above that the explanation that is provided by AI explainability methods is most often an account of what features in the input data are relevant for the classification task the AI system is trained to perform, and the weight the features have in the model. These explanations are general in the sense that they provide information about what input features are used when the system classifies something as X, or when it distinguishes between X and Y. They are not explanations of why the output was X, or X and not Y in a particular instance.

Here it is important to note a specific difference between AI systems and human thinking and clinical reasoning and decision-making. Most AI systems perform their classification tasks bottom up, i.e. from the data points in the input data to a classification. They are not contrastive. They do not explicitly compare different possible outcomes. The possible AI explanation of 'why X and not Y' will therefore differ in structure from the typical human explanation of the same question. The human diagnostic process may be seen as involving essentially an inference to the best possible explanation, i.e. an inference from a set of signs, symptoms and indicators to a conclusion taken to constitute the best possible explanation of these.(Dragulinescu, 2016; Lipton, 2017) Making an inference to the best possible explanation necessarily involves a consideration of whether there are alternative possible explanations that better explain the relevant set of signs, symptoms and indicators. Thus, diagnosis based on inference to best possible explanation specifically answers the question of 'why X and not Y'..

In the clinical context the AI explainability explanation will then in most cases be translated by the health care professional to an explanation the professional think is understandable by the patient[2]. This will in many cases involve a further simplification of the explanation. The explanation the patient receives will therefore often be doubly simplified. First by the AI explainability method employed, and then by the health care professional.

WHAT EXPLANATIONS DO PATIENTS WANT?

What types of explanations are patients likely to want in a diagnostic or treatment choice context when the health care professional has used an advisory AI system? Or to put it differently, what requests for explanation are they likely to raise?

Let us first note that because of the still existing general trust in health care professionals many and

2 We here assume that the professional has sufficient understanding of the explanation provided by the explainability method to be able to translate it and communicate it to the patient. This assumption is probably questionable in many instances but may become more realistic when more AI systems are implemented in health care and health care professionals become more acquainted with AI explainability information. There may also be a specific need to develop AI – Health care professional interfaces that support the understanding and translation of explainability information in the clinical context.(Amann et al., 2020; Holzinger & Müller, 2021)

perhaps even most patients will not put forward any requests for explanation but will accept the diagnosis and / or the treatment option put forward by the health care professional.(Nuffield Trust, 2022) However, some patients will require explanations..

Some of the likely requests for explanations are unrelated to the use of AI and may even lie outside the proper scope of the expertise of health care professionals. The professional will, for example have no proper answer or explanation to offer in response to the question 'Why have I got cancer?' when this question is raised in an existential mode.[3] If we concentrate on requests for explanation more closely linked to the diagnostic or therapy choice process and the involvement of AI in that process likely and very reasonable questions can include (questions here only enumerated in relation to diagnosis but similar questions could be asked in relation to therapy choice):

* Why do you think I have X?[4]
Does all the evidence point towards me having X?
What alternatives did you consider?
What did the AI advice?
* Why did you follow / not follow that advice?
How good is the AI?
Where does the AI get its information about me from, and what information does it use?
I have heard that some AI systems are racially biased, what do you know about this one?
Who has developed this AI, are there any conflicts of interest?
* I am surprised by the result, I thought it was much more likely to be Y? My neighbour has the same symptoms and her doctor told her it was Y.
* Can you provide an explanation of the AI advice?

There are also some relevant questions a patient could ask, but which require some knowledge about machine-learning, for instance:

What are the characteristics of the learning set of data that was used?
Has the AI performance been validated using a completely separate set of test data? Has it been validated against a data set derived from my/our population?

Some of these questions are straightforward requests for an explanation (marked with * above), but the type of explanation that is an appropriate response differ between them, and some may require both an explanation of the professional's thought processes and an explanation of the function of the AI in relation to the specific patient. Some of the other questions may, depending on how they are answered generate requests for explanations, e.g. the question about what information the AI uses may generate several distinct requests for explanation if the answer is 'your patient records, tax and credit card records and your social media data'.

What is the puzzlement or the epistemic interest that lies behind these requests for explanation to use Passmore's analysis? The range of possible epistemic interests is vast, but we will here focus on three distinct interests that are likely to be present when a patient is given a non-trivial diagnosis. One type of puzzlement or epistemic interest concerns the certainty of a diagnosis. The patient is about to make a significant medical decision and wants to have a high degree of epistemic certainty in relation to a major premise for this decision, i.e. the diagnosis. The patient wants to have a better understanding of what the epistemic warrant is for the claim that they suffer from X, or suffer from X and not Y. When AI advice has been involved in the diagnostic process and a patient raises a question about the certainty of the diagnosis, the answer will have to contain both an explanation of the accuracy of the advice, and an explanation of why the health care professional chose either to rely on the advice or not rely on the advice. A related type of puzzlement or epistemic interest could arise from surprise, i.e. the patient did not expect the diagnosis that is presented, but another diagnosis or a bill of clean health, or is surprised by the treatment options that are offered. The patient therefore asks for and need an explanation of why the situation is different from the one they expected. In this case the health care professional must again be able to provide an explanation of the accuracy of the AI advice and the role it played in their diagnostic decision-making, but must also be able to provide a 'contrastive explanation', i.e. an explanation of why the options or possibilities that the patient had in mind initially, before getting the diagnosis are not warranted by the evidence. Providing the contrastive explanation may be more difficult when AI is involved, because even AI systems that provide some information about why they classified the patient in a certain way, often do not provide any explanation of why other possibilities were discarded. A third type of puzzlement can be generated by suspicion and the belief that there is something wrong about the process, the way the outcome is generated, or the outcome itself. This suspicion can be pre-existing, or it can be generated by surprise in relation to the specific outcome. Here the health care professional must be able to explain why the process is robust and leads to reliable outcomes, but will often need to inquire further into the patient's specific concerns before providing this explanation. For example, a patient who is suspicious of commercial involvement in AI development has a different interest and needs a

3 Unless perhaps in the rare situation where professional and patient know each other well and belong to the same faith community.

4 This can be two different questions depending on whether the stress is on 'I' or on 'X' and would therefore call for two different explanations depending on which question is asked.

different explanation than the patient who is worried that health care professionals have not been properly trained to use AI systems in the right way..

In relation to all three types of puzzlement the kind of explanation that can be provided through explainable AI can often form part of the explanation, but as we have illustrated above more will also often be needed to adequately respond to all of the patient's epistemic interests e.g. accounts of the interaction between the AI and the professional or information about the AI system which is not about its inner workings but about how it was developed, what interests it potentially embeds, if and how it has been tested for bias across different groups etc.. The explainable AI explanation may also in itself not fully satisfy the patient's request for an explanation. Some patients will be satisfied by being told that a, b, c … z are the features the AI system uses to classify a patient as having X or needing treatment T, but some will want an explanation for why the system has classified them as having X or needing T. This request for an individual explanation may also be generated when a general explanation is provided and the patient notices that there seems to be a mismatch between that explanation and what they know about their own case, e.g. the explanation of the AI model states that feature f is important for classifying a patient as having X, but the patient knows that they have never had the investigation that provides the data for f. The honest health care professional will also have to acknowledge that the explanation of the AI advice they are providing and interpreting for the patient is a simplification and may not be an accurate reflection of the workings of the system.

In relation to the third type of puzzlement it is arguable that what the patient wants and needs is not merely an explanation but an effective ability to contest the outcome, the diagnosis or the treatment options they are presented with. The patient not only does not understand the outcome, they also think that it is wrong in some way, and they think that if it forms the basis for medical decisions they may be harmed or wronged. They thus have a strong interest in being able to effectively contest the outcome in order to protect against what they see as a potential threat to their health and welfare.(Ploug & Holm, 2020) Analysing the precise requirements of effective contestation of AI advice is beyond the scope of this paper, but they will go beyond explainability and may, in some cases involve a right to a second opinion. (Ploug & Holm, 2020, 2021, 2023)

CONCLUSION

In this paper we have argued that the types of explanation that is provided by current AI explainability methods do not adequately answer many reasonable requests for explanation that patients can make when their diagnosis or treatment choice has involved the use of AI advice. In relation to some requests these types of explanation are only part explanations, in other cases they answer the wrong question, and in yet other cases they are not explanations at all.

It is therefore important that health care professionals realise that patient queries in relation to AI involvement in diagnosis and treatment choice are not always, and perhaps even not in most cases requests for the kind of explanation that AI explainability methods can deliver. It is therefore the role of the health care professional when such queries are raised to inquire into and discern what it is that really puzzles the patient, and tailor the explanation that is provided to the patient's epistemic interests. The need to be able to provide a range of different explanations has implications for how we should train health care professionals to understand and use AI systems. Such training should not focus exclusively technical aspects, such as understanding how AI architectures work or what explainable AI can deliver, but should include the wider contexts of patients' epistemic interests and reasonable requests for relevant explanations.

For the patient who wants to contest an AI outcome or the professionals' adoption of that outcome as good advice their basis for contestation will only rarely be affected by the availability of an explanation generated by an explainability method. That basis is most commonly the belief that there is something questionable about the diagnosis or the treatment choice in the individual case, and the pieces of knowledge that have the potential to dispel that particular kind of puzzlement are often not about the internal workings of the AI system but about quite different things.

REFERENCES

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & the Precise4Q consortium. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

Bonkhoff, A. K., & Grefkes, C. (2022). Precision medicine in stroke: Towards personalized outcome predictions using artificial intelligence. *Brain*, *145*(2), 457–475. https://doi.org/10.1093/brain/awab439

Dragulinescu, S. (2016). Inference to the best explanation and mechanisms in medicine. *Theoretical Medicine and Bioethics*, *37*(3), 211–232. https://doi.org/10.1007/s11017-016-9365-9

Ferdous, M., Debnath, J., & Chakraborty, N. R. (2020). Machine Learning Algorithms in Healthcare: A Literature

Survey. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/ICCCNT49239.2020.9225642

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Holzinger, A., & Müller, H. (2021). Toward Human–AI Interfaces to Support Explainability and Causability in Medical AI. *Computer*, *54*(10), 78–86. Computer. https://doi.org/10.1109/MC.2021.3092610

Lipton, P. (2017). Inference to the Best Explanation. In *A Companion to the Philosophy of Science* (pp. 184–193). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405164481.ch29

Nirvik, P., & Kertai, M. D. (2022). Future of Perioperative Precision Medicine: Integration of Molecular Science, Dynamic Health Care Informatics, and Implementation of Predictive Pathways in Real Time. *Anesthesia & Analgesia*, *134*(5), 900. https://doi.org/10.1213/ANE.0000000000005966

Nuffield Trust. (2022). *Patient experience: Do patients have confidence and trust in clinicians?* Nuffield Trust. https://www.nuffieldtrust.org.uk/resource/confidence-and-trust-in-clinicians

Passmore, J. (1962). Explanation in Everyday Life, in Science, and in History. *History and Theory*, *2*(2), 105–123. https://doi.org/10.2307/2504458

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics—A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine*, *107*, 101901. https://doi.org/10.1016/j.artmed.2020.101901

Ploug, T., & Holm, S. (2021). Right to Contest AI Diagnostics: Defining Transparency and Explainability Requirements from a Patient's Perspective. In N. Lidströmer & H. Ashrafian (Eds.), *Artificial Intelligence in Medicine* (pp. 1–12). Springer International Publishing. https://doi.org/10.1007/978-3-030-58080-3_267-1

Ploug, T., & Holm, S. (2023). The right to a second opinion on Artificial Intelligence diagnosis—Remedying the inadequacy of a risk-based regulation. *Bioethics*, *37*(3), 303–311. https://doi.org/10.1111/bioe.13124

PricewaterhouseCoopers (PwC). (2018). *Explainable AI – Driving business value through greater understanding*.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, *28*(1), Article 1. https://doi.org/10.1038/s41591-021-01614-0

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv:1711.05225 [Cs, Stat]*. http://arxiv.org/abs/1711.05225

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (2016).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-Agnostic Interpretability of Machine Learning* (arXiv:1606.05386). arXiv. https://doi.org/10.48550/arXiv.1606.05386

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. *Proceedings of the 34th International Conference on Machine Learning*, 3145–3153. https://proceedings.mlr.press/v70/shrikumar17a.html

Subramanian, M., Wojtusciszyn, A., Favre, L., Boughorbel, S., Shan, J., Letaief, K. B., Pitteloud, N., & Chouchane, L. (2020). Precision medicine in the era of artificial intelligence: Implications in chronic disease management. *Journal of Translational Medicine*, *18*(1), 472. https://doi.org/10.1186/s12967-020-02658-5

Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Hachette UK.

White House Office of Science and Technology. (2022). *Blueprint for an AI Bill of Rights – Making Automated Systems Work for the American People.* The White House. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

World Health Organisation (WHO). (2021). *Ethics and governance of artificial intelligence for health*. https://www.who.int/publications-detail-redirect/9789240029200

ACKNOWLEDGEMENTS

This research has been supported by a grant from The Independent Research Fund Denmark (10.46540/2027-00140B), a grant from The Poul Due Jensen/Grundfos Foundation and a grant from the Willum Foundation.

We also acknowledge Klitgården Refugium where the first draft of this paper was started.

COMPETING INTERESTS

The authors have no competing interest to declare.