annalise.ai

# Designing Effective Artificial Intelligence Software

## Clinical Poster

# Learning objectives

To raise awareness of the importance of usable AI design, provide examples of model interpretability methods, and to summarise clinician reactions to methods of communicating AI model interpretability in a radiological tool.

# Background

In the past decade, the number of AI-enabled tools, especially deep learning solutions, has exploded onto the radiological scene with the promise of revolutionising healthcare[1]. However, these data-driven models are often treated as numerical exercises and black boxes, offering little insight into the reasons for their behaviour.

Trust in novel technologies is often limited by a lack of understanding of the decision-making processes behind the technology. In medical AI, this problem is twofold - firstly, AI technologies are not widely taught in any medical curriculum so there is limited understanding in practice, and secondly, AI technologies have previously been shown to produce incorrect predictions due to hidden biases in the training data[2][3]. In response to this “black-box” problem in medical AI, there has been a growing call for “explainable” or “interpretable” AI tools to allow more transparency in its thought processes[4][5][6][7][8][9].

Here we present our experiences during the development of the Annalise.ai CXR tool as a commercial product case study, exploring the key steps in the creation of an accurate, user-friendly, and interpretable AI diagnostic tool guided by these principles, with the added benefit of seamless workflow integration. This process requires understanding of the practical requirements for the end user, as well as the software engineering challenges in model development. The onus is on any developer of such tools to organise the AI output in ways that radiologists and other medical practitioners can understand intuitively[10].

# Findings and procedure details

## Design Cycle

"It's just aggravating to have to move and shuffle all these windows... shuffle between the list and your [Brand Name] dictation software... [or] Google Chrome or Internet Explorer, to search for something on there. Everything's just opening on top of each other, which is aggravating."

- UX interview with Interventional Radiologist, USA

The design of the entire user experience of our AI tool has involved radiologists and other clinicians at every step, which has helped generate feedback to ensure that the software is usable in the intended work environment with minimal workflow disruption[11]. The design of our tool is iteratively refined through "rounds" of radiologist feedback involving 4-7 radiologists and shown in Figure 1. Such sessions involve manipulation of the prototype during a structured interview session, and focuses on clinical aspects of the design such as:

- The groupings and names of the 124 clinical findings
- Attitudes to the confidence bar
- Attitudes to the region of interest highlights

The interaction of the widget with work software This design cycle also emphasises interpretability of predictions in recognition of its growing importance. Drawing upon techniques suggested by a growing body of research[12], we explored attitudes to interpretable predictions for clinically important findings in three main ways:

01. Provision of confidence bars
02. Provision of localisation maps
03. Provision of differentials

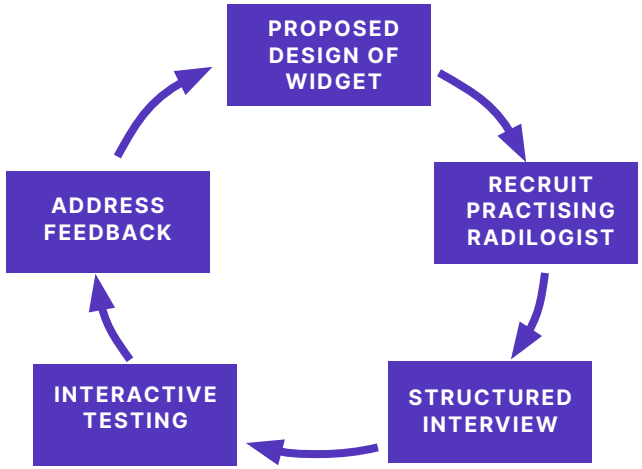


Fig 1: A schematic of the iterative design cycle used in the design of our AI tool....

## Confidence Bar

"The x-rays are sort of black and white, but the actual diagnosis is, you know, it's not black and white and that's one of the criticisms of AI using the ground truth in that you shouldn't re ly on reports because you put the same, x-ray in front of three radiologists, there'll be different opinions."

- UX Interview with Radiologist, Australia

"I like that actually, because again, very few things in medicine are black and white... I really liked that there is a degree of uncertainty built into the system because it means that I can ultimately disagree with or agree with the AI and say, look, I'm not even sure that that thing is present. Neither is the AI."

- UX Interview with Emergency Doctor, Australia

Human radiologists are familiar with soft classifications such as "probable", but often AI tools must artificially binarise predictions into definitely present or definitely absent[13]. Communicating model confidence allows a more interpretable and nuanced approach to the interpretation of a radiograph, where human judgement complements error-prone areas of the AI tool[14].

Our AI tool utilises a confidence bar, which has been refined through multiple rounds into a display of the model prediction relative to the threshold for positivity, and the prediction uncertainty (Figure 2). Figure 3 and 4 both demonstrate the model calling a simple pneumothorax, but Figure 4 is a borderline call, and is ground truth negative for pneumothorax.

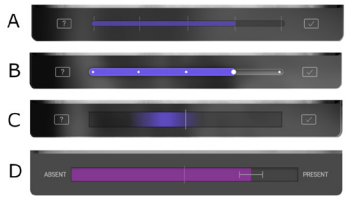


Fig 2: The four main iterations of the integrated confidence bar – A to D being in...

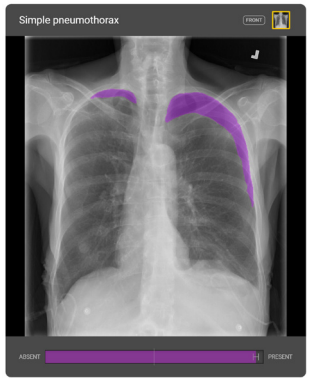


Fig 3: An example of a high confidence finding. A pneumothorax is present with high...

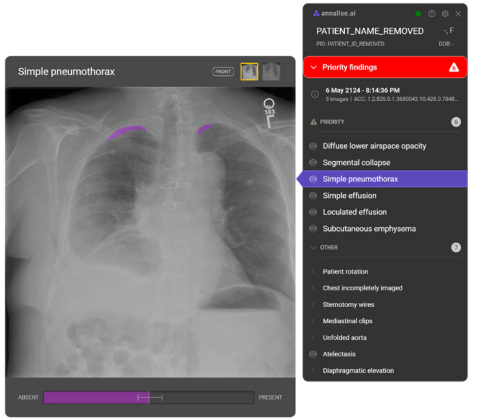


Fig 4: An example of a low confidence finding. A simple pneumothorax is flagged by the...

## Region of Interest (RoI) Map

"So what I like about this is at least it highlights areas to, to be mindful of, or at least your purple blobs there. Anyway,... I think it depicts better what might be an area of concern better than a bunch of words that say the same thing."

- UX interview with Interventional Radiologist, USA

"Yeah, I think it's good. Cause it's like you see the picture with the overlay on top of it, and then you can look at the same picture and see if there are real findings that are not"

- UX interview with Interventional Radiologist, USA

Saliency maps are important for interpretability, providing visual confirmation that an AI is paying attention to the correct region of the image. The classic cautionary tale in medical AI is in melanoma classification, where a photographed ruler or surgical skin markings influenced predictions of malignant skin lesions by a diagnostic AI tool[15][16][17]. Previous work has focused on using saliency maps generated by techniques such as Grad-CAM or Integrated Gradients to query this[18][19][20][21].

Our AI model goes beyond this and provides explicit localisation. The model's Y-net structure means the classification and the RoI map both reflect the model's understanding of the image. This map helps to reassure the operator that the AI model is paying attention to the correct area of the image, and points to the area of the image triggering a finding if it is not obvious. If a RoI were not shown, the operator would instead be forced to scour the image to try and guess the feature that might have triggered the prediction, creating frustration and doubt. Figure 5 provides an example where providing a RoI map aids the clinician in making a decision for a flagged acute rib fracture that was ultimately found to be an old rib fracture.

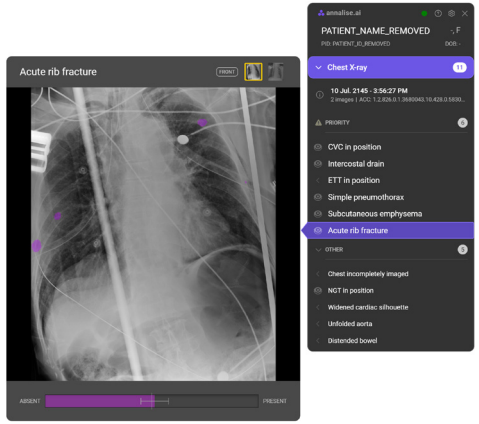


Fig 5: An acute rib fracture is flagged by the model with low confidence. As rib...

## Differentials

"That's probably another, another good one... differential diagnosis. So, you know, there's a finding based on probability, Chucks out say five different differentials. And quite honestly, as a human, I probably would only think of the top two or three. And then I'll look through that list and I'm like, Oh, okay. Maybe the fifth one is a reasonable choice."

- UX Interview with Radiologist, Australia

Providing multiple differentials means the operator can engage with the AI model more organically in a manner that is not simply black and white. This allows the operator to retain the power of decision making in the context of the patient's clinical picture, as well as previous studies and patient history, which the model does not have access to.

Our AI model is deliberately allowed to predict multiple findings for a single radiographic feature, each with different confidences/ probabilities. This behaviour is akin to the organic behaviour of humans in providing differentials for a radiographic finding. For example, an opacity may be labelled simultaneously as a "Focal Airspace Opacity", a "Segmental Collapse", and a "Pulmonary Mass", with each finding having its own confidence. An example of this behaviour can be seen in Figure 6-7, where the clinican is offered two possibilities to consider for an opacity.

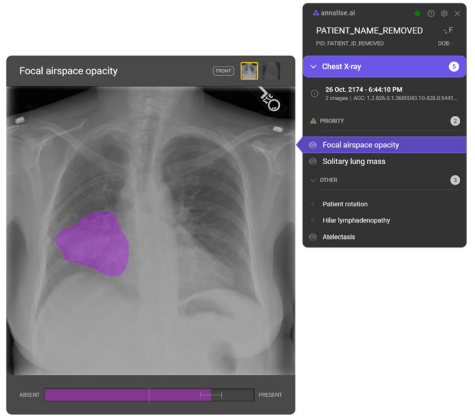


Fig 6: An example of a radiographic opacity with two differentials provided by the AI...

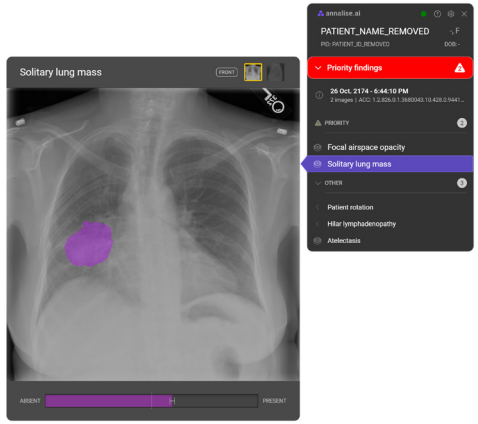


Fig 7: An example of a radiographic opacity with two differentials provided by the AI...

## Conclusion

The inclusion of interpretability techniques has been well-received through testing in multiple rounds of user interviews, reflecting a demand from the broader radiological community to be able to demystify the black box of AI. Future AI work should involve radiologists at all steps of the design process in order to address workflow and UI concerns, especially as regulatory authorities move towards guidelines that aim to ensure a safer and more interpretable AI future.

## Personal information and conflict of interest

C. Tang: Employee: Annalise.ai

J. C. Y. Seah: Employee: Annalise.ai

Q. Buchlak: Employee: Annalise.ai

C. Jones: Employee: Annalise.ai

## References


All images are used with permission from Annalise.AI

All chest radiographs analysed here are from MIMIC-CXR 2.0.0: Johnson, A., Pollard, T., Mark, R., Berkowitz, S., & Horng, S. (2019). MIMIC-CXR Database (version 2.0.0). PhysioNet. https://doi.org/10.13026/C2JT1Q.

[1] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. Springer Science and Business Media LLC; 2018;18(8):500–510 http://dx.doi.org/10.1038/s41568-018-0016-5.

[2] Recht MP, Dewey M, Dreyer K, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. Eur Radiol. Springer Science and Business Media LLC; 2020;30(6):3576–3584 http://dx.doi.org/10.1007/s00330-020-06672-5.

[3] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. Springer Science and Business Media LLC; 2019;17(1) http://dx.doi.org/10.1186/s12916-019-1426-2.

[4] Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? 2017; arXiv:1712.09923v1 [cs.AI]

[5] Geis JR, Brady A, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Insights Imaging. Springer Science and Business Media LLC; 2019;10(1) http://dx.doi.org/10.1186/s13244-019-0785-8.

[6] Baselli G, Codari M, Sardanelli F. Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way? Eur Radiol Exp. Springer Science and Business Media LLC; 2020;4(1) http://dx.doi.org/10.1186/s41747-020-00159-0.

[7] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. Springer Science and Business Media LLC; 2019;17(1) http://dx.doi.org/10.1186/s12916-019-1426-2.

[8] Samek W, Wiegand T, Müller K-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models 2017; arXiv:1708.08296v1 [cs.AI]

[9] Elton DC. Self-explaining AI as an Alternative to Interpretable AI. Artificial General Intelligence. Springer International Publishing; 2020. p. 95–106 http://dx.doi.org/10.1007/978-3-030-52152-3_10.

[10] Zhang Y, Liao QV, Bellamy RKE. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM; 2020. http://dx.doi.org/10.1145/3351095.3372852.

[11] Google. People and AI Guidebook. https://pair.withgoogle.com/. Accessed Feb 2020.

[12] Reyes M, Meier R, Pereira S, et al. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. Radiology: Artificial Intelligence. Radiological Society of North America (RSNA); 2020;2(3):e190043 http://dx.doi.org/10.1148/ryai.2020190043.

[13] Brady AP, Neri E. Artificial Intelligence in Radiology—Ethical Considerations. Diagnostics. MDPI AG; 2020;10(4):231 http://dx.doi.org/10.3390/diagnostics10040231.

[14] Patel BN, Rosenberg L, Willcox G, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. npj Digit Med. Springer Science and Business Media LLC; 2019;2(1) http://dx.doi.org/10.1038/s41746-019-0189-7.

[15] Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated Classification of Skin Lesions: From Pixels to Practice. Journal of Investigative Dermatology. Elsevier BV; 2018;138(10):2108–2110http://dx.doi.org/10.1016/j.jid.2018.06.175.

[16] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. Springer Science and Business Media LLC; 2017;542(7639):115–118 http://dx.doi.org/10.1038/nature21056.

[17] 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. American Medical Association (AMA); 2019;155(10):1135 http://dx.doi.org/10.1001/jamadermatol.2019.1735.

[18] Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. Computer Vision – ECCV 2014. Springer International Publishing; 2014. p. 818–833 http://dx.doi.org/10.1007/978-3-319-10590-1_53.

[19] Philbrick KA, Yoshida K, Inoue D, et al. What Does Deep Learning See? Insights From a Classifier Trained to Predict Contrast Enhancement Phase From CT Images. American Journal of Roentgenology. American Roentgen Ray Society; 2018;211(6):1184–1193 http://dx.doi.org/10.2214/AJR.18.20331.

[20] Singh A, Sengupta S, Lakshminarayanan V. Explainable Deep Learning Models in Medical Image Analysis. J Imaging. MDPI AG; 2020;6(6):52 http://dx.doi.org/10.3390/jimaging6060052.

[21] Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M. D., & Kalpathy-Cramer, J. 2020. Assessing the (Un)trustworthiness of saliency maps for localizing abnormalities in medical imaging. In bioRxiv. https://dx.doi.org/10.1101/2020.07.28.20163899