Expert-Informed, User-Centric Explanations for Machine Learning

Michael Pazzani, Severine Soltani, Robert Kaufman, Samson Qian, and Albert Hsiao
University of California, San Diego
mpazzani@ucsd.edu, ssoltani@ucsd.edu, rokaufma@ucsd.edu, saqian@ucsd.edu, and a3hsiao@ucsd.edu

Abstract

We argue that the dominant approach to explainable AI for explaining image classification, annotating images with heatmaps, provides little value for users unfamiliar with deep learning. We argue that explainable AI for images should produce output like experts produce when communicating with one another, with apprentices, and with novices. We provide an expanded set of goals of explainable AI systems and propose a Turing Test for explainable AI.

Explaining Image Classification

Explaining the decisions of AI has emerged as an important research topic. Considerable progress in image classification using deep learning (Krizhevsky, et al., 2012; LeCun, et al., 2015) has created significant interest in explaining the results of image classification. Although there are many applications for explainable AI (XAI), this paper first focuses on learning to classify images. We then discuss broader implications for explainable AI.

Recent conferences include tutorials and workshops on explainable AI. There are several good surveys of XAI (Chakraborty et al., 2017 & Došilović et al., 2018). This is not one of them. Instead, after working on problems with experts in radiology and ophthalmology and on bird identification, we have concluded that existing techniques leave much room for improvement. The field needs additional directions and methodology, including clarifying XAI's goals, particularly with respect to users, experts, and image classification.

Although some of XAI's original goals were to "explain their decisions and actions to human users" (Gunning & Aha, 2018) the current state-of-the-art is developer-centric rather than user-centric. The dominant method for explaining image classification is assigning an importance score to pixels or regions on a saliency map or heatmap superimposed on an image, visualizing a region's importance with color scales (red, orange, yellow…). Methods developed for creating heatmaps include occlusion sensitivity (Zeiler & Fergus 2014), LIME (Ribeiro, et al., 2016), LRP (Lapuschkin et al., 2016), GradCAM (Selvaraj et al., 2017) and IGOS++ (Khorram et al., 2021). Figure 1 (left) shows heatmaps generated by UCSD researchers for diagnosing glaucoma with IGOS++ (top), identifying bird species with LIME (middle), and diagnosing COVID-19 with GradCAM (bottom). Although heatmaps unquestionably provide useful information to developers (Anders et al., 2022) and perhaps technical auditors (Adebayo et al., 2020), particularly to indicate when the classifier mistakenly focuses on irrelevant regions of images (DeGrave et al., 2020 and Nourani et al., 2019), we argue they do not match what experts naturally produce nor what users expect.

We propose an expanded research agenda that includes:

1. investigating how people, particularly experts, explain their conclusions to others,
2. investigating the preferences of users for different types of explanations, and
3. developing systems that output the types of explanations experts produce and users prefer.

Investigation of how people explain their conclusions draws techniques from ethnography, anthropology, and cognitive science. How do experts communicate their findings, and what artifacts do they use to explain them? Surveys on explanation from psychology (Miller, 2019, Hoffman et al., 2018) and philosophy (Lu et al., 2020) have not emphasized expert interpretation of images.

We argue that heatmaps have several problems:

1. Heatmaps are not typically what experts create when they communicate with others.
2. Heatmaps do not appear to be what users prefer.
3. Despite many approaches to generating heatmaps, alternatives are rarely compared quantitatively or in psychology experiments with actual users.
4. A single deep net and heatmap finds one sufficient way to classify but ignores other regions and features considered important by experts.

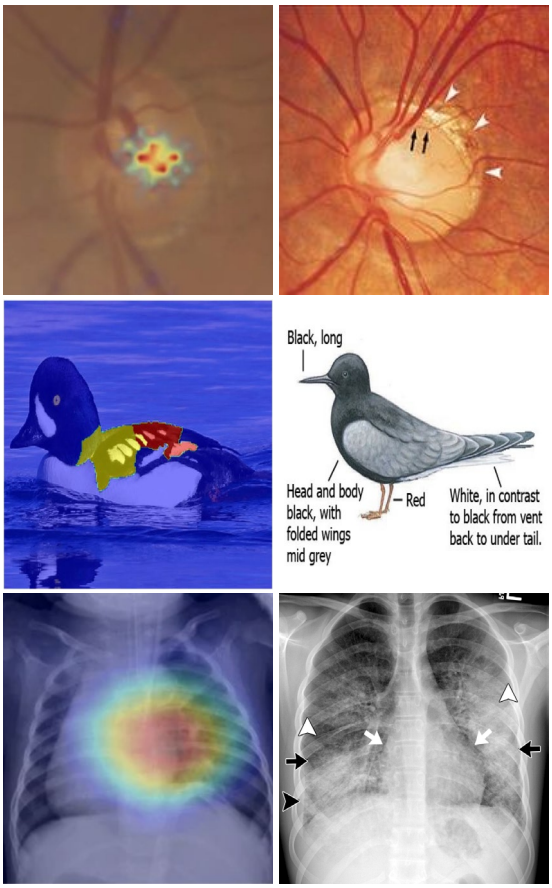

Figure 1. XAI annotates regions of interest (left) compared to expert-created explanations (right).

Experts and Image Classification

In contrast to the left column of Figure 1, the right column shows explanations produced by experts to communicate with others. The left image describes an unusual glaucoma case from Jonas et al. (1998) indicating "parapapillary atrophy (arrowheads) and rim notching (arrows)." The middle image from Morcombe & Stewart (2010) describes important features of a bird, and the right image from Kligerman et al. (2020) addresses four different radiological features as "small right pleural effusion (black arrowhead) and septal thickening (white arrowhead) and subpleural portions of lung (black arrows)."

After several years of reading radiology and ophthalmology journals and bird guides, we have yet to encounter a heatmap used to explain image classification except in the developer-oriented context of describing deep learning. It is as though image explanations in AI have access to a paintbrush or a highlighter but lack arrows and text boxes.

In addition to looking at artifacts, we have interviewed experts during video meetings about bird species identification from photos or diagnosis from radiographs and then analyzed the videos and transcripts. Figure 2 shows screen captures from two sessions.

While experts do indicate regions of interest, in discussion they also describe what it is about each region that makes it informative. For the bird on the left, the description includes "*Looks to be a Bell's Vireo. You can tell it's a vireo by the … stronger and thicker legs than a lot of stuff like a warbler.* [circles legs] … *a thin but slightly thicker bill than a warbler* [circles bill] … *and then Bell's vireo by it doesn't have the bold spectacles here, just kind of some faint spectacles and kind of a broken eye ring* [circles eye area], *kind of weak wing bars* [circles wing area], *a kind of longish tail* [circles tail] … *and just kind of overall plain, gray, gray, whitish with maybe a little bit of greenish tones, but not very bright. I think the easiest confusion would be gray vireo. I think that gray area tends to have more of just a broken eyering… I don't think they have any of the greenish wash to the back or the wings and tail either.*"

For the radiograph on the right, the discussion includes "*There are multiple masses here. It's some sort of metastatic cancer… this is going to be probably an enlarged lymph node* [draws semi-circle]. *The normal contour of the aorta probably is this* [draws vertical line] … *There are the branch pulmonary arteries* [draws horizontal lines] *These are nodules…*[draws polygons]*...Probably metastatic cancer you know they're dense and there are many and varying in size.*"

A complete analysis of these conversations is beyond the scope of this paper, but it is obvious the birding expert is describing the bird at multiple places in the hierarchy and drawing contrast with likely confusions (vireo vs. warbler, Bell's vireo vs. gray vireo). The radiologist is highlighting important features to justify his diagnosis. Most importantly, they are describing what it is about the region of interest that makes it interesting.

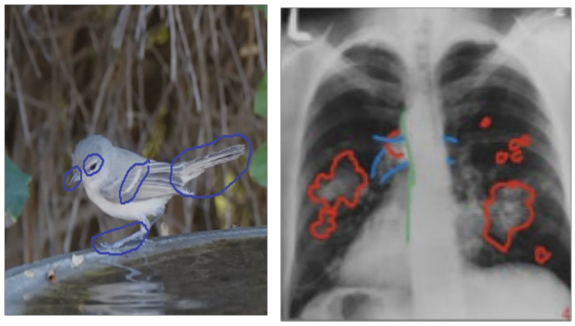

Figure 2. Screen captures from video interviews.

Of course, there has been some work in identifying intermediate concepts in images that can serve as part of an explanation. For example, TCAV (Kim et. al., 2018) can identify whether a deep net has used an intermediate concept but does not identify where in a particular image that concept appears. Concept bottleneck models (Koh, 2020) learn to

recognize whether features are present in images and then use these features for an overall classification without identifying where in the image the features occur. Semantic segmentation (Noh et al., 2015) divides an image into segments but does not indicate how these segments lead to a classification for the image. Similarly, image captioning (Vinyals et al., 2015) identifies objects within an image but does not indicate the location of objects or produce an overall classification of an image.

We now propose two goals that we believe expert-informed user-centric explainable AI should achieve.

> **Explainable AI systems should have the goal of producing explanations like those of experts.**

This naturally leads to using Turing's (1950) imitation game to evaluate explainable AI systems.

> **Explainable AI systems should be evaluated according to whether their explanations are indistinguishable from those of human experts.**

Biessmann & Treu (2021) have proposed a Turing Test for transparency, though it lacks ecological validity. The task was to have people distinguish positive from negative movie reviews. However, rather than allowing people to explain concepts such as sarcasm, subjects had to perform like AI systems, marking three words in the review as most relevant for their decision. Instead of getting people to act like AI systems, we propose to get XAIs to act like human experts.

Users and Image Classification

We now turn our attention to what users want from an explanation. Here we summarize two experiments performed at UCSD. In the first, 21 expert bird watchers were recruited from mailing lists that report rare bird sightings in Southern California. Subjects were shown various annotations, such as heatmaps and labeled arrows, and asked two questions on a 7-point Likert scale ranging from "strongly disagree" to "strongly agree." Feedback was collected on "This explanation emphasizes the areas of the bird that I think are important for identification" and "I would recommend using this explanation to help identify this bird." The heatmap annotations were produced by GradCAM to compare an established XAI algorithm with annotations like those from bird guidebooks. The subjects exhibited a significant preference for labeled arrows (median ratings of 7 for "correct emphasis" and 6.5 for "helpful") over heatmaps (median ratings of 3 and 2). This leads us to our third goal.

> **Explainable AI systems should meet the expectations of users for helpful explanations.**

We distinguish user-centric explainable AI (UCXAI) from developer-centric explainable AI (DCXAI). A further example illustrates the difference between DCXAI and UCXAI. We trained two deep nets on the same data, differing only by the initial random weights. Figure 3 shows the areas each net finds important as shown by the GradCAM heatmaps. Each net has found only one of the two important field marks birdwatchers use to distinguish this bird from similar ones. This is not an issue with GradCAM but rather that deep learnings find one sufficient way of distinguishing classes, not all ways. However, we would argue that both field marks should be reported to users who care more about how to distinguish this bird from similar ones than how a particular neural net operates.

Figure 3. Heatmaps on the same image from two deep nets.

In the second study, subjects were 336 UCSD undergraduate students from psychology, cognitive science, or linguistics courses who were not expert bird watchers. The task was to learn to distinguish two similar bird species such as Western Grebes and Clark's Grebes. Subjects were asked to distinguish three pairs of similar bird species, one pair at a time. Subjects were shown a bird, asked to guess its classification, and then shown the correct classification. One group of subjects received feedback with photos of the correct bird with labeled arrows pointing to its distinguishing features. A second group saw heatmaps that highlighted distinguishing features. These heatmaps were drawn by hand as a best-case scenario and corresponded to the features identified by arrows. A third group saw no explanation, just feedback on the correct class. We measured the number of trials until the subject was able to correctly identify 9 out of 10 photos in a running window of 10 trials before moving to the next bird pair. There was no significant difference in the number of trials taken for this task between the group that received a heatmap explanation and the group that received feedback without an explanation for any of the bird pairs. The labeled arrow explanation emerged as the most useful type of feedback: the median number of trials to complete the task for each bird pair in the group that received a labeled arrow explanation was significantly lower than the number of trials for the groups receiving heatmap or no explanations. The results are summarized in Figure 4.

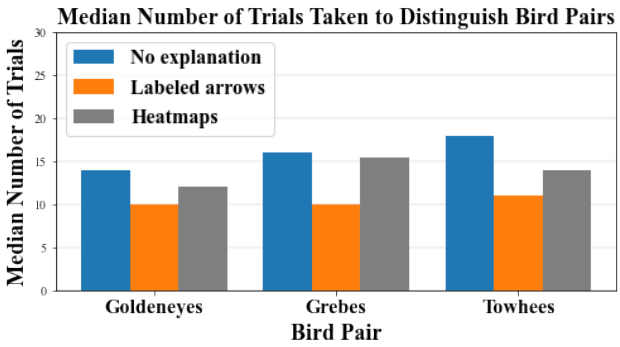

Figure 4. Median number of trials by explanation type.

Our fourth goal for XAI involves both finding regions of interest and describing these regions.

> **Explainable AI systems should be able to locate and identify distinguishing features for users.**

We have begun investigating whether XAI methods can be used to identify and locate features. In addition to the class, we label each image with additional features, such as whether the wing has a solid color, has spots, or has a wingbar. Using multitask learning (Caruana, 1997), we simultaneously provide feedback on the species and features such as wing pattern and bill size. Note that we do not train on the location of the wingbar but let the deep learner determine whether the bird has a wingbar and use existing XAI methods to find regions that are important for determining whether there is a wingbar. Figure 5 (left) shows an example of using GradCAM to visualize the important pixels for wingbars. Figure 5 (right) shows an example using GradCAM to find the most important pixels in determining that the bill is large. Figure 5 is a first step toward labeling features that experts use in explanations and that novices find helpful in learning to classify.

Although we do not present evidence here, we agree with Miller (2018) and Hoffman et al. (2019) that there is not a one-size-fits-all explanation.

> **Explainable AI should adapt explanations to the user's knowledge and experience.**

A special case of this involves novice users.

> **Explainable AI systems should help novices learn to be experts.**

We have taken a step toward this, but much more remains. Again, using multitask learning, we learn separate concepts for the family (e.g., hawk) and the species (red-shouldered hawk). The heatmap (produced by LIME) for why the bird is a hawk (Figure 6 left) focuses on the beak and eye, while the heatmap for why it is a red-shouldered hawk also highlights the wing (Figure 6 right). Of course, this is just a small step toward UCXAI. We assume a novice would want to know why it is a hawk and a more advanced person why it is a red shouldered hawk. Multi-task learning has no knowledge of the user's mental model, nor can it have a dialogue where it determines that the user already knows that the hawk has sharp beak but is not aware that the eyes are close together.

Figure 5. Heatmap for wingbars (left) and long bill (right).

Figure 6. Heatmap for hawk (left) and red-shouldered hawk (right).

Quantitatively Evaluating XAI

We argued that explainable AI systems should produce expert-like explanations and be evaluated on how indistinguishable their explanations are from experts. However, since most current systems identify the importance of pixels and regions, as an intermediate step there should be quantitative ways to determine regions that correspond to those considered important by experts even if there is no label for the region. Many papers show a few representative images and argue why one method is better than another. To illustrate, Figure 7 shows heatmaps produced by GradCAM and occlusion sensitivity (Zeiler & Fergus 2014) on the same network trained to identify whether the bird has wingbars. Adebayo et al. (2018) have proposed some coarse metrics that explainable AI systems should meet, but these are not detailed enough to evaluate slight differences in algorithms that will result in incremental improvements to identifying regions.

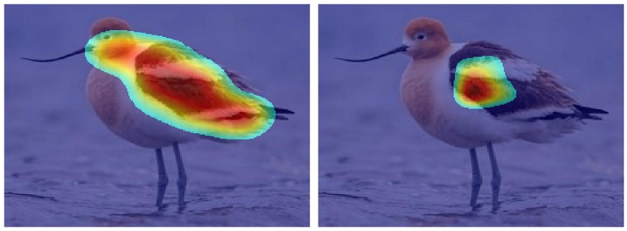

Figure 7. Heatmap for wingbars produced by occlusion sensitivity (left) and GradCam (right).

A measure indicating that the region identified by one approach is better would allow developers to refine explanation algorithms as they do for accuracy. One approach is to have experts mark important regions and evaluate the expert region overlap with XAI regions using a metric such as Dice coefficient. Indeed, this approach is used for U-nets for medical segmentation (Ronneberger et al., 2015). U-nets are a form of image-to-image transformation with training data containing images annotated with regions; the goal is to identify regions on new images. We argue for this methodology for XAI systems, not by giving region annotations in training, but only using them for evaluations. Such a study evaluating how well XAI methods can identify important regions on chest X-rays by Arun et al. (2021) concluded "A variety of saliency map techniques used to interpret deep neural networks trained on medical imaging did not pass several key criteria for utility and robustness." Recently, Arras et al. (2021) have provided metrics and a testbed based on VQA for explainable AI.

Ultimately, user testing of approaches along the lines of Turing's imitation game will be fruitful. This may be premature, however, since current approaches are so far from what experts produce and novices find useful. Current techniques identify important regions without giving them meaningful labels or identify meaningful features but not their locations. We argue both are needed for UCXAI.

Broader Implications

There is a long history of explanation in Artificial Intelligence— much of it user-centered (e.g., Clancey, 1983; Buchanan & Shortliffe, 1984; Swartout & Moore, 1993; Pazzani et al., 1997; Pazzani & Bay, 1999; Herlocker et al., 2000; Leake, 2014; Schank, 2013; Pearl & Mackenzie, 2018;). However, much of recent XAI for inscrutable models (Weld & Bansal, 2019) such as deep learning has used the intuitions of developers as a guide for creating explanations. If the goal is to create explanations for developers, then the developer's intuition is appropriate. Nonetheless, working without experts in a domain may mislead developers into thinking their results will have real world utility (Roberts et al., 2021). DCXAI has also resulted in some types of "explanations" such as heatmaps we have discussed extensively but also lists of words or features with importance scores that had never been thought of as explanations before in the philosophy or psychology of explanation. Older methods such as the permutation method for determining feature importance in "impenetrable" random forests (Breiman, 2001) do the same task but do not refer to them as explanations.

If UCXAI is the goal, investigating the explanations produced by experts, developing systems that replicate these explanations and evaluating the reactions of users to these explanations is the appropriate methodology. Instead of using Amazon Mechanical Turk to annotate videos of cars driving as in Kim et al., (2018), we would suggest recording driving instructors as they explain to new drivers why certain actions should be performed. Instead of an XAI system explaining "the vehicle slowed down because the light controlling the intersection is red," one might get more useful explanations such as "If you're caught behind a brake-happy driver, leave extra distance between your vehicle and theirs so that you don't end up rear-ending them" (from www.wikihow.com/Drive-Defensively).

Conclusion

We argue that many existing explainable AI systems are developer centric. Expert-informed, user-centric explainable AI introduces issues that require additional research: How do we produce explanations like those of experts? How do we help novices learn to be experts? We proposed using Turing's imitation game to evaluate how indistinguishable explainable AI explanations are from those of experts.

Our argument that XAI needs to expand its research agenda can be summarized as XAI needs to answer "what" in addition to "where." Ultimately, we believe causality and "why" need to be addressed.

Acknowledgments

This work was supported with funding from the DARPA Explainable AI Program under a contract from NRL and from NSF grant 2026809. We would like to thank Dave Gunning for stimulating XAI research. Discussions with David Aha, Pat Langley, Eric Vorm, Justin Karneeb, Robert Hoffman, David Kirsh, Lise Getoor, Matt Turek, David Rankin, Dorrit Billman, Tom Fawcett, Aadil Ahamed, Kamran Alipour, Sateesh Kumar, Jordan Levy, Justin Huynh, Che-wei Lin, Nick Lin, Samira Masoudi Linda Zangwill, Mark Christopher, Rui Fan and David Danks were helpful in developing the ideas in this paper. I'm grateful to Rajesh Gupta and Jeffery Ellman for creating an institute that facilitates multidisciplinary research.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I.J., Hardt, M., and Kim, B. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. 2020. Debugging Tests for Model Explanations. In *Advances in Neural Information Processing Systems*.

Anders, C. J., Weber, L., Neumann, D., Samek, W., Müller, K. R., and Lapuschkin, S. 2022. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion,* 77, 261-295.

Arras, L., Osman, A., and Samek, W. 2021. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*.

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M. and Adebayo, J. 2021. Assessing the un trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence* 2021: e200267.

Biessmann, F., and Treu, V. 2021. A Turing Test for Transparency. In *ICML 2021 Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*

Breiman, L. 2001. Random Forests. *Machine Learning* 45, 5–32.

Buchanan, B. G., and Shortliffe, E. H. 1984. *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, MA.

Caruana, R. 1997. Multitask learning. *Machine learning*, *28*1, 41-75.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1721-1730.

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava M., Preece A., Julier S., Rao R., Kelley T., Braines D., Sensoy M., Willis C., and Gurram, P. 2017. Interpretability of deep learning models: A survey of results. In *2017 IEEE smartworld, ubiquitous intelligence and computing, advanced and trusted computed, scalable computing and communications, cloud and big data computing, Internet of people and smart city innovation. pp*. 1-6.

Clancey, W. J. 1983. The epistemology of a rule-based expert system—a framework for explanation. *Artificial intelligence*, 203, 215-251.

DeGrave, A. J., Janizek, J. D., and Lee, S. I. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 1-10.

Došilović, F. K., Brčić, M., and Hlupić, N. 2018, May. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics MIPRO*. pp. 0210-0215. IEEE. MIPRO 2018.

Gunning, D., and Aha, D. 2019. DARPA's explainable artificial intelligence XAI program. *AI Magazine*, *40*2, 44-58.

Herlocker, J. L., Konstan, J. A., and Riedl, J. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. pp. 241-250.

Hoffman, R., Miller, T., Mueller, S. T., Klein, G., and Clancey, W. J. 2018. Explaining explanation, part 4: a deep dive on deep nets. *IEEE Intelligent Systems,* 333, 87-95.

Jonas, J., Cursiefen, C., and Budde, W. 1998 Optic Neuropathy Resembling Normal-Pressure Glaucoma in a Teenager With Congenital Macrodiscs. *Arch Ophthalmol.*;11610:1384–1386.

Khorram, S., Lawson, T., and Fuxin, L. 2021. iGOS++ integrated gradient optimized saliency by bilateral perturbations. *In Proceedings of the Conference on Health, Inference, and Learning*. pp. 174-182.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and sayres, R. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors TCAV. *International Conference on Machine Learning*.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Akata, Z. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision ECCV*. pp. 563-578.

Kligerman, S., Raptis, C., Larsen, B., Henry, T. S., Caporale, A., Tazelaar, H., and Kanne, J. 2020. Radiologic, pathologic, clinical, and physiologic findings of electronic cigarette or vaping product use–associated lung injury EVALI: evolving knowledge and remaining questions. *Radiology*, 2943, 491-505.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. 2020, November. Concept bottleneck models. *In International Conference on Machine Learning*. pp. 5338-5348.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Curran Associates Inc.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K. R., and Samek, W. 2016. The LRP toolbox for artificial neural networks. *The Journal of Machine Learning Research*, *17*1, 3938-3942.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature*, *521*7553, 436–444. https://doi.org/10.1038/nature14539

Lee, E., Braines, D., Stiffler, M., Hudler, A., and Harborne, D. 2019, May. Developing the sensitivity of LIME for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications* Vol. 11006, p. 1100610. International Society for Optics and Photonics.

Leake, D. B. 2014. *Evaluating explanations: A content theory.* Psychology Press.

Lu, J., Lee, D., Kim, T. W., and Danks, D. 2020. Good explanation for algorithmic transparency. In *Proceedings of the 2020 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society.*

Lundberg, S. M., and Lee, S. I. 2017. *A unified approach to interpreting model predictions.* In Advances in Neural Information Processing Systems. pp. 4765-4774

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267. 1-38.

Morcombe, M. and Stewart, D. 2010 eGuide to the Birds of Australia The eGuide to the Birds of Australia. iPhone app version 1.0, mydigitalearth.com, South Africa.

Noh, H., Hong, S., and Han, B. 2015. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*. pp. 1520-1528.

Nourani, M., Kabir, S., Mohseni, S., and Ragan, E. D. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* Vol. 7, No. 1. pp. 97-105.

Pazzani, M., Mani, S. and Shankle, W. R. 1997. *Comprehensible knowledge-discovery in databases.* In M. G. Shafto and P. Langley Ed., In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*. pp. 596-601. Mahwah, NJ:Lawrence Erlbaum.

Pazzani, M. J. and Bay, S. D. 1999. The Independent Sign Bias: Gaining Insight from Multiple Linear Regression. In *Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society*.

Pearl, J., and Mackenzie, D. 2018. *The book of why: the new science of cause and effect.* Basic books.

Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. pp. 1135-1144.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., McCague, C., Beer, L. and Weir-McCall, J.R., 2021. *Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans.* Nature Machine Intelligence, 33. pp.199-217.

Ronneberger, O., Fischer, P., and Brox, T. 2015, October. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. pp. 234-241. Springer, Cham.

Schank, R. P. 2013. *Explanation patterns: Understanding mechanically and creatively*. Psychology Press.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. pp. 618-626.

Swartout, W. R., and Moore, J. D. 1993. Explanation in second generation expert systems. In *Second generation expert systems*. pp. 543-585. Springer, Berlin, Heidelberg.

Turing, A. 1950, Computing Machinery and Intelligence. *Mind*, LIX 236: 433–460,

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3156-3164.

Weld, D. S., and Bansal, G. 2019. *The challenge of crafting intelligible intelligence*. Communications of the ACM, 626, 70-79.

Welinder P., Branson S., Mita T., Wah C., Schroff F., Belongie S., Perona, P. “Caltech-UCSD Birds 200”. California Institute of Technology. CNS-TR-2010-001. 2010

Zeiler, M. D., and Fergus, R. 2014, September. Visualizing and understanding convolutional networks. In *European conference on computer vision*. pp. 818-833. Springer, Cham.