Cognitive Differences in Human and AI Explanation

Robert A. Kaufman (rokaufma@ucsd.edu)
University of California, San Diego, Department of Cognitive Science

David J. Kirsh (kirsh@ucsd.edu)
University of California, San Diego, Department of Cognitive Science

Abstract

How do humans explain and cognize visual information? Why do AI explanations in radiology, despite their remarkable accuracy, fail to gain human trust? In a study of 13 radiology practitioners, we found that AI explanations of x-rays differ from human explanations in 3 ways. The first concerns visual reasoning and evidence: how humans get other humans to see an interpretation's validity. Machine learned classifications lack this evidentiary grounding, and consequently XAI explanations like heat maps fail to meet many users' needs. The second concerns the varying needs of interlocutors. Predictably, explanations suitable for experts and novices differ; presuppositions on explainee knowledge and goals inform explanation content. Pragmatics matter. The third difference concerns how linguistic terms and phrases are used to hedge uncertainty. There is no reason XAI might not satisfy these human requirements. To do so, however, will require deeper theories of human explanation.

Keywords: Explainable AI; Explanation; Radiology; Visual Reasoning

Introduction

How do humans explain and cognize visual information? Here, we present findings derived from observation and ethnographic study of 13 expert (attending) radiologists and radiology residents explaining their findings and impressions to end practitioners such as other attending radiologists, radiology residents who are midway through residency, and medical students doing a rotation in radiology.

We found that Explainable AI (henceforth XAI) explanations of COVID-19 x-rays differ from human explanations in three ways that teach us something generalizable about the joint activity of explaining how to interpret an image.

The first concerns visual reasoning and evidence. Humans get other humans to see the validity of an interpretation by explaining *why* they see what they see. That is, they are familiar with the process of directing attention to relevant details, providing evidence for claims, and linking what they see to why it matters. Machine learned classifications lack this evidentiary understanding, with the consequence that popular visualizations such as heat maps do not meet many users' explanatory needs. Language plays an important role in directing how a listener visually inquisitions an x-ray, how they move from heat map to x-ray and back again, and how they attribute meaning to what they see. Even the best XAI image classifiers fail in this regard.

The second way XAI explanations differ from those of humans concerns their sensitivity to the needs of different interlocutors. When an explainee receives an explanation, they are engaged in a conversation and bound by rules for cooperative conversation (Grice, 1975), which put constraints on what should be said and how. Predictably, an explanation suitable for an expert is not suitable for a novice, and vice versa. The common ground (Clark, Schreuder, & Buttrick, 1983) between groups is different. Accordingly, the Gricean maxims of *quantity* and *relevance* would predict briefer exchanges between experts. A more surprising observation, however, is that two explainees often need different types of explanations because the more novice of the two will typically see the context of explanation as an educational moment that goes beyond visual interpretation of a particular image. They want more generalizable knowledge; they may even want knowledge unconnected to the specific case, an 'explanation' that tells them about x-ray interpretation or reporting itself. The pragmatics of explanation matters.

The third difference concerns hedging—how an explainer conveys uncertainty. Linguistic hedges like 'sometimes' or 'likely' may convey uncertainty throughout an explanation to provide a comprehensive reliability measure. All this matters for a theory of human visual interpretation and visual explanation, but it also has a practical side: AI systems in Radiology, despite their remarkable accuracy, currently fail to gain human trust because their explanations are inadequate (Holzinger, Biemann, Pattichis, & Kell, 2017; Holzinger, Langs, Denk, Zatloukal, & Müller, 2019). This paper is about the ways current XAI in radiology fail, and how they can be improved by modeling them after humans.

Background: XAI in Radiology

Current diagnostic AIs for radiology classify images by applying dozens of statistical measures over a full grayscale image. Radiologists forming diagnostic interpretations are also sensitive to grayscale changes, but they tend to focus on edges, blobs, areas of major contrast, and textures describable in natural language. Machines are sensitive to changes within convolution windows of arbitrary size (Ribeiro, Singh, & Guestrin, 2016; Lapuschkin, Binder, Montavon, Müller, & Samek, 2016; Selvaraju et al., 2017) regardless of whether these correspond to attributes describable in natural language. The implication is that subtle shading or textural changes that are statistically informative to the machine may be uninformative to humans. This makes it challenging to explain the

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

evidentiary basis of a machine's classification.

The most common way radiology XAI systems attempt to explain an image interpretation, such as an x-ray with a 'COVID-19' classification, is by providing another image – a heat map. This is meant to help users understand what regions (specific pixels) the machine regards as of greater or lesser importance. Common examples include LIME (Ribeiro et al., 2016), GradCAM (Selvaraju et al., 2017), and LRP (Lapuschkin et al., 2016). These are generally paired with an overall classification label, such as COVID-19 pneumonia, and a probability measure for certainty. Figure 1 shows a heat map explanation for x-ray interpretation using GradCAM from Chexpert, a classification system for chest x-rays (Irvin et al., 2019).

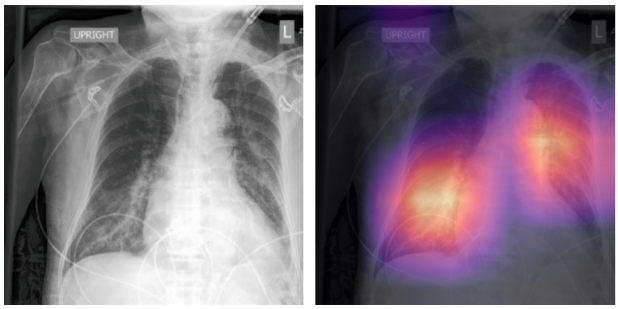

Figure 1: A heat map visual explanation for chest x-rays produced by Chexpert using GradCAM (right) and the original x-ray image (left) it is laid over. Paired with the visualization is an accompanying classification and probability measure "Pulmonary edema, p = 0.824" (Irvin et al., 2019).

Chexpert is impressive in its interpretative accuracy, as are many other classification systems using similar techniques (Rajpurkar et al., 2017; Karim et al., 2020). We argue that these 'explanations', however, fall short of those given by human radiologists. First, they fail to draw attention to the visual evidence in the radiograph in the way human explainees need in order to understand the basis for the interpretation. Further, they fail to form a set of logical premises that connect the visual information to a clinically-meaningful radiological impression using steps of justification.

Explainees need to know what to look at, and in what order. XAIs fail in this regard: they do not direct the temporal order of attention. A novice might look anywhere in a heat map and fail to contextualize what they see because they do not see the connections between different regions correctly.

In a human-human explanation, by contrast, one person calls attention to specific regions, and within those regions to specific features. As an explainer moves through an image, focusing on what is deemed relevant at each step, they create an argument that constitutes a chain of evidence similar to step-by-step reasoning in language (Schwartz, Panicek, Berk, Li, & Hricak, 2011). The primary difference is that when reasoning in language, it is necessary to identify attributes using descriptions, whereas in a visual explanation, descriptors are replaced by reference to visual regions displaying those attributes. This assumes the explainee has enough background knowledge to recognize what is salient. If they do not, they can ask in language what they should be seeing.

When this temporal ordering of joint attention is successful and paired with enough information to derive meaning, the explainee should understand the grounds for an image's classification. In radiology this classification process is called an interpretation, the conclusion being an impression. When the sequence is wrong, or an explainer and explainee lack the requisite common ground to jointly attend to the salient attributes and make meaning of them, the two are misaligned. Common ground fails to build up in the right way. As the two move apart neither can be certain they know what the other expects them to know.

The most obvious way to improve XAI for image understanding is to rely on language or some visual highlighting mechanism to call attention to attributes one assumes the explainee knows. Indeed, there have been efforts to group pixels based on perceptual factors like continuity and similarity and tie these to named shapes (Koontz & Gunderman, 2008). Pattern-recognition of these shapes might then be meaningfully related to radiological concepts drawn from memory and built through expertise (Wood, 1999). Connecting visual features to an overall impression could then be assisted with.

Assuming the question of temporal order has also been resolved, such efforts may one day improve XAI for image classification. Provided, of course, the system is appropriately calibrated to the features different users can actually recognize. This is no small proviso. Experts see many more attributes than simple features like blobs, edges, and a few describable textures. Calibration is a hard and very human problem. When two people talk they soon determine what they can assume the other knows. They also have methods for recognizing what they do not share in common. A real question, therefore, is whether an XAI can adapt and calibrate to common ground.

Method

To explore the key factors impacting how radiologists explain their impressions to colleagues, we a) ethnographically observed how radiologists and radiology residents interpret and explain radiographs; b) interviewed and administered a survey to further assess factors which affect human to human explanations in radiology. Study procedures occurred in 90 minute sessions over video conference.

Participants

Thirteen (n = 13) participants took part in the study and each participant completed all study aspects. Seven participants were attending radiologists and six were radiology residents. Participant experience ranged from a first-year radiology resident to an attending radiologist with 25 years of experience. All participants were employed within large hospital systems dispersed throughout the United States.

Observation: X-ray Interpretation

During the x-ray interpretation task, participants were observed interpreting and explaining chest radiographs of potential COVID-19 patients via video conference.

Explanations were performed orally, as if the given explainee presented the participant with a radiograph and asked them for a second opinion - one of the most common contexts within which explanations occur. Participants were told that each patient was presenting clinical symptoms of COVID-19 such as cough, fever, and shortness of breath.

Explanations were unidirectional and the participants were never face-to-face with the receivers; instead, they were told before each image to which receiver they should tailor their explanation. This allowed explanations to match current radiology XAI systems that do not allow for interactive dialogue. Even though participants were not interacting with explainees themselves, they reported that the study setup felt natural to them and they didn't have issues tailoring each explanation.

Each participant saw approximately 12 radiographs to interpret and explain; images were randomized and half were repeated to multiple explainee types. By repeating images we are able to examine how explanations differ when given to different explainees. For repeat images, participants were asked to explain as if they were seeing the image for the first time. Given that the interpretation itself is not the basis of inquiry, we do not believe that repeating biased the explanation given. This was confirmed by participants during interview. Pairwise comparisons were performed to ensure that there were no differences between explanations of images that were and were not repeated and any differences between participants or images were controlled for during analysis.

The x-rays interpreted by our subjects came from one of the large COVID-19 open datasets pairing radiological images and diagnostic information (Chowdhury et al., 2020). Our video observations closely match the natural patterns of communication employed at the time of collection (Matalon et al., 2020).

Explanation Coding and Nomenclature

Participant explanations were transcribed and broken into segments to reflect different types of information conveyed during the interpretation explanation process. Codes were assigned based on how certain types of words and phrases corresponded with each segment. Content was analyzed to reflect the information communicated within each segment and how the segments unfold over time. This allowed us to evaluate how thoroughly each level was covered by explanations to different receiver groups, and by what means.

For codes relating to radiology terminology, we used the Fleischner Society's Glossary of Terms (Hansell et al., 2008) along with contemporary references to COVID-19 image findings (Ng et al., 2020) and the Radlex Radiology lexicon (Langlotz, 2006).

We conform to the standard definition of radiological 'findings' as those statements that describe observations - what is seen in an image using theory-based jargon (Hall, 2000). For instance, a 'ground-glass opacity' is a radiological finding. Radiological 'impressions' are theoretical conclusions; they communicate what the findings mean in terms of pathology, like COVID-19 pneumonia.

Codes relating to the expression of uncertainty, or linguistic hedging, were taken with permission from the collection of terms and phrases used by Hanauer et al. in their analysis of uncertainty in clinical documents (Hanauer et al., 2012) as well as the certainty descriptors in the Radlex Radiology lexicon (Langlotz, 2006).

Results

We elaborate upon the content of an explanation in radiology and illustrate how this content changes by end practitioner.

The Role of Presuppositions

The background knowledge and expertise of an explainee is one of the more obvious factors a radiologist takes into account when tailoring an explanation. We measured participants' presuppositions of x-ray interpretation (process) and radiology terminology (terms) expertise by asking them to rate each group on a 1-7 likert scale (no expertise to super expert).

Table 1: Presuppositions of Explainee Expertise

	Attending		Resident		Student	
	Mean	SD	Mean	SD	Mean	SD
Terms	6.2	0.3	4.7	0.3	2	0.2
Process	5.8	0.3	3.9	0.3	1.7	0.2

Results follow the expected trend: the higher a receiver's role in the healthcare system, the more expertise they are presupposed to have (Table 1). Presupposition rates given by attending radiologists and radiology residents did not differ significantly. These results provide the foundational evidence underlying the assumption that more ground needs to be covered in explanations to explainee groups who are presupposed to have less expertise. Past work has noted differences in end practitioner domain expertise, but to our knowledge no radiology XAI systems have been implemented which take into account these differences.

Explanation Content as Segments

To analyze content, we break human multimodal explanations into linguistic segments that can be counted. These segments are linguistic units (words or short phrases) which combine to form the full explanation (Passonneau & Litman, 1997). Segments were determined by identifying the categories of information which explanations progressively cover, from low-level visual features to abstract impressions using domain-specific radiology jargon. We also measure 'elaborations', which are segments that are categorized by their function in the explanation, such as providing additional information or next steps. Once segments were determined,

associated words and phrases were quantitatively measured. We posit that explanations themselves do not need to cover all of the segments explicitly. Instead, presuppositions about a receiver's knowledge and needs dictate what information should be included. Figure 2 presents a simplified example with segments highlighted.

"This blob[1] might be[5] a ground glass opacity[2] by the hazy shape[2a]. This makes me think there's an infection[3] like COVID[3]. Given the pandemic this would be my impression[3a] and I'd order a follow-up CT[4]. If it's an artifact[6], I would check their lung apices for signs of...[7]"

Figure 2: A simplified example explanation with segments highlighted. The segments are: 1.) Identifying a ROI, 2.) Abstracting the ROI features to radiological 'finding terms', 2a.) If needed, assist with this connection through 'finding elaboration', 3.) Inferring 'impression terms' from findings, 3a.) If needed, assist with this through 'impression elaboration', and 4.) Contextualize the inferences and add 'orders' as next steps. Throughout, the segments are modified to include 5.) Certainty via 'hedging' terms and phrases, and 6.) Alternative conclusions via counterfactuals. 7.) The x-ray interpretation process is expanded upon via 'process elaborations' if needed.

The first segment in Figure 2 establishes a visual reference point, a region of interest (ROI), on the image. Identifying a ROI enables joint attention to visual attributes if the explainee has the ability to recognize the attributes. Typically a ROI is pointed out by using a gesture and a simple linguistic descriptor or two with indexical, as in "this blob".

Second, the visual attributes of the ROI are connected to radiological finding terms, such as "ground glass opacity". This allows the ROI to be further identified using domain-specific language that may help the practitioner understand what they are seeing. Some explainees have difficulty connecting ROIs to finding terms. In these instances, an explainer may include a 'finding elaboration' - a pedagogical extra - where information is included to help them understand how they know they are seeing a particular finding.

Next, impressions are induced from the constellation of findings identified in the image. This is important for clinical sensemaking, as impressions are the primary information that helps other providers identify next steps and form a treatment plan (Hall, 2000). Forming impressions from findings does not occur in a vacuum; the clinical context, patient history, and alternative interpretations of findings may iteratively inform the likelihood of different impressions. Sometimes a single impression cannot be identified and instead a differential is communicated (Dahnert, 2017). Similar to finding elaborations, 'impression elaborations' include extra information given by an explainer to help a particular receiver understand what a particular finding means in terms of an impression. They help with the inferential process and may not be required for all receivers.

Fourth, impressions are contextualized within the clinical context and 'orders', or next steps, may be included in explanations to help the receiver know what should be done once the interpretation process has concluded. This may include information such as orders for additional images or instructions to other physicians.

Fifth, indicators of uncertainty - hedges - are vitally important in medical settings. The weight or confidence an explainer has in their finding, impression or interpretation has been shown to impact medical decision making (Hanauer et al., 2012; Khorasani et al., 2003). In our data, we found the use of linguistic hedging in all stages of an explanation. Their inclusion as a fifth segment carries no information about when they may appear in an explanation.

Sixth, if-statements and alternatives are another type of linguistic segment that may appear at any point in the interpretation process. They offer a differential or other way of interpreting the image. These often take the form of contrastive or counterfactual explanations, such as "if the shape were more precise it would have been a B, but since it is not it must be an A" (Miller, 2019; Wang, Yang, Abdul, & Lim, 2019).

The seventh type of segment - process elaboration - is another form that may appear at any point in an explanation. These elaborations are included to assist explainees on any of the component tasks involved in the interpretation process, from reading a radiograph to understanding what to do when certain findings or impressions are identified.

We found that explanations of decisions tend to be communicated in a linear fashion via these segments, however, this is not always the case. Explanations may sequentially follow the explainer's process of discovery which may be implicit and nonlinear. For example, one participant described their deductive (top-down) discovery process *"forming a hypothesis based on the clinical context ... searching for evidence to confirm or reject the hypothesis, and then assessing for [alternatives]"*. Another described an inductive (bottoms-up) process of looking for *"areas that seem off,"* forming a hypnosis based on what they find, and then gathering further evidence to test the hypothesis. We refer to the Select and Test model of medical reasoning for how radiological impressions may be concluded upon (Ramoni, Stefanelli, Magnani, & Barosi, 1992).

Explanations Change With End Practitioner

There was a clear and intuitive difference in the way explanations were presented to attending radiologists, radiology residents, and medical students. Overall differences manifest in nearly all segments of an explanation, providing evidence that explanation content changes by end practitioner.

Analysis was done using linear mixed-effects models in order to account for individual participant differences and for the effect of individual x-ray stimuli used in the study (Bates,

Mächler, Bolker, & Walker, 2014). Comparisons between a model with random effects for these variables and that same model with a fixed-effect for end-practitioner type are shown. This fixed effect of end practitioner type was added to models omitting only that variable to assess for significant differences between explainee type. There were no statistically significant differences between explanations *produced by* attending radiologist and radiology resident participants, thus these explainers are presented conjointly.

Unique terms for each segment are counted to accurately represent content coverage and control for repeat utterances. As some words convey more uncertainty than others, a weighted total was calculated for hedging to reflect the amount of uncertainty conveyed. Elaborations were counted as it relates to their corresponding segment descriptions. The corresponding descriptive statistics are reported in Table 2 and test statistics are reported in Table 3. In Table 3, 'Att-Res' should be read as 'Attending Radiologist explaining to a Radiology Resident', and so on.

Table 2: Explanation segment counts by end practitioner.

	Attending		Resident		Student	
	Mean	SD	Mean	SD	Mean	SD
Total words	110.26	12.71	151.25	16.46	280.28	25.91
Findings (terms)	1.72	0.18	2.13	0.21	2.64	0.35
Findings (elab)	0.23	0.07	0.78	0.15	1.85	0.27
Impress. (terms)	3.21	0.37	3.53	0.26	4.31	0.47
Impress. (elab)	1.59	0.21	2.00	0.23	2.64	0.24
Orders (elab)	0.36	0.09	0.38	0.09	0.67	0.13
Hedging (weighted)	7.95	1.57	10.40	1.58	12.72	2.05
Process (elab)	0.31	0.10	0.53	0.11	2.67	0.44

Overall, we find that information was included within all segments of an explanation to all groups – albeit at different levels of frequency. This implies that a baseline level of information exists that is necessary to connect visual findings with impressions regardless of prior expertise. This does not imply that all segments should be explained at all times; some contexts and some end practitioners may require certain types of information to be added or omitted to achieve common ground and fulfill additional goals.

We also find clear evidence that explanations to different end practitioners contain different amounts and types of information. In general, the amount of content within each explanation segment increases as we move from end practitioners with more presupposed expertise (attending radiologists and radiology residents) to those with less (medical students). This supports the principle tenets of common ground theory. Crucially, content differences were segment-dependent, implying that in order to meet the explanatory needs of each group, different information will need to be presented.

Table 3: Comparing end practitioner segment counts.

	Chi-sq	df	Overall	Att - Res	Att - Stu	Res - Stu
Total words	50.10	2	***	trend	***	***
Findings (terms)	8.40	2	*	ns	**	trend
Findings (elab)	35.88	2	***	*	***	***
Impress. (terms)	5.81	2	trend	ns	*	ns
Impress. (elab)	10.84	2	**	ns	**	*
Orders (elab)	5.59	2	trend	ns	*	*
Hedging (weighted)	8.78	2	*	ns	**	ns
Process (elab)	38.34	2	***	ns	***	***

*'***' p <0.001, '**' p <0.01, '*' p <0.05, 'trend' p <0.1*
Overall: Pr(>Chi-sq), Group Differences: Pr(>| t |)

We found few significant differences between explanations to radiology residents and attending radiologists, though we observe clear trends. We believe this is due to the near-expert status of radiology residents and the high prevalence of COVID-19 at the time of the study. With a less common clinical case we may see more pronounced differences.

We know more information was provided to some groups, but what role does this additional information play? In some cases, additional information may be included to point out content that may be missed or misunderstood by those with less expertise. We believe this to be the case with hedging and added elaborations. In other cases, such as the inclusion of many finding and impression terms, counterfactuals were used to illustrate differences between *fact* and *foil* findings and why some impressions were included in a differential over others. Contrastive explanation strategies, often through counterfactuals, can be especially effective at establishing causal attribution (Miller, 2019).

Discussion

Our analysis reveals stark differences between human-given and machine-derived explanations. These lead to several critical areas to be addressed for future XAIs that are viable for real-world deployment.

Visual Reasoning: Moving Beyond ROIs

We show that radiology practitioners move through their explanations in segments of increasing abstraction, ascribing meaning to visual elements of a radiograph at multiple levels of the interpretation process, providing additional information and adjusting information fluidly. These steps are important for establishing common ground and justifying the interpretation in terms of a shared process. Common ground builds up over time; it has a logic that relies on shared capacities for inference and judgment. In the case of visual reasoning, this involves helping explainees see what they need to and in the right order.

By contrast, heat maps and similar XAI techniques identify potentially important ROIs but do not determine the logical or temporal sequence an explainee should take when looking at ROIs. Moreover, such techniques do not explain the reason *why* certain pixels are important. There is no explanation of the relation of one ROI to another or to the clinical case.

This is quite unlike the information conveyed by radiology practitioners who have a near-linear approach to controlling attention. Moreover, by using known terms when identifying the findings in a region they encourage contextual thinking. "Ground glass" for example, is understood as caused by medical conditions, such as infection, chronic interstitial disease, or acute alveolar damage (Hansell et al., 2008). Its semantics is close to physiology and the meaning of pulmonary disease, which invokes a constellation of inferences. Because it is not just a measure of statistical importance, it is more intuitive than statistical heat maps for someone trying to understand what it is about a patient that reveals they have COVID-19 pneumonia. It makes it easier to cross-examine the x-ray to look for meaningful connections.

A similar opportunity for sensemaking is created by elaborations, counterfactuals, and talking of next steps. All these linguistic types provide guidance and contextualization. Elaborations come in many forms depending on explainee need, from helping form a diagnosis to passing along tacit knowledge on how to go about the interpretation process itself. Once an impression or differential is decided upon, radiology practitioners also help contextualize that information within the clinical context as needed. Elaborations help to structure joint attention and develop common ground.

Moving beyond highlighting potential ROIs, we believe that XAIs should convey information at all levels of the interpretation process. This can provide an intuitive justification for a diagnosis as well as enable an end practitioner to cross-examine the XAI. Examples include identifying features as findings using domain-specific terminology like "ground glass", visualizing how a constellation of findings contributes to an impression, or suggesting possible next steps given a case.

Explanation Pragmatics: Receiver Needs Vary

Radiology practitioners tailor their explanations to different explainees and are sensitive to their specific needs. We observed that human explainers vary the amount of information within each segment based on explainee expertise and thoughtfully choose what to elaborate.

Current XAIs provide no sensitivity, leaving more novice end practitioners unsupported with the task of connecting the meaning of different segments of information and how to utilize information in the larger ecosystem of patient care. More expert users, meanwhile, may be provided with unnecessary information that is redundant or distracting.

Hedging Uncertainty

Finally, humans differ from XAI in radiology in the way they convey uncertainty. Hedging is found at nearly all levels of the interpretation process, including on the process itself if necessary. This is important, as uncertainty attached to different types of information has different implications on how that information should be used. For example, if an explainer's uncertainty stems from being unsure of the clinical meaning of a finding it has drastically different implications than uncertainty stemming from not knowing if the finding is real or an artifact.

Most current XAIs only convey overall quantitative certainty - COVID with 90% probability. Research on human decision-making with probabilities illustrates that people cannot meaningfully distinguish certainty quantifications at fine granularity nor without conforming to heuristic biases (Kahneman & Tversky, 2013). Reasoning is more qualitative, and past work suggests that explanations with probabilities tied to causal events are more useful (Miller, 2019).

More work needs to be done to understand the variety of ways explainers have of marking uncertainty and the different meanings they convey.

Conclusions

To understand the inadequacies of Explainable AI (XAI) for radiological interpretations, we examined how radiology practitioners explain their impressions to attending radiologists, radiology residents, and medical students. Through ethnographic and quantitative analysis, we dissected their explanations into seven types of linguistic segments. We found that segments were presented in levels of increasing abstraction, directing attention and ascribing meaning to visual attributes of an image to facilitate clinical sensemaking. Explanations were sensitive to the knowledge, needs and goals of the explainees and information was added or subtracted as needed to achieve common ground.

XAI systems do not at present explain like humans. The most popular forms rely on heat maps that give no guidance on how to attend to features, how to make sense of the relations between features, and what features mean in the larger clinical context. This XAI approach reflects a failure in three areas: it fails to accommodate how humans reason and make sense visually; how we hedge our uncertainty in a qualitative manner, and how we are sensitive to the many needs and goals our explainees may have.

Acknowledgments

We would like to thank the radiology practitioners who participated in the study and Michael Pazzani for his support. Funding was provided through NSF grant #2026809 and the DARPA Explainable AI Program under contract from NRL.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Chowdhury, M. E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., ... others (2020). Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, *8*, 132665–132676.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, *22*(2), 245–258.

Dahnert, W. F. (2017). *Radiology review manual*. Lippincott Williams & Wilkins.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Hall, F. M. (2000). Language of the radiology report: primer for residents and wayward radiologists. *American Journal of Roentgenology*, *175*(5), 1239–1242.

Hanauer, D. A., Liu, Y., Mei, Q., Manion, F. J., Balis, U. J., & Zheng, K. (2012). Hedging their mets: the use of uncertainty terms in clinical documents and its potential implications when sharing the documents with patients. In *Amia annual symposium proceedings* (Vol. 2012, p. 321).

Hansell, D. M., Bankier, A. A., MacMahon, H., McLoud, T. C., Muller, N. L., & Remy, J. (2008). Fleischner society: glossary of terms for thoracic imaging. *Radiology*, *246*(3), 697–722.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... others (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 590–597).

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part i* (pp. 99–127). World Scientific.

Karim, M., Döhmen, T., Rebholz-Schuhmann, D., Decker, S., Cochez, M., Beyan, O., et al. (2020). Deepcovidexplainer: Explainable covid-19 diagnosis based on chest x-ray images. *arXiv preprint arXiv:2004.04582*.

Khorasani, R., Bates, D. W., Teeger, S., Rothschild, J. M., Adams, D. F., & Seltzer, S. E. (2003). Is terminology used effectively to convey diagnostic certainty in radiology reports? *Academic radiology*, *10*(6), 685–688.

Koontz, N. A., & Gunderman, R. B. (2008). Gestalt theory: implications for radiology education. *American Journal of Roentgenology*, *190*(5), 1156–1160.

Langlotz, C. P. (2006). *Radlex: a new method for indexing online educational materials.* Radiological Society of North America.

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

Matalon, S. A., Souza, D. A., Gaviola, G. C., Silverman, S. G., Mayo-Smith, W. W., & Lee, L. K. (2020). Trainee and attending perspectives on remote radiology readouts in the era of the covid-19 pandemic. *Academic radiology*, *27*(8), 1147–1153.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*.

Ng, M.-Y., Lee, E. Y., Yang, J., Yang, F., Li, X., Wang, H., ... others (2020). Imaging profile of the covid-19 infection: radiologic findings and literature review. *Radiology: Cardiothoracic Imaging*, *2*(1), e200034.

Passonneau, R. J., & Litman, D. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, *23*(1), 103–139.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... others (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Ramoni, M., Stefanelli, M., Magnani, L., & Barosi, G. (1992). An epistemological framework for medical knowledge-based systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(6), 1361–1375.

Ribeiro, M. T., Singh, S., & 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).

Schwartz, L. H., Panicek, D. M., Berk, A. R., Li, Y., & Hricak, H. (2011). Improving communication of diagnostic radiology findings through structured reporting. *Radiology*, *260*(1), 174–181.

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. In *Proceedings of the ieee international conference on computer vision* (pp. 618–626).

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–15).

Wood, B. P. (1999). Visual expertise. *Radiology*, *211*(1).