The false hope of current approaches to explainable artificial intelligence in health care

Marzyeh Ghassemi, Luke Oakden-Rayner, Andrew L Beam

The black-box nature of current artificial intelligence (AI) has caused some to question whether AI must be explainable to be used in high-stakes scenarios such as medicine. It has been argued that explainable AI will engender trust with the health-care workforce, provide transparency into the AI decision making process, and potentially mitigate various kinds of bias. In this Viewpoint, we argue that this argument represents a false hope for explainable AI and that current explainability methods are unlikely to achieve these goals for patient-level decision support. We provide an overview of current explainability techniques and highlight how various failure cases can cause problems for decision making for individual patients. In the absence of suitable explainability methods, we advocate for rigorous internal and external validation of AI models as a more direct means of achieving the goals often associated with explainability, and we caution against having explainability be a requirement for clinically deployed models.

Lancet Digit Health 2021; 3: e745–50

Department of Electrical Engineering and Computer Science and Institute for Medical and Evaluative Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA (M Ghassemi PhD)**; Vector Institute, Toronto, ON, Canada** (M Ghassemi)**; Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia** (L Oakden-Rayner)**; CAUSALab and Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA, USA** (A L Beam PhD)**; Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA** (A L Beam)

Correspondence to:
Dr Andrew L Beam, Department of Epidemiology, Harvard T H Chan School of Public Health, Boston, MA 02115, USA
andrew_beam@hms.harvard.edu

Introduction

Artificial intelligence (AI), powered by advances in machine learning, has made substantial progress across many areas of medicine in the past decade.[1–5] Given the increasing ubiquity of AI techniques, a new challenge for medical AI is its so-called black-box nature, with decisions that seem opaque and inscrutable. In response to the uneasiness of working with black boxes, there is a growing chorus of clinicians, lawmakers, and researchers calling for explainable AI models for high-risk areas such as health care.[6,7]

Although precise technical definitions of explainability lack consensus,[8,9] many high-level, less precise definitions have been put forth by various stakeholders. For example, the General Data Protection Regulation laws in the EU state that all people have the right to "meaningful information about the logic behind automated decisions using their data".[10,11] Similar discussions have taken place in the clinical literature, in which it has been argued that clinicians might feel uncomfortable with black-box AI,[12] leading to recommendations[13] that AI should be explainable in a way that clinical users can understand. Indeed, Tonekaboni and colleagues report that surveyed clinicians "viewed explainability as a means of justifying their clinical decision-making".[14]

We believe that the desire to engender trust through current explainability approaches represents a false hope: that individual users or those affected by AI will be able to judge the quality of an AI decision by reviewing a local explanation (that is, an explanation specific to that individual decision[8]). These stakeholders might have misunderstood the capabilities of contemporary explainability techniques—they can produce broad descriptions of how the AI system works in a general sense but, for individual decisions, the explanations are unreliable or, in some instances, only offer superficial levels of explanation. In practice, explanations can be extremely useful when applied to global AI processes, such as model development, knowledge discovery, and audit, but they are rarely informative with respect to individual decisions.

As such, we suggest that end users of explainable AI, including clinicians, lawmakers, and regulators, be aware of the limitations of explainable AI as it currently exists, especially as it relates to policy, use, and reporting. We argue that if the desire is to ensure that AI systems can operate safely and reliably, the focus should be on rigorous and thorough validation procedures.

Current approaches to explainable AI

Attempts to produce human-comprehensible explanations for machine learning decisions have typically been divided into two categories: inherent explainability and post-hoc explainability.

For machine learning models for which the input data are of limited complexity and clearly understandable, quantifying the relationships between these simple inputs and the outputs of the model is termed inherent explainability. An example of this would be in a linear regression model, where a coefficient measures the strength and direction of the relationship between the weight of a car and the fuel efficiency. The coefficient itself characterises the decision in an understandable way by describing how much each additional kilogram reduces fuel efficiency on average.

The intuitive simplicity of inherently explainable models is appealing, but even these explanations are hampered by the presence of unrecognised confounders. Work in the human–computer interaction community has identified that increased transparency can hamper users' ability to detect sizable model errors and correct for them, "seemingly due to information overload",[15] even for clear-box or inherently explainable models. Further work has found that even data scientists "over-trust and misuse interpretability tools" and that few such experts were able to accurately describe visualisations output by interpretability tools.[16]

In contrast to inherently explainable models, in many modern AI use cases, the data and models are too complex and high-dimensional to be easily understood; they cannot be explained by a simple relationship

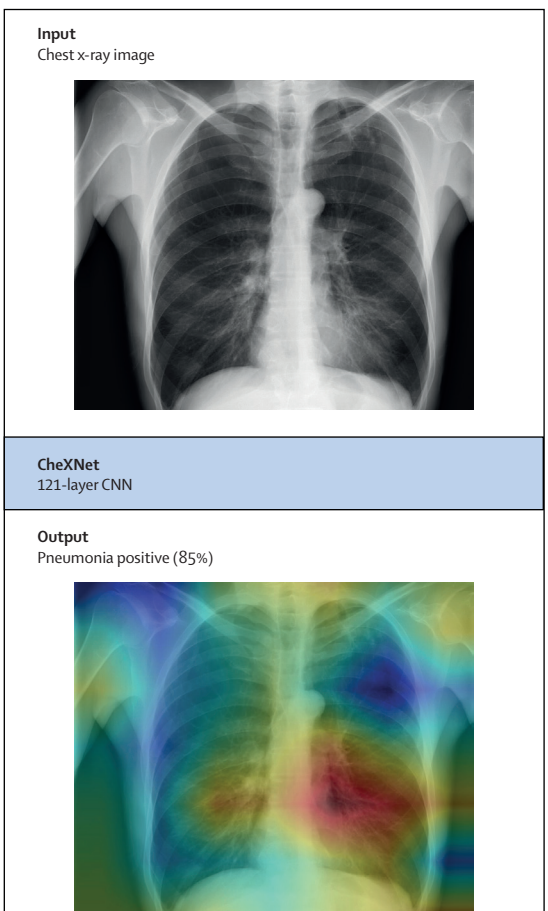

***Figure 1*: Heat map produced by a post-hoc explanation method for a deep learning model designed to detect pneumonia in chest x-rays**
Brighter colours (red) indicate regions with higher levels of importance according to the deep neural network, and darker colours (blue) indicate regions with lower levels of importance. Reproduced with permission from Rajpurkar et al.[21] CNN=convolutional neural network.

See Online for appendix

between inputs and outputs. Examples include models designed to analyse images, text, and sound data. In these scenarios, the focus has been on attempting to dissect the model's decision making procedure, a process called post-hoc explainability. To show post-hoc explainability, we use medical imaging as an illustrative example and explore the most commonly used form of post-hoc explainability in this setting: heat maps. Heat maps (or saliency maps)[17–19] highlight how much each region of the image contributed to a given decision and are illustrative because they provide a simple means of understanding some of the limitations of post-hoc explainability techniques. Although they are popular for medical imaging models, they are well known to be problematic in the broader explainability literature.[20]

As an example, the saliency map shown in figure 1, from Rajpurkar and colleagues,[21] highlights the areas of the image deemed most important for the diagnosis of pneumonia. Even the hottest parts of the map contain both useful and non-useful information (from the perspective of a human expert), and simply localising the region does not reveal exactly what it was in that area that the model considered useful. The clinician cannot know if the model appropriately established that the presence of an airspace opacity was important in the decision, if the shapes of the heart border or left pulmonary artery were the deciding factor, or if the model had relied on an inhuman feature, such as a particular pixel value or texture that might have more to do with the image acquisition process than the underlying disease.

This interpretability gap of explainability methods relies on humans to decide what a given explanation might mean. Unfortunately, the human tendency is to ascribe a positive interpretation: we assume that the feature we would find important is the one that was used (this is an example of a famously harmful cognitive error called confirmation bias). This problem is well summarised by computer scientist Cynthia Rudin: "You could have many explanations for what a complex model is doing. Do you just pick the one you 'want' to be correct?".[22] The ability of localisation methods to mislead human users is compellingly demonstrated by Adebayo and colleagues,[20] who show that even untrained networks can produce saliency maps that appear reassuring (appendix). Moreover, Gu and Tresp[23] showed that common visual explanations remain unchanged even when precise modifications are made to the input that substantially alter the model's predictions (a process known as an adversarial attack), even when those attacks lead to incorrect model predictions (figure 2). It is hard to credit the explanatory ability of a technique that appears believable even when the model is wrong or even completely untrained.

The interpretability gap exists beyond imaging as well. As an example, we see similar problems with contextual language models such as SciBERT,[24] trained on seemingly innocuous sources such as PubMed, which have been shown to have deeply problematic associations about gender and race.[25] Although explanations for language tend to revolve around highlighting the words in the text that contributed to the decision, this does not reveal the associative meaning the model has learned for those words. As with heat maps, the human tendency is to assume that a model has used words in the same way we would. However, deeper investigation often reveals that these models rely on unacceptable shortcuts, such as strongly associating the word doctor with maleness and using this reductionist interpretation to inform decision making.

Beyond heat maps, many other approaches have been developed to produce explanations in complex medical data, including methods such as feature visualisation and prototypical comparisons. Feature visualisation involves the production of synthetic inputs that most strongly activate specific parts of a machine learning

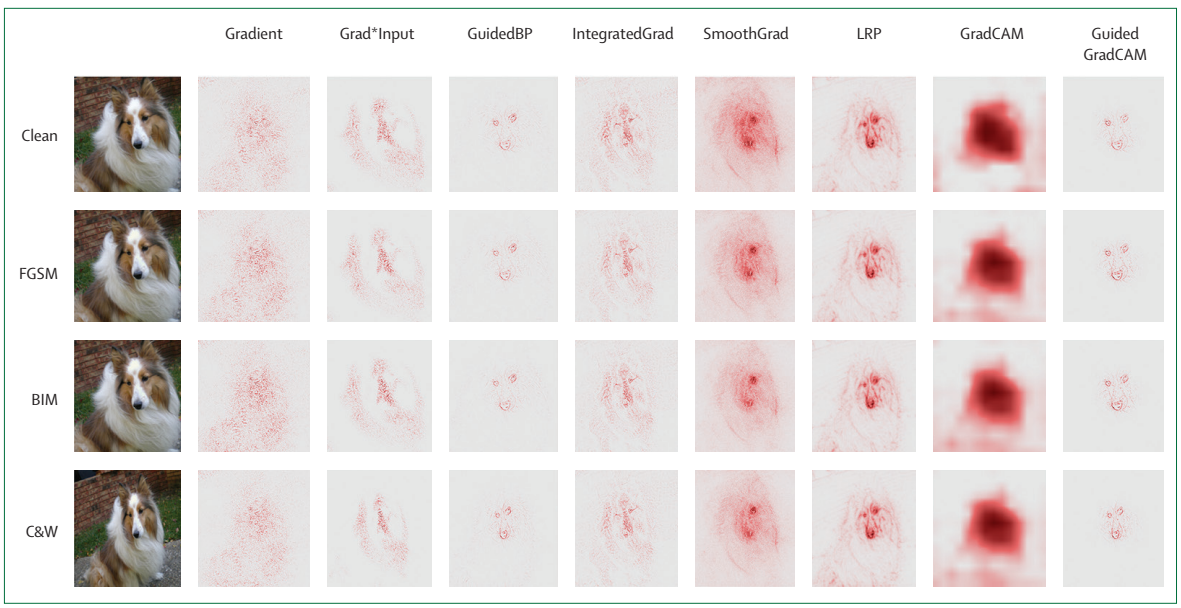

***Figure 2*: Saliency maps produced by popular methods**
Each column shows a different type of explainability method that highlights the most relevant pixels in the images below, which, in each row, are subject to different adverserial pertubations. The top row shows the correctly classified image and saliency maps, and rows 2 to 4 show incorrectly classified images after adversarial perturbations. Reproduced with permission from Gu and Tresp.[23] BIM=basic iteractive method. C&W=Carlini & Wagner. FGSM=fast gradient sign method. GradCAM=gradient-weighted class activation mapping. GuidedBP=guided backpropagation. LRP=layerwise relevance propagation.

model.[26] Each model decision can then be described as a combination of a series of features that were detected in the input. In practice, these synthetic inputs rarely correspond exactly to specific human-interpretable features and are subject to the exact same concerns as heat maps: if a synthetic input looks roughly like a feature a human would use to make a decision (for example, a fur-like texture feature in a dog-detecting AI model), a human must still interpret whether this implies the model made a good decision.

These concerns also extend to other well known post-hoc explanation methods such as locally interpretable model-agnostic explanations (LIME)[27] and Shapley values (SHAP).[28] LIME seeks to understand decisions at the individual level by permuting the input example (altering it in minor ways) and identifying which alterations were most likely to change the decision. In the case of image analysis, this is done by occluding parts of the image, the explanation consisting of a heat map that indicates the image components that were most important for the decision. Such explanations suffer from interpretability gaps in the same way as saliency mapping. Methods such as LIME and SHAP are generic and not specific to images and are routinely used on a wide variety of health-care data, including structured data from electronic health-care records[29] and electroencephalogram waveform data.[30]

Prototypical explanations are interesting in that they are generally considered to be a form of inherent explainability. The model is not only trained for the task itself, but to also identify prototypical elements of each class and then quantify how much of each component it identified for the given decision. Examples include comparing the relevant parts of an image (such as the beak and claws of a bird) to a prototype,[31] producing a text-based description of the decision by referencing canonical descriptive features,[32] or identifying a training instance that is most similar to a test instance according to the trained model.[33] This type of learned explanation has only been recently proposed and has yet to be applied broadly, but still requires human interpretation (ie, were the right canonical elements selected? Was the proportion of each element appropriate?).

All of these examples reveal another major challenge: explanations have no performance guarantees. Indeed, the performance of explanations is rarely tested at all, and most tests that are done rely on heuristic measures rather than explicitly scoring the explanation from a human perspective.[34] This is problematic because explanations, such as those shown in figures 1 and 2, are only approximations to the model's decision procedure and therefore do not fully capture how the underlying model will behave. As such, using post-hoc explanations to assess the quality of model decisions adds an additional source of error—not only can the model be right or wrong, but so can the explanation. Rudin takes this further, saying that post-hoc explanations "must be wrong"; that they are by definition not completely faithful to the original model and must be less accurate with respect to the primary task.[35] In this context, should researchers prefer the full, complex model, which, as humans, we cannot understand but has a high, validated

performance or do we seek to modify that performance with an explanation mechanism, potentially resulting in diminished and unvalidated accuracy?

What are explanations for?

These limitations do not render explainability methods useless, but they do challenge the use of these techniques for certain purposes. If we look at the policy positions and user preferences mentioned earlier, or the intuitive expectations that AI is made explainable, we see a desire to generate trust and inform the choices of individual users or the subjects of AI decision making. However, on an individual level, the explanations we can produce for the behaviour of complex AI systems are often confusing or even misleading. Selbst and Barocas[36] state that, although explainability methods can provide some insight into the decision making process of models, they rarely elucidate whether a given decision was sensible or not. Selbst and Barocas distinguish between explainability techniques that provide descriptive accounts of how the model behaved and normative evaluations that can answer whether that behaviour was justified. Although most discussions and policies call for normative evaluations, current techniques are only capable of descriptive accounts and it is our own intuition that often "serves as the unacknowledged bridge" between the two.[36]

In the example of heat maps, the important question for users trying to understand an individual decision is not where the model was looking but instead whether it was reasonable that the model was looking in this region. By conflating these questions and allowing intuition to bridge the gap, there is a serious risk of introducing harmful biases into decision making. There is a great deal of evidence that humans tend to over-trust computer systems,[37–39] and evidence suggests models that use explainability techniques can hamper people's ability to detect when a model makes serious mistakes[15] or unreasonably increase their confidence in an algorithmic decision,[40,41] giving the veneer of authenticity and resulting in decreased vigilance and auditing of such systems.

This tendency is particularly problematic as another goal of explainability is to detect and avoid algorithms biased towards certain populations.[25] Many systematic biases that reflect societal prejudices (eg, discriminatory policies against women and minority ethnic groups) are encoded in the data from which the AI system learns. Left unchecked, an AI system could operationalise these biases on a large scale. It is implied that explainability could allow us to catch discriminatory behaviour more readily. Unfortunately, as outlined above, this possibility is not reflected in the current state of explainability research, and reliance on explanations might even decrease our vigilance for these behaviours.

Rather than seeing explainability techniques as producing valid, local explanations to justify the use of model predictions, it is more realistic to view these methods as global descriptions of how a model functions. If, for example, a clinical diagnostic model appears to perform well in a specific test set but the heat maps show that the model is consistently distracted by regions of the images that cannot logically inform the diagnosis, then this finding can indicate that the test set itself is flawed and that further forensic investigation is required. An example of this use was when explanatory heat maps revealed that an AI model trained to detect skin cancer was focusing more on the surgical skin markings present on the images rather than the skin lesions.[42] Similarly, there have been notable successes in using explainability methods to aid in the discovery of knowledge, for example, when heat maps were used to identify novel features of diabetic retinopathy progression in ophthalmological fundal eye examination[43] and new radiographic features that are predictive of knee pain.[44] In this sense, we can see it is the aggregate behaviour of these explanations that is informative, not the unquantifiable effect a single reassuring or aberrant explanation will have on an individual prediction.

Better and more equitable outcomes

Although explanations cannot provide a normative evaluation of our models, that does not mean we are forced to accept their black-box predictions without scrutiny. As we have argued, it is the aggregated behaviour of the models that can be informative and, as such, the only effective way to justify the decisions of AI systems is thorough, careful, meticulous safety and validation efforts. Instead of requiring local explanations from a complicated AI system, we should advocate for 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.

The medical system is already extremely adept at evaluation and validating various kinds of black-box systems, as many drugs and devices function, in effect, as black boxes. An often cited example is acetaminophen, which, despite having been used for more than a century, has a mechanism of action that remains only partially understood.[45] Despite competing explanations for how acetaminophen works, we know that it is a safe and effective pain medication because it has been extensively validated in numerous randomised controlled trials (RCTs). RCTs have historically been the gold-standard way to evaluate medical interventions, and it should be no different for AI systems. In recognition of this, many RCT reporting guidelines are being updated to incorporate AI-specific recommendations.[46]

RCTs are not the only mechanism used in health technology assessment to ensure safety, efficacy, and equity. As an example, for an investigation into racial bias in a machine learning system,[47] even completely transparent understanding of the algorithm in question did not reveal the racial bias inherent to the model because it was the problem formulation itself that was

flawed. Instead, it was an aggregate analysis of the inputs, outputs, and outcomes associated with the model that identified the bias. In this context, explainability techniques can serve as a valuable tool for analysis and an adjunct to algorithmic audit,[48] for which the appropriate audience for explanations is not the users or subjects of AI, but rather the developers, auditors, and regulators of these systems.

Conclusions

AI will have an extraordinary impact on medicine in the coming decades, and we should do all we can to ensure that this technology is implemented in a way that maximises patient benefit. However, despite its intuitive appeal, explainability for patient-level decision making is unlikely to advance these goals in meaningful ways. Explainability methods cannot yet provide reassurance that an individual decision is correct, increase trust among users, nor justify the acceptance of AI recommendations in clinical practice.

That is not to say that explainability methods have no role in AI safety. These methods are incredibly useful for model troubleshooting and systems audit, both of which can be used to improve model performance or identify common failure modes or biases. Current explainability methods should be seen as tools for developers and auditors to interrogate their models and, unless there are substantial advances in explainable AI, we must treat these systems as black boxes, justified in their use not by just-so rationalisations, but instead by their reliable and experimentally confirmed performance.

Presently, the hope for human-comprehensible explanations for complex, black-box machine learning algorithms that can be used safely for bedside decision making remains an open challenge. In light of this challenge, we strongly recommend that health-care workers exercise appropriate caution when using explanations from an AI system and urge regulators to be judicious in listing explanations among the requirements needed for clinical deployment of AI.

Contributors

All authors contributed equally to the conception, writing, and editing of the Viewpoint, and had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Acknowledgments

The authors wish to thank Pranav Rajpurkar for the permission to reprint figure 1, Jindong Gu for permission to reprint figure 2, and Julius Adebayo for permission to use a modified figure in the appendix. ALB was supported by the National Heart, Lung, and Blood Institute (7K01HL141771-02).

References

1 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25:** 44–56.

2 Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. *arXiv* 2019; published online Dec 5. https://arxiv.org/abs/1806.00388 (preprint).

3 Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018; **2:** 719–31.

4 Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018; **319:** 1317–18.

5 Beam AL, Kohane IS. Translating artificial intelligence into clinical care. *JAMA* 2016; **316:** 2368–69.

6 Gastounioti A, Kontos D. Is it time to get rid of black boxes and cultivate trust in AI? *Radiol Artif Intell* 2020; **2:** e200088.

7 Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol Artif Intell* 2020; **2:** e190043.

8 Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv* 2017; published online Feb 28. http://arxiv.org/abs/1702.08608 (preprint).

9 Lipton ZC. The mythos of model interpretability. *Commun ACM* 2018; **61:** 36–43.

10 European Parliament. 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. *OJEU* 2016; **59:** 294.

11 Miller K. AI decisions: do we deserve an explanation? June 29, 2020. https://www.futurity.org/ai-decisions-right-to-explanation-2394872-2/ (accessed Sept 9, 2021).

12 Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept "black box" medicine? *Ann Intern Med* 2020; **172:** 59–60.

13 Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med* 2020; **3:** 47.

14 Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. *arXiv* 2019; published online May 13. http://arxiv.org/abs/1905.05134 (preprint).

15 Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H. Manipulating and measuring model interpretability. *arXiv* 2021; published online Aug 15. https://arxiv.org/abs/1802.07810 (preprint).

16 Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Vaughan JW. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI conference on human factors in computing systems. New York, NY, USA: Association for Computing Machinery, 2020: 1–14.

17 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. Cambridge, MA, USA: Institute of Electrical and Electronics Engineers, 2017: 618–26.

18 Tulio Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. *arXiv* 2016; published Aug 9. https://arxiv.org/abs/1602.04938 (preprint).

19 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; **30:** 4765–74.

20 Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *Adv Neural Inf Process Syst* 2018; **31:** 9505–15.

21 Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv* 2017; published online Nov 14. http://arxiv.org/abs/1711.05225 (preprint).

22 Bornstein AM. Is artificial intelligence permanently inscrutable? Sept 1, 2016. http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable (accessed Feb 27, 2020).

23 Gu J, Tresp V. Saliency methods for explaining adversarial attacks. *arXiv* 2019; published online Aug 22. http://arxiv.org/abs/1908.08413 (preprint).

24 Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. *arXiv* 2019; published online Sept 10. https://arxiv.org/abs/1903.10676 (preprint).

25 Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. In: Proceedings of the ACM conference on health, inference, and learning. New York, NY, USA: Association for Computing Machinery, 2020: 110–20.

26 Olah C, Satyanarayan A, Johnson I, et al. The building blocks of interpretability. *Distill* 2018; **3:** e10.

27 Biecek P, Burzykowski T. Local interpretable model-agnostic explanations (LIME). In: Explanatory model analysis. New York, NY, USA: Chapman and Hall/CRC, 2021: 107–23.

28 Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society. New York, NY, USA: Association for Computing Machinery, 2020: 180–86.

29 Khedkar S, Gandhi P, Shinde G, Subramanian V. Deep learning and explainable AI in healthcare using EHR. In: Dash S, Acharya BR, Mittal M, Abraham A, Kelemen A, eds. Deep learning techniques for biomedical and health informatics. Cham, Germany: Springer International Publishing, 2020: 129–48.

30 Alsuradi H, Park W, Eid M. Explainable classification of EEG data for an active touch task using Shapley values. In: HCI international 2020—late breaking papers: multimodality and intelligence. Cham, Germany: Springer International Publishing, 2020: 406–16.

31 Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. *Adv Neural Inf Process Syst* 2019; **32:** 8930–41.

32 Gale W, Oakden-Rayner L, Carneiro G, Palmer LJ, Bradley AP. Producing radiologist-quality reports for interpretable deep learning. *arXiv* 2018; published online June 1. https://arxiv.org/abs/1806.00340 (preprint).

33 Schmaltz A, Beam A. Exemplar auditing for multi-label biomedical text classification. *arXiv* 2020; published online April 7. http://arxiv.org/abs/2004.03093 (preprint).

34 Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). Cambridge, MA, USA: Institute for Electrical and Electronics Engineers, 2018: 80–89.

35 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; **1:** 206–15.

36 Selbst AD, Barocas S. The intuitive appeal of explainable machines. *Fordham Law Rev* 2018; **87:** 1085–139.

37 Skitka LJ, Mosier KL, Burdick M. Does automation bias decision-making? *Int J Hum Comput Stud* 1999; **51:** 991–1006.

38 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017; **24:** 423–31.

39 Howard A. Are we trusting AI too much? In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. New York, NY, USA: Association for Computing Machinery, 2020: 1.

40 Ghassemi M, Pushkarna M, Wexler J, Johnson J, Varghese P. ClinicalVis: supporting clinical task-focused design evaluation. *arXiv* 2018; published online Oct 13. http://arxiv.org/abs/1810.05798 (preprint).

41 Eiband M, Buschek D, Kremer A, Hussmann H. The impact of placebic explanations on trust in intelligent systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, 2019: 1–6.

42 Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol* 2019; **155:** 1135–41.

43 Arcadu F, Benmansour F, Maunz A, Willis J, Haskova Z, Prunotto M. Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit Med* 2019; **2:** 92.

44 Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med* 2021; **27:** 136–40.

45 Kirkpatrick P. New clues in the acetaminophen mystery. *Nat Rev Drug Discov* 2005; **11:** 883.

46 Liu X, Faes L, Calvert MJ, Denniston AK. Extension of the CONSORT and SPIRIT statements. *Lancet* 2019; **394:** 1225.

47 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366:** 447–53.

48 Raji ID, Smart A, White RN, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *arXiv* 2020; published online Jan 3. http://arxiv.org/abs/2001.00973 (preprint).