# Do Explanations Help or Hurt? Saliency Maps vs Natural Language Explanations in a Clinical Decision-Support Setting

Anonymous EMNLP submission

## Abstract

As AI models are becoming more powerful, their adoption is becoming more widespread, including in safety-critical domains. Explainable AI (XAI) has the aim of making these models safer to use, for instance by making their decision-making process more transparent. However, current explainability methods are seldom evaluated in the way they are intended to be used: by real-world end users. To address this, we conducted a large-scale user study with 85 clinicians in the context of human-AI collaborative chest X-ray analysis. We evaluated three types of explanations: saliency maps, natural language explanations, and their combination. We specifically examine how different explanation types influence users depending on whether the AI is correct. We find that the quality of explanations, i.e., how much correct information they entail, and how much this aligns with AI correctness, significantly impacts the usefulness of the different explanation types.

## 1 Introduction

A significant barrier to the adoption and regulatory approval of deep learning models in medical imaging is the limited understanding of the decision-making processes underlying these models (Langlotz et al., 2019). The combination of lack of model robustness (Papernot et al., 2016), bias (algorithms are prone to amplifying inequalities that exist in the world) (Obermeyer et al., 2019; Hajian et al., 2016), and the high stakes in clinical applications (Vayena et al., 2018) prevent black-box algorithms from being used. XAI is proposed as a promising solution to address the inherent issues of model robustness, bias, and lack of transparency in medical imaging (Borys et al., 2023).

While various XAI methods have been proposed to increase the transparency of AI models, such as saliency maps (Saporta et al., 2022), counterfactual explanations (Schutte et al., 2021), and natural language explanations (Kayser et al., 2022), the practical utility of these approaches within clinical settings remains poorly understood. While there is an abundance of literature and regulatory frameworks that advocate the significance of interpretability in medical applications (Frasca et al., 2024), only a few studies investigate how useful these explanations are for end-users, with some studies suggesting that these methods may not work as well as anticipated (Adebayo et al., 2018; Hoffmann et al., 2021; Margeloiu et al., 2021; Shen and Huang, 2020).

Evaluating the effectiveness of XAI explanations is a challenging task, as there can often be a variety of correct ways to explain a decision and the criteria for judging their quality are diverse (e.g., plausibility, faithfulness, clarity (Miller, 2019)). As one main value of explanations is their utility to end-users, it's crucial to evaluate explanations with human subjects. As explanations are prone to confirmation bias and user preference doesn't always correspond to desired explanation quality requirements, proxies are developed for evaluating explanation usefulness (Ehsan and Riedl, 2020; Liao et al., 2022; Liao and Varshney, 2021; Ehsan et al., 2021).

We address this by carrying out a large-scale human subject study to evaluate the usefulness of natural language explanations (NLEs), saliency maps (SM), and a combination of both (COMB), in the setting of imperfect AI and imperfect XAI. Specifically, we investigate how different explanation types impact users in a clinical decision-support system (CDSS) setting, with respect to both AI accuracy and XAI quality. As the main purpose of AI in medical applications is arguably to enhance practitioners in CDSS settings (Langlotz, 2019; Agrawal et al., 2019), our proxy for the usefulness of explanations is how much they improve human performance in a human-AI collaborative chest X-ray analysis. In our study, 85 clinicians analyse 80 images each, under the setting of four different AI models. SMs, the prevailing mode of interpretability in medical imaging (Van der Velden et al., 2022), attribute importance weights to regions in an image. We compare them to NLEs, for which there have been calls to deploy them in clinical practice (Reyes et al., 2020), and which are known to be user-friendly and able to explain more complex reasoning (Kayser et al., 2022). We also study whether a combination of them further enhances human performance.

Our results show that explanation correctness (EC) is an important factor in deciding whether AI explanations are helpful or harmful to end users. When the AI is correct, incorrect explanations are detrimental to human-AI performance, but equally, when the AI is incorrect, correct explanations mislead users into agreeing with the AI. We also find that the combination of NLEs and SMs is the most useful, and interestingly is better than SMs even though NLEs on their own are significantly worse.

## 2 Related Work

**XAI in medical imaging** XAI methods can be broadly classified into post-hoc explainers or self-explaining models, i.e. approaches that either explain trained black-box AI models, or approaches that are inherently designed and trained to be explainable. Both types have been applied widely in medical imaging applications (Irvin et al., 2019; Thomas et al., 2019; Verma et al., 2020; Koh et al., 2020; Gale et al., 2018). In this study we focus on SMs (post-hoc), a common XAI method for medical imaging (Irvin et al., 2019; Thomas et al., 2019), and NLEs (self-explainable), which are user-friendly, can convey complex reasoning, are promising for clinical applications (Reyes et al., 2020), and ever more widespread with the rise of large language models.

**Human-AI collaboration in medical imaging** The rapid advancements in AI spurred discussions about its capability to automate processes and outperform humans in specific tasks. However, a parallel discourse is centered on how AI can enhance, rather than replace, humans, a domain referred to as human-AI collaboration. This has been studied in areas such as content generation and moderation (Lee et al., 2022; Zhang et al., 2022; Jhaver et al., 2019; Lai et al., 2022), and visual recognition (Colin et al., 2022; Kim et al., 2022). Especially in medical imaging, where concerns around safety and trust make autonomous deployment of AI models challenging, there is an emphasis on how AI can collaboratively support medical professionals. Clinical Decision Support Systems (CDSSs), where AI models offer recommendations to humans for specific tasks, are a common form of human-AI collaboration in clinical practice.

DCSSs have been getting increasing attention in radiology. Existing studies investigate this form of human-AI interaction by looking at how the sequential order of human and AI decisions affect performance (Fogliato et al., 2022), what influence the assertiveness of AI suggestions has (Calisto et al., 2023), or which kind of users benefit the most from it (Gaube et al., 2023). A recent large-scale study conducted by Agarwal et al. (2023) shows that, in most cases, human performance is enhanced when using DCSSs.

In this work, we built upon this literature by evaluating the usefulness of different XAI explanations in the context of a DCSS for chest X-ray analysis. However, in contrast to previous works, we specifically focus on imperfect AI and XAI by controlling the accuracy of both AI predictions and explanations.

**Evaluating XAI** Evaluating AI explanations is less straightforward than evaluating e.g., prediction performance. The lack of unique ground truth, the wide range of interpretability goals, as well as the human-computer interaction aspect, make this more difficult. Thus, differences in the effectiveness of existing XAI methods are not well understood (Gaube et al., 2023). For these reasons, a growing body of work is evaluating XAI methods through the lens of human subject studies, following one of three predominant methodologies.

**User Preference** Some studies directly measure human participants' preferences for XAI explanations. For instance, Adebayo et al. (2020) simulated a quality assurance context, requesting participants to assess the deployment readiness of AI algorithms, which came with different kinds of explanations. However, Hase et al. (2020) demonstrated that user preference does not correlate with how well users can predict model behavior, a proxy for how transparent the model is. Additionally, there are concerns that humans might fall prey to confirmation bias, the tendency to believe that the system used the features they think are most important (Rudin et al., 2021-03-20). There is also evidence that XAI methods can unreasonably increase the

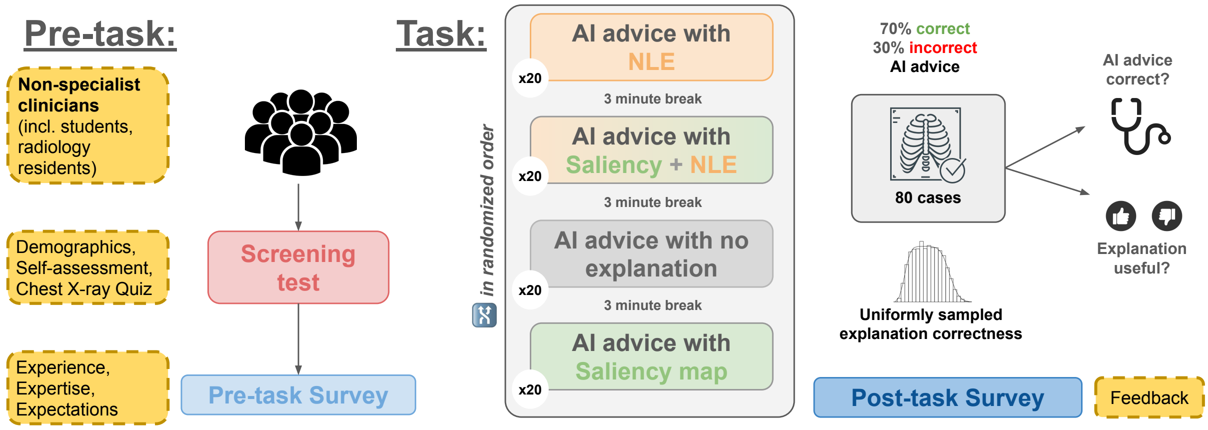


Figure 1: Study design overview.

confidence in a model's prediction (Kunkel et al., 2019; Schaffer et al., 2019; Ghassemi et al., 2018; Eiband et al., 2019).

**Model Predictability.** Arguably, the closest proxy for full model transparency is to measure how well humans can predict a model's predictions on unseen data. If users achieve 100% accuracy, it would mean the model is entirely transparent to them. While this method is common to evaluate XAI explanations (Alqaraawi et al., 2020; Colin et al., 2022; Yang et al., 2019; Shen and Huang, 2020), its applicability to radiology is limited, as predictions are highly nuanced and multiple labels can apply and each come with their own, different explanations.

**Human-AI Collaboration.** Another approach to evaluate the usefulness of XAI explanations is to measure how much they improve human performance in the AI-human collaborative setting. The goal of XAI in this setting is to guide the user to appropriate evidence when the model is correct, or shed light on faulty AI decision-making when it is wrong. Chu et al. (2020) measured the impact of XAI methods in helping users predict age given images of human faces. Kim et al. (2022) analyzed performance changes in a bird classification task under the guidance of various XAI techniques. In clinical applications, where practitioners see a need for explanations to justify "their decision-making in the context of a model's prediction" (Tonekaboni et al., 2019), this evaluation method is particularly well suited and hence also used in this work. Existing work most similar to ours is by Morrison et al. (2023), who are the first to look at NLEs and consider imperfect XAI. We differ by the task (safety-critical CDSS vs bird classification), continuous EC scores instead of binary, considering EC as how well it explains the AI advice even when incorrect, looking at SMs SMs+NLEs).

**Evaluating XAI in Clinical DCSS.** Few works looked at the usefulness of XAI in clinical applications. Du et al. (2022) consider a simple, 5-feature set-up to compare explanation-based and feature attribution methods in the CDSS setting. Rajpurkar et al. (2020); Ahn et al. (2022) provide visual explanations when evaluating the usefulness of a DCSS, but they do not look at the effect that XAI explanations had. Gaube et al. (2023) find that SMs improve the diagnosis performance for non-task experts, but they do not compare it to other XAI methods. Tang et al. (2023) look at AI tools for lung nodule detection in chest X-rays. They compare having no AI help, to having just the AI prediction, AI prediction with confidence score, and AI prediction with confidence score and localisation map. They find that while AI prediction helps, neither type of AI with the above forms of explanations (e.g., confidence score, localisation map) leads to any significant improvement over no AI.

Our work is the first to study and compare the effect of different explanation types, and the interaction with advice and explanation correctness, on the complex vision task of chest X-ray analysis.

## 3 Methods

We evaluate the usefulness of SMs, NLEs, and their combination in a clinical decision-support context. We also control for AI advice correctness and explanation correctness (EC). EC captures to what extent the information provided in an explanation is clinically correct. We obtain the ground-truth for both advice and explanation correctness from the annotations of three expert radiologists. EC is rated on a 7-point Likert scale, evaluating both individual and combined explanation effectiveness. The annotator interface to provide these metrics is shown in Figure 24 in the Appendix. The study design is outlined in Figure 1.

### 3.1 Study Overview

Our pre-registered, IRB-approved user study entails both quantitative and qualitative measurements involving 85 clinical participants. The study was developed through iterative pilot studies and consultations with expert clinicians. Our goal is to evaluate the *usefulness* of XAI explanations. We consider usefulness to be the ability of an explanation to help users discern whether a model prediction is correct. A natural way to evaluate usefulness is in a human-AI collaborative setting, i.e., CDSSs.

Our CDSS provides a suggestion for each image, consisting of a single radiographic finding predicted by the AI, i.e., the *AI advice*. To simplify our design, we focus on one finding per image, and communicate to participants that this is not necessarily the only or most important finding.

We study the following four scenarios: (i) No XAI (participants receive the AI model's advice without any explanation), (ii) SM (participants receive the model's advice and an SM), (iii) NLE (participants receive the model's advice and an NLE), (iv) COMB (participants receive the model's advice, an SM, and an NLE).

We consider the case of an imperfect AI and XAI, as we want to explicitly study how *good* or *bad* explanations can help users identify whether the model is correct. We simulated an environment where the model has an accuracy of 70%, to strike a balance between having a reasonable representation of correct and incorrect model predictions and not making the model appear overly unreliable. We also sample image-explanation pairs to ensure that the overall distribution of EC scores is as uniform as possible (to get a good representation of different EC levels). The distributions are shown in the appendix in Figure 9 and 8.

In each of the four randomly shuffled sessions, participants are shown 20 examples, which consist of a chest X-ray, the patient context, the AI advice, e.g., "Pneumonia", and a scenario-specific explanation (see a snapshot of the user interface in Figure 25). They are then asked to express their agreement with the AI advice ("Not present", "Maybe present", or "Definitely present"). We also ask them whether they found the explanation useful in their decision-making (e.g. "How useful was the AI model's explanation in helping you decide whether the AI was right or wrong in suggesting pneumonia."). This encourages them to engage with the explanation and it enables us to quantify the relationship between *perceived* and *actual* explanation usefulness.

To mitigate order effects and user fatigue, we randomize the order of the tasks for each participant, ensuring that each task is equally distributed as the first, second, third, or last. We also enforce three-minute breaks between each session, where we give participants the option to follow a guided meditation. We also emphasize multiple times that the users are engaging with different AI models in each task, to avoid carry-over effects where a person's engagement with explanation type A affects their perception of the CDSS and therefore their subsequent engagement with explanation type B. Finally, we introduce an incentive of doubling the compensation for participants who perform in the top 20%. The aim of this ensure users are dedicated through the 80 examples. At the end of the four tasks, users fill out a post-study survey. Here we ask them about their experience with the different AI explanations and measure how their attitude towards AI has been affected.

### 3.2 Participant Recruitment

As our aim is to study the effect of different explanation types in an imperfect (X)AI setting, we recruit participants with foundational competence in reading chest X-rays, who are knowledgeable enough to not rely wholly on the AI system, but are still likely to engage with the AI's predictions and explanations. Indeed, Gaube et al. (2023) found that increasing expertise in radiology leads to an increased likelihood of dismissing AI suggestions. Furthermore, CDSSs are generally seen as most useful for people who have medical training but are not experts in the task at hand (Bussone et al., 2015). This is particularly relevant in scenarios where there is a scarcity of expert radiologists, and non-expert clinicians benefit from collaborating with AI systems (Gaube et al., 2023).

For the above reasons, our primary target group for this study are medics who have undergone training in reading chest X-rays, but who are not specialist radiologists. All potential participants fill out a screening document, which contains a self-assessment as well a quiz on three chest X-rays that fulfil the medical student curriculum of the Royal College of Radiologists (UK) (an example is shown in Figure 25). These X-rays contain examples of pneumonia, pleural effusion, and lobe collapse, which are the most common classes in our dataset. We then select our final batch of participants based on these forms. In order to determine the sample size, we ran four pilot studies and used the estimated effects to run a power analysis using the model described in 1. We found that 80 participants should provide significant power. We ended up recruiting 85 participants, as we sent out extra invitations to account for dropouts. Our participants range from medicine students to radiology residents (see detailed characteristics in Appendix B. We recruit participants via mailing lists and networks focusing mainly on COUNTRIES ANONYMIZED. Participants are compensated for their time with an voucher worth an equivalent of $38 for the one-hour study. The entire task is conducted online on a custom streamlit platform that we will make publicly available for future use.

### 3.3 Model Implementation

We train a model following the Ratchet architecture (Kayser et al., 2022). It consists of a DenseNet vision encoder (Huang et al., 2017) that generates 7x7 1024-dimensional feature maps of the image. These are then used to perform multi-label image classification, and given as prefixes to a transformer decoder for NLE generation. The NLE is further conditioned on the predicted label. For each positively predicted class an NLE is generated.

The model was trained on the MIMIC-NLE dataset (Kayser et al., 2022). The NLEs are all directly extracted from radiology reports that were recorded during routine clinical practice. Each NLE links a finding to its evidence in a radiographic scan, including details about location, size, severity, certainty, and differential diagnoses. Examples of NLEs are shown in the Appendix in Figure 10. The model obtained a weighted AUC of 0.75. Note that the main purpose wasn't to maximize model performance. Instead, we specifically focus on the case of imperfect AI, where a model, for various reasons, such as limited or biased data, does not perform optimally. This is different from existing work in human-AI collaboration, where they often consider AI models that outperform humans to investigate how they could be used to improve human performance (Tschandl et al., 2020; Fogliato et al., 2022). Nonetheless, our model still performs favorably on existing benchmarks, ensuring that our model and the generated explanations are of a realistic standard (Irvin et al., 2019).

The NLEs that the model generates are learned in a purely supervised way. They, therefore, capture the nuances around assertiveness and the certainty of findings that naturally occur in clinical practice. For this reason, we consider assertiveness an integral part of the NLEs, as opposed to a design factor that can be studied by itself (Calisto et al., 2023).

For SMs, we implement Grad-Cam (Selvaraju et al., 2017) following Gildenblat and contributors (2021). We ran it on our model trained for both image classification and NLE generation. We chose Grad-CAM as it is widely used and previous work has shown that out of the commonly used saliency techniques, it is the most accurate one for medical imaging (Saporta et al., 2022). We have also qualitatively verified it by comparing it to Grad-Cam++, HiResCam, AblationCAM, and XGradCAM.

### 3.4 Data Selection

In this section, we describe how we obtained the set of 80 images used in our study.

#### 3.4.1 Acquiring AI Predictions

We used a multi-label classification AI trained on the MIMIC-CXR dataset, which assigns a logit to each of the 10 classes. We established thresholds for each class by maximizing the Youden Index to optimize the balance between sensitivity and specificity. The selected classes for our study—pneumonia, atelectasis, pulmonary edema, fluid overload/heart failure, aspiration, and alveolar haemorrhage—were chosen for their clinical significance and detectability in chest X-rays alone, after consultations with radiologists.

#### 3.4.2 Expert annotation

Even though our chest X-rays are paired with radiology reports, we follow existing work (Gaube et al., 2023; Ahn et al., 2022; Seah et al., 2021) and have three experienced radiologists annotate our AI advice and explanations.

Radiologists classify each AI-predicted finding as *Not present*, *Maybe present*, or *Definitely present*, based on established medical imaging standards. They also rate the correctness of textual and heatmap explanations on a 7-point Likert scale, evaluating both individual and combined explanation effectiveness. The majority vote determines the advice correctness, while explanation correctness scores are averaged and mean-centered. More details, including the user interface used by our annotators, are shown in Appendix I.

#### 3.4.3 Selecting the study examples

From the annotated set, we carefully selected 80 images, ensuring a similar distribution of correct and incorrect AI predictions across all our classes. We also excluded ambiguous cases with significant annotator disagreement. Additionally, we sample examples such that the distribution of EC scores is as uniform as possible.

For our selected sample we obtain pairwise kappa scores of $0.451$, $0.458$, and $0.502$ between the three annotators (grouping "Maybe present" and "Definitely present" as positive). Note that if we leave out "Maybe present" votes, we get perfect kappa scores because of the above exclusion criteria. Further details on our selected samples are given in the Appendix C.

#### 3.4.4 Distributing examples across participants and tasks

These 80 images were evenly distributed across four tasks and multiple participants, ensuring each image was equally represented across all tasks. This method prevents task-specific biases and maintains a consistent 70% accuracy rate for AI advice across different explanation types.

## 4 Results

### 4.1 Statistical Model

We model our results using a Generalized Linear Mixed-Effects Model that predicts human accuracy for each instance. The model is given below:

$$\begin{aligned} l_{ij} =& \beta_0 \\ & + \beta_a * \text{(AC)} \\ & + \beta_t * \text{(Explanation Type)} \\ & + \beta_{t \times a} * \text{(Expl. Type)} \times \text{(AC)} \\ & + \beta_{t \times e} * \text{(Expl. Type)} \times \text{(EC)} \\ & + \beta_{t \times e \times a} * \text{(Expl. Type)} \times \text{(EC)} \times \text{(AC)} \\ & + u_{Participant} \\ & + u_{Image} \end{aligned} \tag{1}$$

This model predicts the log-odds of the human accuracy $l_{ij}$ for the $i$-th participant on the $j$-th image. As fixed effects, we consider *advice correctness* AC (i.e., whether AI advice is correct or not), *explanation type* (None, NLE, SM, and combined), *explanation correctness* EC and different interactions of these effects. As random effects, we include the participants (which can have different skill levels) and the images (which can have different difficulty levels). A rationale for the different interaction terms is given below:

- (Explanation Type) × (AC): We are assuming that different explanation types have a different impact on human accuracy when advice is correct or incorrect. For example, explanation types prone to confirmation bias will have a particular effect when the advice is incorrect.

- (Expl. Type) × (EC): Note that we do not include (EC) as a main effect. This is because (EC) between different explanation types is not directly comparable (e.g. NLEs contain more specific information and therefore can contain both more correct information and more false information). Therefore we consider (Explanation Correctness) as a type-specific metric and need to include the interaction term.

- (Expl. Type) × (EC) × (AC): We need to model this interaction as (EC) is strongly correlated to (AC) (the (EC) scores for incorrect advice are much lower).

We test the model statistically and find that both random and fixed effects should be included. In particular, we perform a likelihood ratio test (LRT) between the model in (1) and a baseline model disregarding explanation correctness and interactions and find that the full model yields significantly better fit $\chi^2_{12} = 28.21, p = .005$ (see Appendix A).

### 4.2 Main Hypotheses

Our main goal is to understand how different explanation types affect human accuracy, which is our proxy for explanation usefulness. More specifically, we are interested in how explanation and advice correctness factor into this. In the context of imperfect XAI, we consider the following classification of EC. Qualitative examples representing the different subtypes are given in Figure 10

- Explanations are *insightful* when their correctness aligns with advice correctness: *Convinc-*

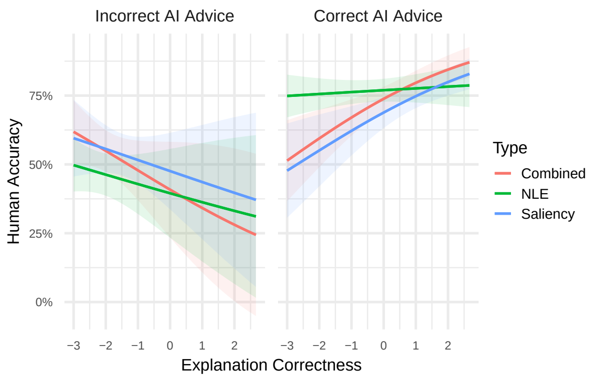


Figure 2: Human Accuracy given AC and EC.

*ing* explanations are correct when the AI advice is correct; *Revealing* explanations are incorrect when the AI advice is incorrect.

- Explanations are *deceptive* when their correctness misaligns with advice correctness: *Misleading* explanations are correct when the AI advice is incorrect; *Confusing* explanations are incorrect when the AI advice is correct.

**EC needs to *align* with AC:** Our results show that insightful explanations, i.e., where EC aligns with AC, are helpful in the decision-support setting. Figure 2 illustrates how higher EC scores harm human accuracy when the AI prediction is incorrect (*deceptive* explanations) and benefits human accuracy when the AI advice is correct (*insightful* explanations). These effects are less strong for NLEs than for the visual methods.

In Figure 3 we look at human accuracy by explanation type for the four EC scenarios described earlier. We consider high EC to be the upper half of EC scores by explanation type, and low EC is the lower half.

We observe that as a general trend human accuracy is harmed when explanations are deceptive, and people would be better off using no explanation. For SMs, human accuracy goes down 4.9% ($p < .05$) when AC and EC don't align. For combined explanations, it goes down 3.9% ($p = .06$). On the contrary, for insightful EC scores, human accuracy goes up 4.3% ($p < .005$) for combined explanations. These effects are not seen for NLEs, suggesting that the visual explanations are more helpful to users to discern whether an AI's decision-making is flawed.

**When insightful, combine SM and NLE:** For insightful explanations, combining SMps and NLEs provides significant improvements compared to the other conditions: 6.3% ($p < .005$) against No XAI, 7.1% ($p < .005$) against NLEs, and 4.5% ($p < .05$) against SM. This suggests that participants can integrate the information from both visual and textual cues to identify when an AI is wrong or right. Interestingly, even though insightful NLEs on their own are worse than "No AI", combining them with visual explanations leads to a significant boost.

**NLEs on their own lead to overreliance:** Across AC and EC scores, differences between our four conditions cancel each other out and we observe no significant differences (see Figure 17 in the Appendix. However, in the case of incorrect advice, there is a significant drop in human accuracy for NLEs compared to combined (-7.3%, $p < 0.05$) and SM (-6.2%, $p < 0.05$). This suggests that NLEs make people more likely to agree with the AI when it is incorrect. Especially when EC is comparatively high but the AI advice is incorrect, people are 10.1% ($p < 0.05$) more likely to agree with the AI than without explanation. This also means that for the scenario of correct advice and comparatively low EC explanations, NLEs lead to higher performance (6.6%, $p < 0.05$ versus SAL and 5.7% $p < 0.05$ versus combined), as people are more likely to agree with low EC NLEs. Overall, people agree with the AI 67.3% of the time when it's accompanied by an NLE, compared to 63.8% on average for the other explanation types. This aligns with our survey results, which show a clear user preference for NLEs, as well as the perception that the NLE model was the most correct one (participants were not aware that they all have the same share of correct/incorrect advice). This could suggest that the assertiveness (Calisto et al., 2023) and/or human-like (Breum et al., 2024) nature of NLEs could lead people to overly trust and rely on AI.

#### 4.2.1 Additional Results

In further analyses, we study the time participants require to reach a diagnostic decision (decision speed), their decision confidence and the perceived helpfulness of different explanation types. We find that with increasing complexity of explanations (NLE > Saliency > No XAI), participants require more time to reach a decision. Further, we find that the measured confidence is similar across explanation types, but increases significantly as explana-

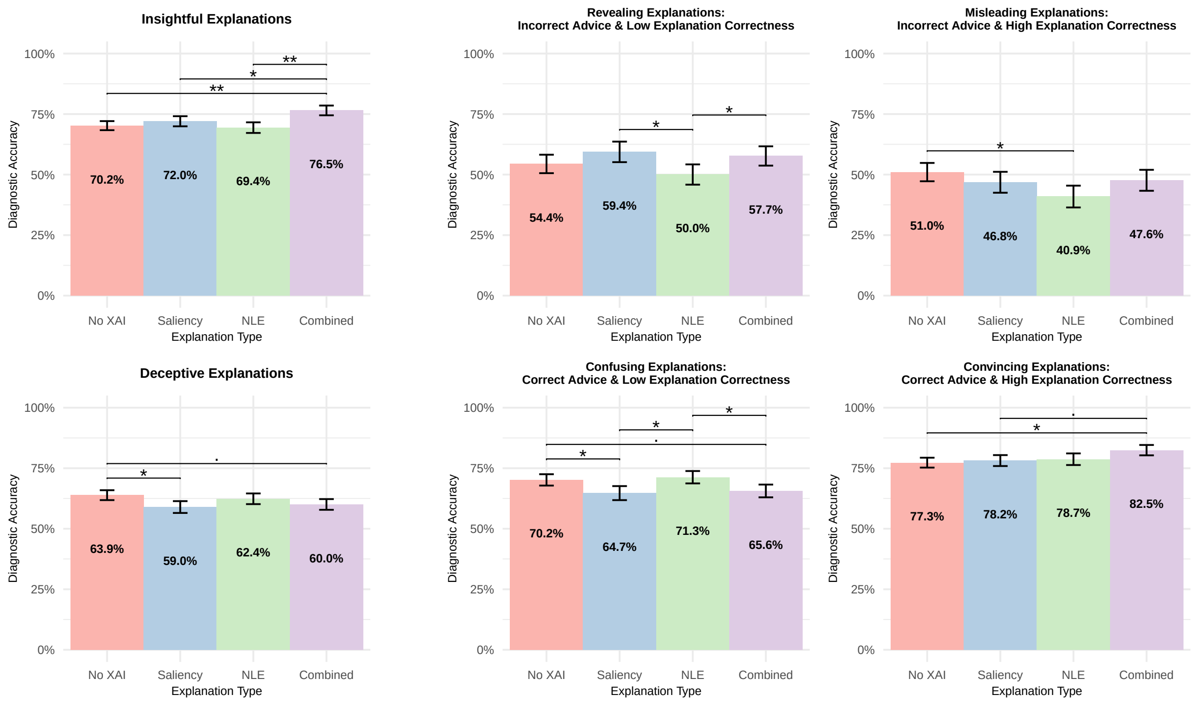


Figure 3: Main results. The error bars represent standard errors. $.p < 0.1$, $*p < 0.05$, $**\,p < 0.01$, statistically non-significant are left unmarked.

tions get more *insightful*. Finally, we observe that higher quality NLEs are rated as more useful and we find an effect of perceived usefulness on the diagnostic accuracy that resembles that of confidence. We discuss results in more detail in Appendix **??**.

### 4.3 Post-Survey Insights

In our post-task survey, we ask users about their experience with the different explanation types. There is a strong tendency towards preferring NLEs the most, and saliency maps the least, as shown in Table 1. Participants also perceive the model with saliency maps to be on average 17% less accurate than the model with NLEs. This confirms our finding that users overestimate (and therefore overrely) on the model with NLEs. They deem the model with saliency maps as more inaccurate, but perhaps that caution is warranted given the artificially flawed model. Participants also evaluated each explanation type across five key characteristics (the exact questions can be found in Appendix F) of explanations, with NLEs scoring the highest on all 5 (Figure 4).

Table 1: Ranking of models.

| | $\mu$Rank | #1 | #2 | #3 | #4 |
|---|---|---|---|---|---|
| NLE | 1.85 | 38.9% | 38.9% | 20.0% | 2.2% |
| Comb. | 2.05 | 40.0% | 23.3% | 27.8% | 8.9% |
| No XAI | 2.98 | 14.4% | 21.1% | 16.7% | 47.8% |
| SM | 3.11 | 6.7% | 16.7% | 35.6% | 41.1% |

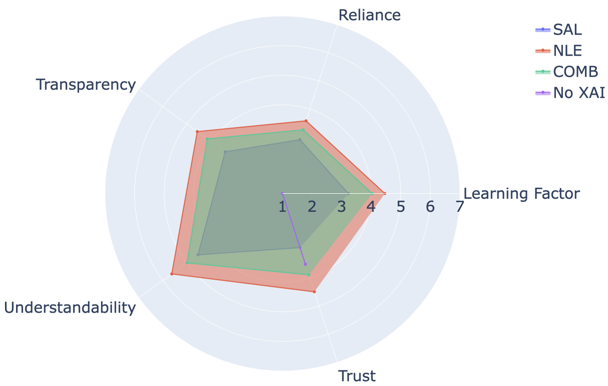


Figure 4: Five attributes of explainability methods, ranked on a 7-point Likert scale.

## 5 Summary and Outlook

In this work, we conducted a large user study to assess the usefulness of SMs, NLEs, and their combination in a clinical setup with imperfect AI and XAI. We showed that EC and its alignment with AC are significantly affecting the usefulness of explanations. Textual explanations alone are prone to lead to overreliance, but joint with saliency maps are showing the most promise.

## Limitations

The present study presents a distinct insight into how users engage with AI explanations in a specific scenario. We aim to evaluate imperfect AI and imperfect XAI explanations in a clinical decision-support setting, rather than validating a clinical end product. It provides a snapshot, rather than a longitudinal study, leaving unexplored how interaction with models and explanations change over time. Similarly, the data used in this study consists of chest X-rays in a limited number of classes, hence more research is needed to understand how generalizable the results are for other classes and types of X-rays. It is worth noting that recruitment biases such as self-selection can impact the participants who chose to engage in this study. Methodologically, to mitigate order effects and fatigue, we implemented breaks between sessions and clearly stated that participants interacted with a different AI in each session. Additionally, to incentivize performance, we announced beforehand that the top 20% of participants completing the survey would gain double earnings.

## References


Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging tests for model explanations. *Advances in Neural Information Processing Systems*, 33:700–712.

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Working Paper 31422, National Bureau of Economic Research.

Ajay Agrawal, Joshua S Gans, and Avi Goldfarb. 2019. Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, 33(2):31–50.

Jong Seok Ahn, Shadi Ebrahimian, Shaunagh McDermott, Sanghyup Lee, Laura Naccarato, John F Di Capua, Markus Y Wu, Eric W Zhang, Victorine Muse, Benjamin Miller, et al. 2022. Association of artificial intelligence–aided chest radiograph interpretation with reader performance and efficiency. *JAMA Network Open*, 5(8):e2229289–e2229289.

Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 275–285.

Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M Friedrich, and Felix Nensa. 2023. Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches. *European journal of radiology*, page 110786.

Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 152–163.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE.

Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C Nascimento. 2023. Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*.

Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. 2022. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems*, 35:2832–2845.

Yuhan Du, Anna Markella Antoniadi, Catherine McNestry, Fionnuala M McAuliffe, and Catherine Mooney. 2022. The role of xai in advice-taking from a clinical decision support system: A comparative user study of feature contribution-based and example-based explanations. *Applied Sciences*, 12(20):10323.

Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, pages 449–466. Springer.

Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing human-centered perspectives in explainable ai. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pages 1–6.

Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who goes first? influences of human-ai workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1362–1374.

Maria Frasca, Davide La Torre, Gabriella Pravettoni, and Ilaria Cutica. 2024. Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. *Discover Artificial Intelligence*, 4(1):15.

William Gale et al. 2018. Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv preprint arXiv:1806.00340*.

Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias FC Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, et al. 2023. Non-task expert physicians benefit from correct explainable ai advice when reviewing x-rays. *Scientific reports*, 13(1):1383.

Marzyeh Ghassemi, Mahima Pushkarna, James Wexler, Jesse Johnson, and Paul Varghese. 2018. Clinicalvis: Supporting clinical task-focused design evaluation. *arXiv preprint arXiv:1810.05798*.

Jacob Gildenblat and contributors. 2021. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam.

Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *SIGKDD*.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? *arXiv preprint arXiv:2010.04119*.

Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. 2021. This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. *arXiv preprint arXiv:2105.02968*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Jeremy Irvin et al. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. 33(01).

Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35.

Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. 2022. Explaining chest x-ray pathologies in natural language. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 701–713. Springer.

Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. Hive: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*, pages 280–298. Springer.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *ICML*.

Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12.

Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Curtis P Langlotz. 2019. Will artificial intelligence replace radiologists?

Curtis P. Langlotz et al. 2019. A roadmap for foundational research on artificial intelligence in medical imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. *Radiology*, 291(3).

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19.

Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 147–159.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. 2021. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267.

Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2023. The impact of imperfect xai on human-ai decision-making. *arXiv preprint arXiv:2307.13566*.

Ziad Obermeyer et al. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464).

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, et al. 2020. Chexaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv. *NPJ digital medicine*, 3(1):115.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. 2020. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology: Artificial Intelligence*, 2(3).

Noelia Rivera-Garrido, MP Ramos-Sosa, Michela Accerenzi, and Pablo Brañas-Garza. 2022. Continuous and binary sets of responses differ in the field. *Scientific reports*, 12(1):14376.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2021-03-20. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *arXiv preprint arXiv:*.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. 2022. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878.

James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your ai: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 240–251.

Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. 2021. Using StyleGAN for visual interpretability of deep learning models on medical images. *arXiv preprint arXiv:2101.07563*.

Jarrel C. Y. Seah, Cyril H. M. Tang, Quinlan D. Buchlak, Xavier G. Holt, Jeffrey B. Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F. Lambert, Ben Hachey, Stephen J. F. Hogg, Benjamin P. Johnston, Christine Bennett, Luke Oakden-Rayner, Peter Brotchie, and Catherine M. Jones. 2021. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multi-reader multicase study. *The Lancet Digital Health*, 3(8):e496–e506.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization.

Hua Shen and Ting-Hao Huang. 2020. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 168–172.

Jennifer SN Tang, Jeffrey KC Lai, John Bui, Wayland Wang, Paul Simkin, Dayu Gai, Jenny Chan, Diane M Pascoe, Stefan B Heinze, Frank Gaillard, et al. 2023. Impact of different artificial intelligence user interfaces on lung nodule and mass detection on chest radiographs. *Radiology: Artificial Intelligence*, 5(3):e220079.

Armin W. Thomas, Hauke R. Heekeren, Klaus-Robert Müller, and Wojciech Samek. 2019. Analyzing Neuroimaging Data Through Recurrent Deep Learning Models. *Frontiers in Neuroscience*, 13.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234.

Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. 2022. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470.

Effy Vayena, Alessandro Blasimme, and I. Glenn Cohen. 2018. Machine learning in medicine: Addressing ethical challenges. *PLoS medicine*, 15(11).

Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:*.

Fan Yang, Mengnan Du, and Xia Hu. 2019. Evaluating explanation without ground truth in interpretable machine learning. *arXiv preprint arXiv:1907.06831*.

Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

## A Model Selection

Here we provide details on the statistical model we used to analyze our main results. The statistical model was selected based on the nature of the task and experiment design at hand and then verified using inferential statistics.

To establish the significance of our main model (1), we compare it against a baseline model that disregards explanation types. The model equation is as follows:

$$\begin{aligned} l_{ij} = &\beta_0 \\ &+ \beta_a * \text{(Advice Correctness)} \\ &+ u_{Participant} \\ &+ u_{Image} \end{aligned} \tag{2}$$

**Fixed Effects.** We first select fixed effects while including random effects. As reported in the main paper, we use an LRT to test whether the added variables improve model fit. We further find the AIC (Akaike Information Criterion) is improved: 5504.3 to 5500.1.

**Random Effects.** The study design strongly suggests the inclusion of random effects $u_{Image}$ and $u_{Participant}$ as these introduce dependencies between observations. For both models, we study the random effect variances and compare the model with and without its random effects. For the baseline model (2) we find that $Var(u_P) = 0.056$ and $Var(u_I) = 0.400$. Further, the LRT is significant suggesting the inclusion of random effects: $\chi^2_2 = 227.86$, with $p < .0001$. We repeat this analysis for the full model (1). We find $Var(u_P) = 0.059$ and $Var(u_I) = 0.295$, which are qualitatively $> 0$. The LRT comparing this model with and without random effects is significant, $\chi^2_2 = 144.43$, $p < .0001$. In addition, we test incrementally only including $u_{Image}$ in comparison to a model with both random effects. Analysis of both models suggests that $u_{Participant}$ should be included. Hence, we only consider models with both random effects included.

## B Selected Participants

We provide descriptive information on the 85 participants included in this study in Figures 5, 6, and 7.

## C Study X-ray sample

In this section, we provide additional data on the process of annotating X-rays and sampling the set

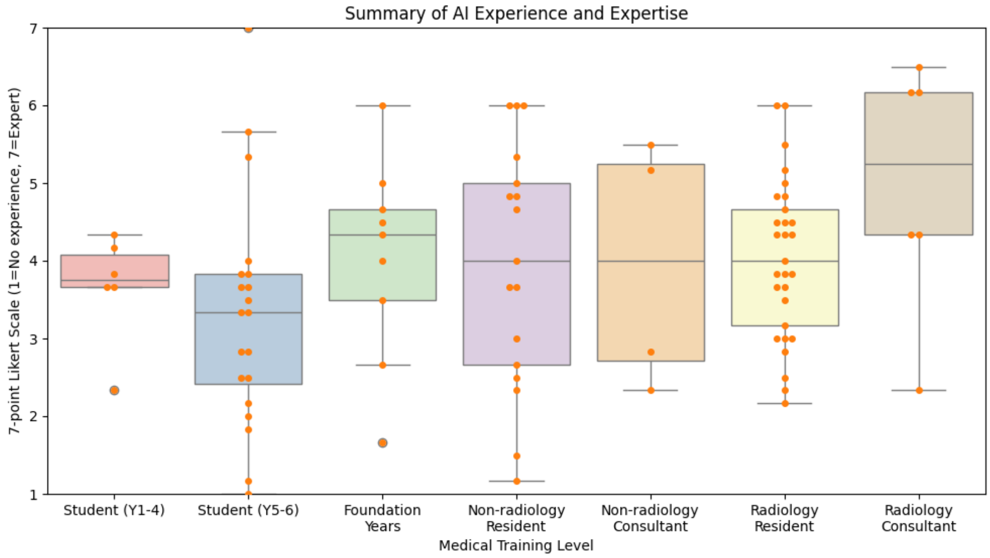


Figure 5: Self-assessed levels of experience and expertise in computer vision, NLP, explainable AI, and clinical decision-support systems.

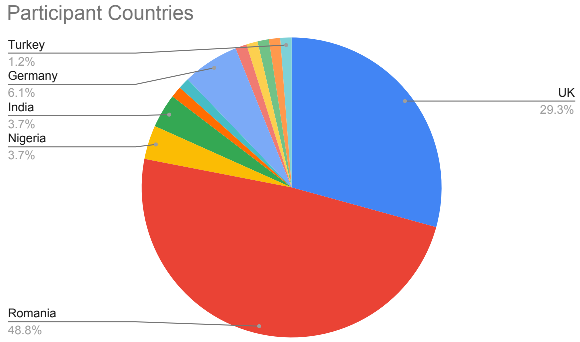


Figure 6: Countries where participants have spend the most time "studying or practicing" medicine.

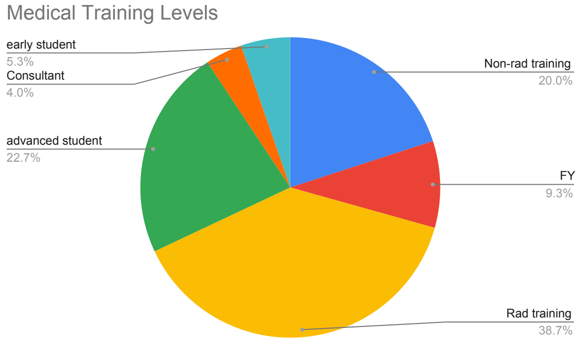


Figure 7: Medical Training Level of Participants.

of 80 scans included in this study.

# D Qualitative Examples

Figure 10 contains representative examples showcasing how Explanation Correctness (EC) affects clinicians' diagnostic accuracy. Each scenario includes the original chest X-ray (left) and the X-ray overlaid with a saliency map (right), along with the corresponding AI advice, Natural Language Explanation (NLE), mean EC scores, and the participants' overall average diagnostic accuracy for that image given different explanation types.

# E Participant Behavior Analysis

This section (Figures 11 to 16 contains further insights into participant behavior performance.

# F Participant Survey

## F.1 Questions about level of AI expertise

Participants have to agree to each of the following statements on a 7-point Likert scale from "Strongly Disagree" to "Strongle Agree".

- I understand the principles behind computer vision models (i.e., AI algorithms used for analysing images) and how they work.
- I am familiar with language models (i.e. AI algorithms used to understand and generate language) and how they work.
- I understand the concepts of explainable AI (XAI), i.e., methods that try to make AI algorithms' decision-making more transparent (for example: heatmaps).

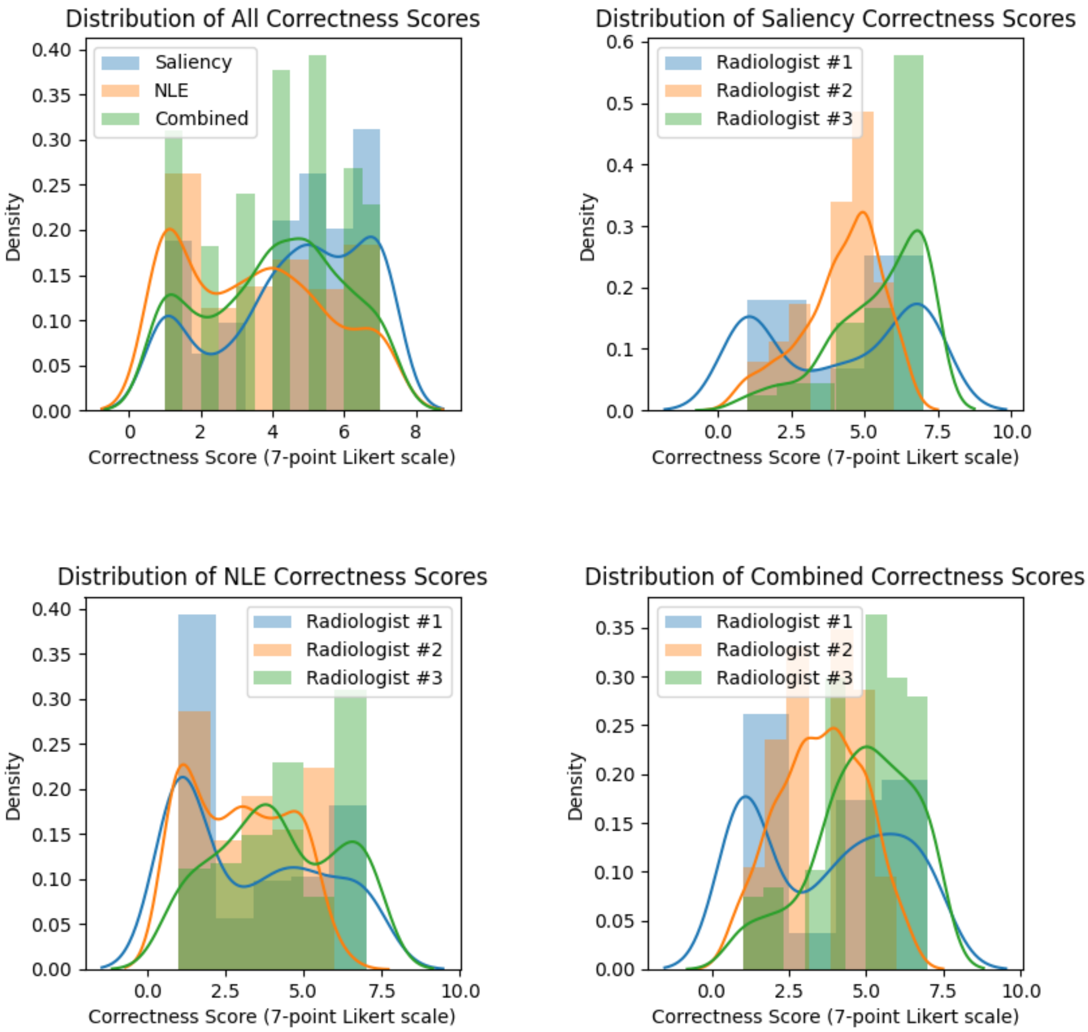


Figure 8: The graphs show the distribution of explanation correctness scores assigned to the different explanation types. In total, 3 explanations (NLE, SM, COMB) were annotated for 160 images.

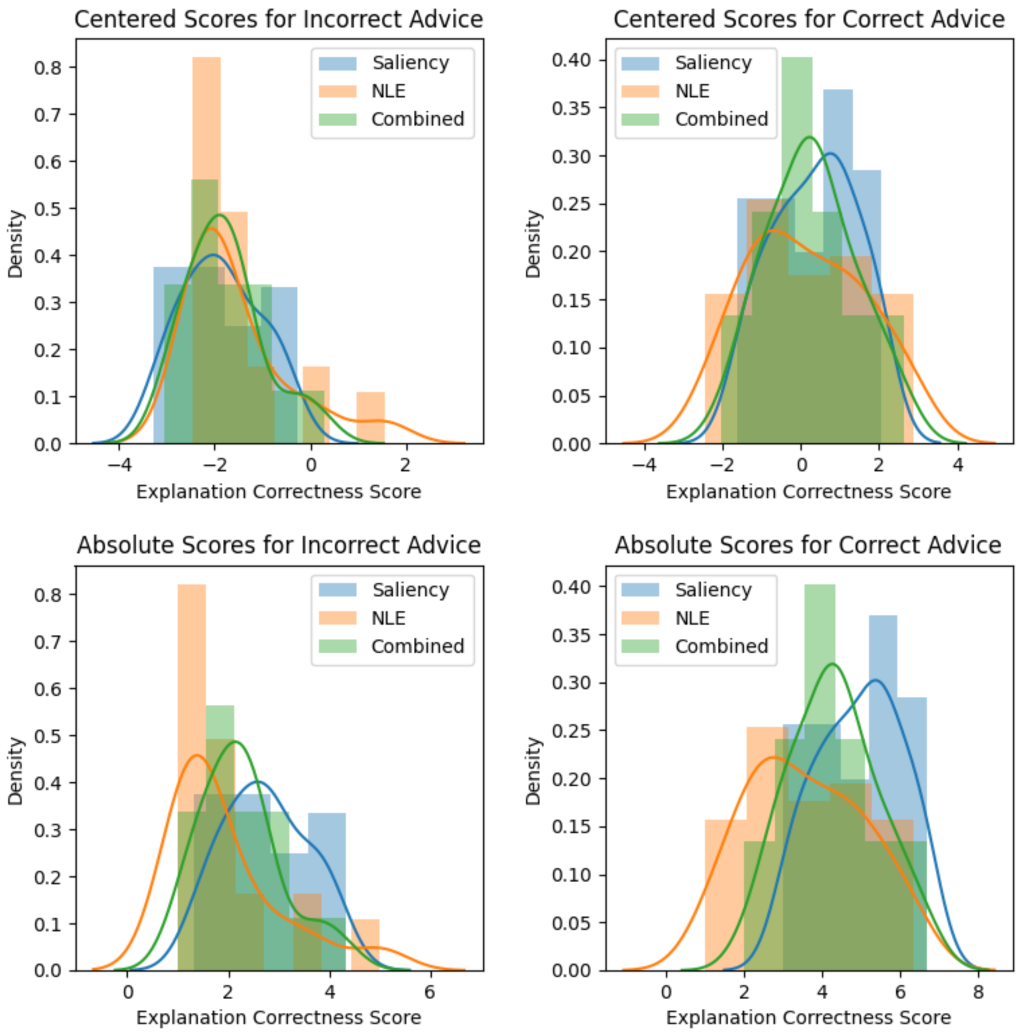


Figure 9: An illustration of the distribution of explanation correctness scores included in the study. The images were selected to ensure that the distribution is as uniform as possible (representing all EC scores equally). It can be seen that annotators assigned higher EC values to SM compared to NLE.

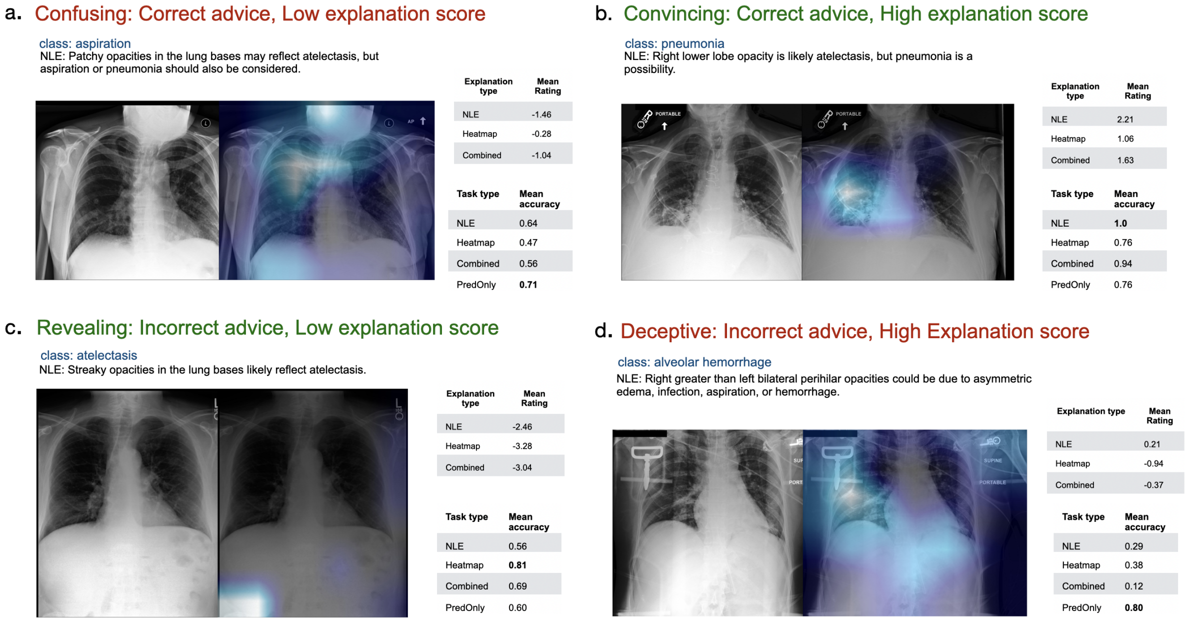


Figure 10: (a) *Confusing* (Correct advice, Low explanation score): The AI correctly identifies aspiration but provides a poorly rated explanation, leading to lower diagnostic accuracy compared to relying on the AI prediction alone. (b) *Convincing* (Correct advice, High explanation score): The AI correctly identifies pneumonia and provides a highly rated explanation, resulting in high diagnostic accuracy. (c) *Revealing* (Incorrect advice, Low explanation score): The AI incorrectly suggests atelectasis, but the poorly rated explanation helps clinicians identify the error, leading to higher accuracy compared to relying on the AI prediction alone. (d) *Deceptive* (Incorrect advice, High explanation score): The AI incorrectly suggests alveolar haemorrhage and provides a highly rated yet misleading explanation, leading clinicians to agree with the incorrect prediction and resulting in the lowest diagnostic accuracy.

- I regularly use AI-powered chat tools (e.g. ChatGPT).
- I regularly interact with methods that make AI algorithms' decision-making more transparent.
- I regularly use AI-based decision-support tools for medical imaging.

### F.2 Questions about attitude towards AI

Below are the 9 statements that were used to evaluate participants' attitude towards AI in terms of trust, ethical concern, and performance expectation. We use the same Likert scale as above.

Trust

- I'm not comfortable using an AI if I don't fully understand how it makes a decision.
- The use of AI should always be accompanied by the option for human review and intervention.
- I trust AI-based recommendations as much as those from human experts in a clinical setting.

**Ethical Concerns**

- I am not concerned about the ethical implications of using AI in healthcare.
- Due to the dangers of AI, its adoption should be minimised.
- The development of AI in healthcare should be tightly regulated.

**Performance Expectations**

- It won't take long until AI will drastically transform healthcare.
- AI in its current form is still far from being ready to be used in clinical practice.
- I believe AI can improve the accuracy of diagnoses in healthcare.

### F.3 Explanation Type Feedback Questionnaire

To capture participants' objective feedback of explanation types we asked the following questions for each type (only the "trust" question for "No XAI")

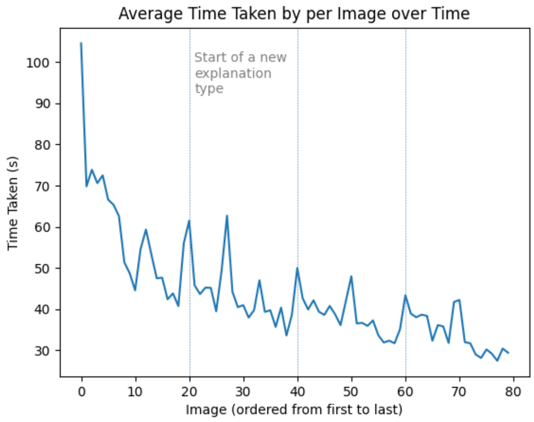


Figure 11: This plot shows the average decision speed (time taken per image) and how it changed over time. The overall trend is that participants become faster over time. We can also see spikes at the start of each new task, when they are introduced to a new explanation type.

- I trusted this AI.
- The explanations that were provided for the diagnoses were difficult to understand.
- It was transparent to me how the AI came to a diagnosis.
- I didn't rely on the AI's explanations to decide whether I agree with the diagnosis or not.
- I have learned something from the AI's explanations and they helped me become more proficient in reading chest X-rays.
- How accurate do you think this AI was (in %)?

For all but the last question users had to respond on the same 7-point Likert scale as described above.

# G Additional Results

In Figure 17 we show the effect of explanation types (given correct and incorrect advice) on human accuracy.

# H Exploratory Analysis

## H.1 Perceived Usefulness

**Hypotheses.** Participants report the *perceived usefulness* of all explanations. We seek to understand the association of this perceived usefulness with the *actual* usefulness, measured by differences the diagnostic accuracy. Further, we wish to understand if some explanation types are perceived as more useful than others. Finally, we are interested in the effect of explanation quality on the perceived usefulness.

**Modeling.** We model human accuracy by augmenting our main model (1) with the perceived usefulness and its first-order interaction effects:

$$
\begin{aligned}
l_{ij} = &\beta_0 \\
&+ \beta_a * (\text{AC}) \\
&+ \beta_t * (\text{Explanation Type}) \\
&+ \beta_p * (\text{Perceived Usefulness}) \\
&+ \beta_{t\times a} * (\text{ET}) \times (\text{AC}) \\
&+ \beta_{t\times e} * (\text{ET}) \times (\text{EC}) \\
&+ \beta_{p\times a} * (\text{PU}) \times (\text{AC}) \\
&+ \beta_{p\times e} * (\text{PU}) \times (\text{EC}) \\
&+ \beta_{p\times t} * (\text{PU}) \times (\text{ET}) \\
&+ \beta_{t\times e\times a} * (\text{ET}) \times (\text{EC}) \times (\text{AC}) \\
&+ u_{Participant} \\
&+ u_{Image}.
\end{aligned} \tag{3}
$$

We find this model yields significantly better model fit than our main model (1) indicating that the perceived usefulness adds above and beyond the observed effects based on the explanation correctness (and other variables), $\chi^2_4 = 40.923$, $p < .0001$.

**Perceived usefulness increases with explanation quality.** We find that the perceived usefulness increases with an increasing explanation correctness for NLEs and by extension for combined explanations (see Figure 18). However, this trend is not visible for saliency maps, which is a surprising finding.

**Perceived usefulness interacts with advice correctness.** We use model (3) to study the effect of perceived usefulness on the diagnostic accuracy and find that such effect is present, albeit heavily moderated by the correctness of the advice. Interestingly, when AI advice is incorrect, higher perceived usefulness is associated with worse diagnostic accuracy as participants fail to detect that the explanation is misleading. This effect resembles that of the explanation quality. It noteworthy though that the misleading nature of deceptive explanations *does* indeed translate from explanation correctness into self reported measures of perceived usefulness. Beyond this joint effect of advice cor-

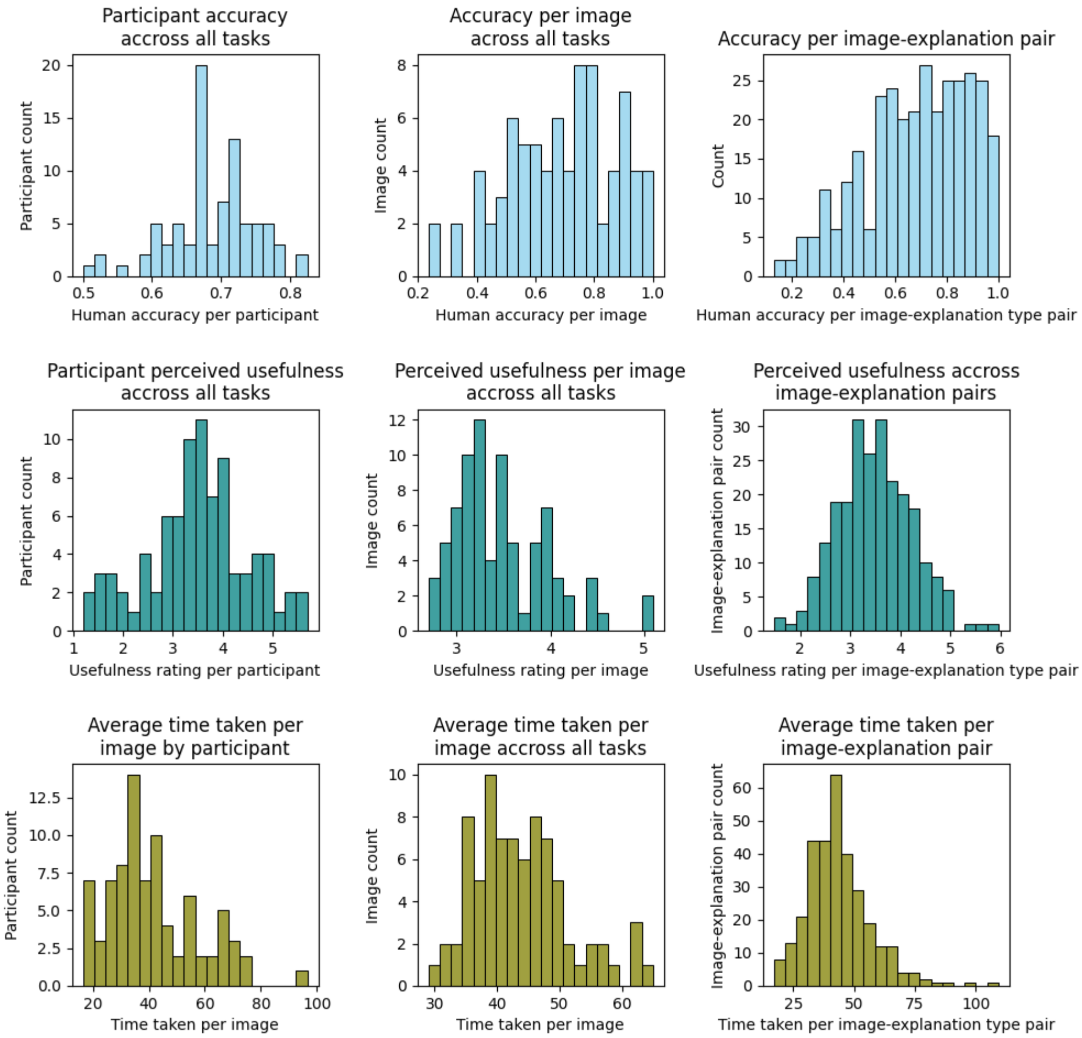


Figure 12: This 3x3 plot illustrates the distributions of accuracies, perceived usefulness, and decision speed by: participant, image, and image-explanation pairing.

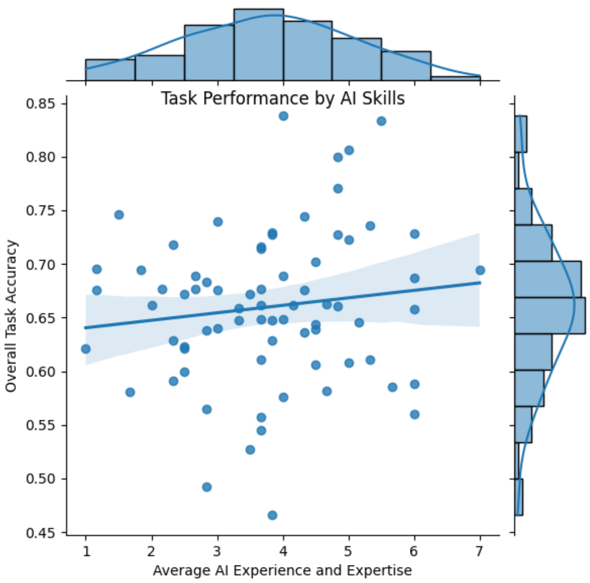


Figure 13: A participant's AI experience and understanding compared to their diagnostic accuracy across all tasks.

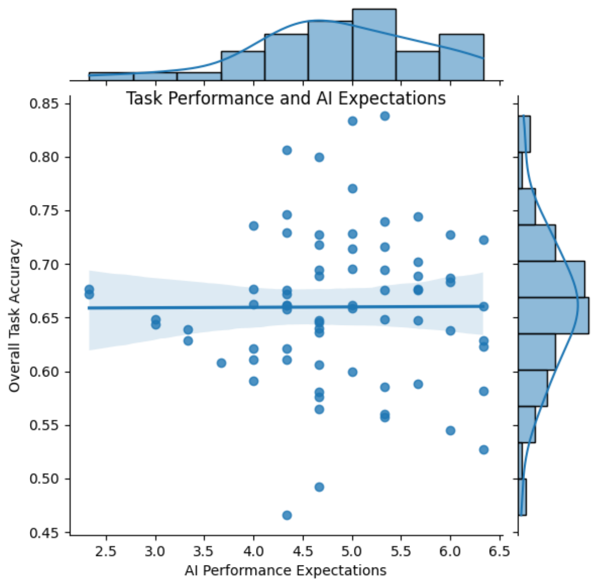


Figure 14: A participant's expectation of AI compared to their diagnostic accuracy across all tasks.

rectness and perceived usefulness, we do not see a clear trend between different types of explanations.

### H.2 Confidence

We study agreement confidence, which we define as the share of participants deeming a finding as "Maybe present" ( *low confidence*) or "Definitely present" (*high confidence*).

**Modeling.** We model whether participants indicate "Definitely present" (Confidence $=1$) or "Maybe present" (Confidence $=0$) using a binomial generalized linear mixed model:

$$\begin{aligned}\eta_{ij} =& \beta_0 \\ &+ \beta_t * (\text{Explanation Type}) \\ &+ \beta_{ac} * (\text{Advice Correctness}) \\ &+ \beta_{ec} * (\text{Explanation Correctness}) \\ &+ \beta_{ec \times ac} * (\text{AC}) \times (\text{EC}) \\ &+ u_{Participant} \\ &+ u_{Image}\end{aligned} \tag{4}$$

where $\eta_{ij}$ are the log-odds of confidence for participant $i$ and image $j$. We compare (4) against a null model

$$\eta_{ij} = \beta_0 + u_{Participant} + u_{Image} \tag{5}$$

and find, our model is significant, $\chi^2_6 = 13.454$, $p = 0.036$.

**Confidence increases with insightful explanations.** We distinguish between *insightful explanations* and *deceptive explanations*. The former are *high quality* explanations for correct advice, as well as, *low quality* explanations for incorrect advice, as they reveal the poor model workings. A deceptive explanation is *high quality* for incorrect advice and *low quality* for correct advice. As presented in Figure 21, we find that *deceptive explanations* are associated with *low* confidence. With increasing insightfulness of the explanations, confidence increases.

**Explanation Types do not predict confidence.** An interesting question is whether some types of explanations are associated with higher agreement confidence as reported by participants. As can seen in Figure 21, there is no statistically significant evidence supporting this. While there is no variation for correct advice, NLEs are associated with higher confidence ratings than combined explanations ($\approx$ 12%). However, this difference is not significant.

### H.3 Efficiency

We study the time participants require to make a diagnostic decision based on the presented information. Besides the diagnostic accuracy, the time taken to examine a radiological study and reach a decision is an important metric as it influences the cost and efficiency of the diagnostic procedure.

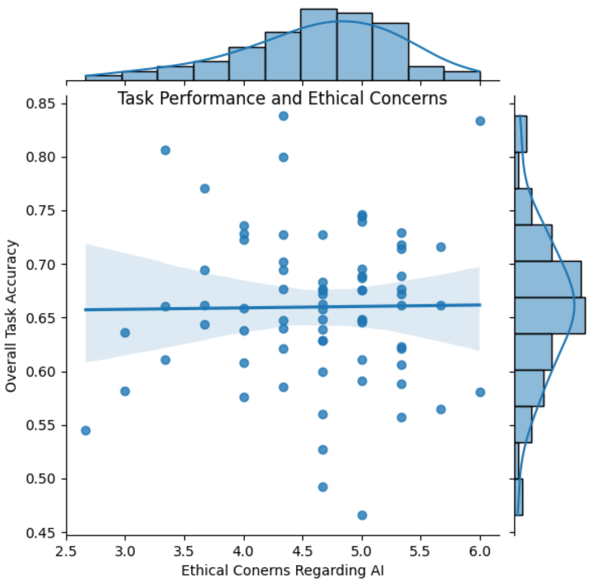


Figure 15: Participant's level of ethical concerns regarding AI compared to their diagnostic accuracy across all tasks.

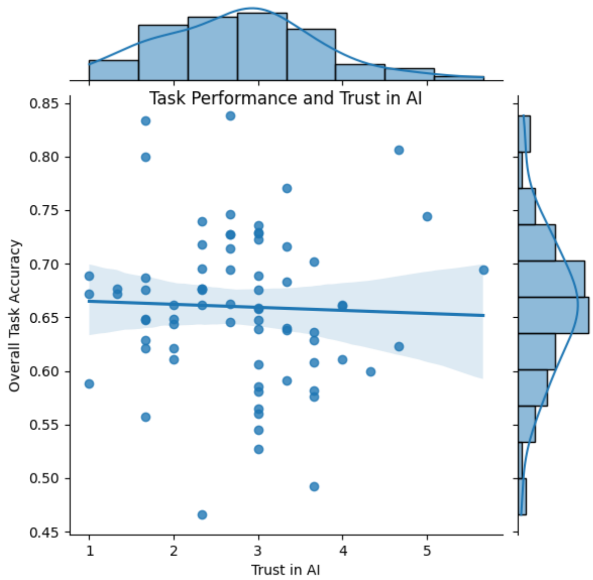


Figure 16: A participant's trust in AI compared to their diagnostic accuracy across all tasks.

The median time taken per study is 35.05 seconds with an inter-quartile range of $[24.25, 55.24]$. As some users might have paused the experiment (evident in very few, very long time intervals), the time taken per study does not necessarily measure the time required to reach a diagnostic decision. Hence, we decide to limit our analysis to observations below 5 min. This excludes 0.6% of observations.

**Modeling.** We use a Gamma Linear Mixed Model to answer our hypotheses in regards to the decision time. As decision times are still over-dispersed, we model the $\log\log$ Decision Time. We build our model as

$$\begin{aligned}\eta_{ij} =&\beta_0 \\ &+ \beta_t * (\text{Explanation Type}) \\ &+ u_{Participant} \\ &+ u_{Image}\end{aligned} \tag{6}$$

and compare against the null model

$$\eta_{ij} = \beta_0 + u_{Participant} + u_{Image}. \tag{7}$$

We find that the larger model fits the data better $\chi^2_3 = 47367.00$, $p < .0001$ and hence base our analysis upon this.

**Hypotheses.** We aim to investigate two hypotheses.

1. Does the complexity of the type of explanations predict the time required to reach a diagnostic decision?

2. Does the explanation correctness influence the decision speed? In particular, we expect higher quality explanations to *increase* speed when the advice is correct. We also expect higher quality explanations to *decrease* speed when advice is incorrect, as *conflicting*, deceptive information are shown.

**Complexity reduces decision speed.** We model the decision speed (as described above) and obtain 95% confidence intervals for the adjusted means as shown in Figure 22. We observe that the most complex explanations (NLE and combined) reduce decision speed by 8s per image (26.8%). Saliency maps reduce the decision speed by only 4s (13.8%). All pairwise comparisons are significant with $p < .001$ with the exception of combined explanations and NLEs (Bonferroni-Holm adjusted, log-log domain). One could argue that the help provided by the explanations reduces the decision times. However, we find that the additional time spent on processing the explanations outweighs such effect - if present: With the increasing complexity of the explanation, the decision speed reduces substantially. The main factor seem to be the NLEs ($t_{\text{Combined}} \approx t_{\text{NLE}}$ and $t_{\text{NLE}} > t_{\text{Saliency}}$ ).

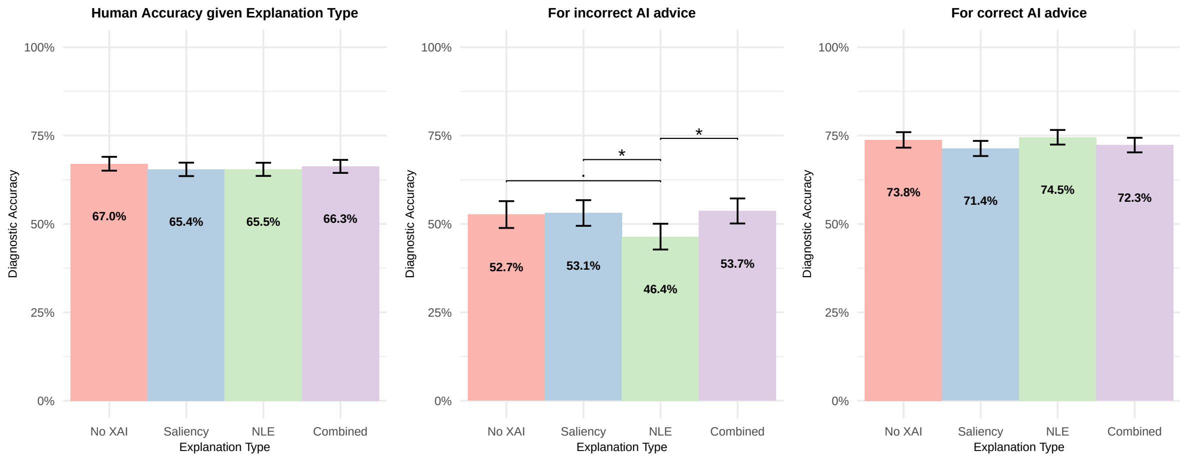


Figure 17: Human accuracy given explanation types (a) for both incorrect (b) and correct (c) advice.

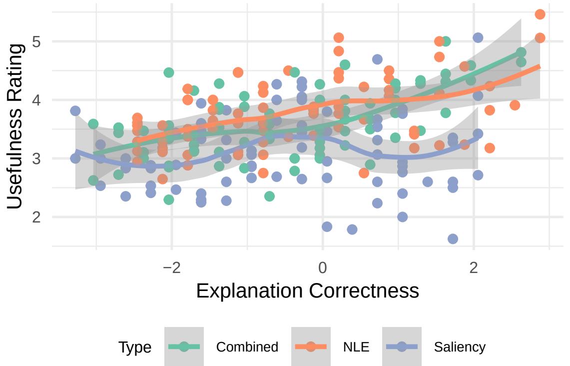


Figure 18: The perceived usefulness of NLEs and combined explanations increases with explanation quality (observed trends). Saliency maps do not follow this trend.

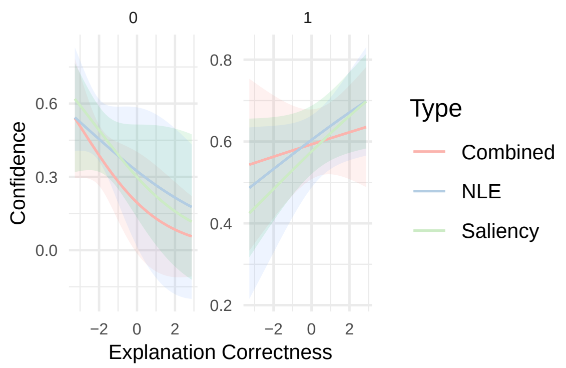


Figure 20: We find that insightful explanations increase reported confidence. Left panel is *incorrect advice*, right panel *correct advice*.

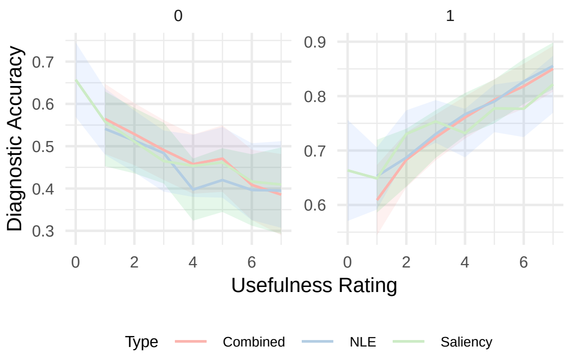


Figure 19: The diagnostic accuracy increases with the perceived usefulness of explanations when AI advice is correct (*right*).


Figure 21: We find no significant effect of explanation type on confidence.

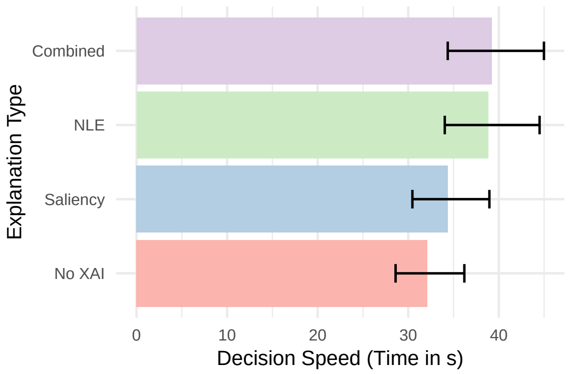


Figure 22: The adjusted mean decision speed (95% CI) is smallest the shortest for “No XAI“.

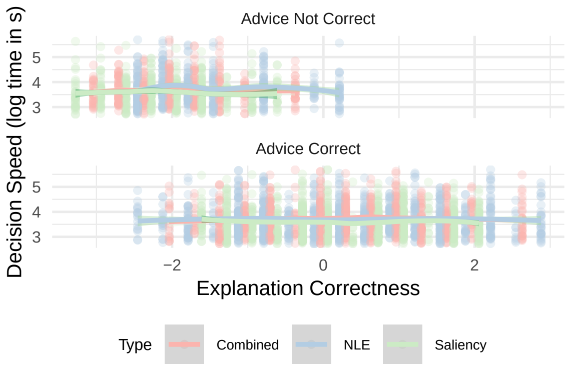


Figure 23: The explanation quality does not have an effect on the decision speed. Neither in the top panel (AI advice incorrect) nor the bottom panel (AI advice correct) a clear trend between explanation correctness and decision speed is visible.

**Explanation Correctness does not influence decision speed.** We find that the correctness of explanations does not significantly influence the decision time. In Figure 23, we show that the log decision time is almost constant across explanation correctness. We find this is true across variations of Advice Correctness and Explanation Type. Additionally, a GLMM including explanation correctness does not significantly improve the model likelihood.

## I Annotation process

When evaluating the AI advice, annotators are presented with a chest X-ray and a single class predicted by the AI (e.g. “pneumonia”). They are then asked whether they think the class is “Not present” (the finding can not be seen so is not worth mentioning or it can be mentioned negatively. For example: “No signs of pneumonia.”), “Maybe present” (while the evidence is inconclusive and/or there is some ambiguity, it’s worth mentioning in the radiology report that the finding may be present. For example: “Bibasilar opacities may represent atelectasis or pneumonia.”), or “Definitely present” (the finding is clearly present and will be noted in the radiology report. For example: “There are clear signs for pneumonia.”), following a common convention in evaluating the presence of chest X-ray findings (cite MIMIC-CXR, Chexpert). Both the annotators and study participants are instructed to interpret the labels as above.

The annotators also evaluate the textual explanation and heatmap for each prediction. Given that explanations can vary significantly in information richness Rivera-Garrido et al. (2022), we argue that a continuous scale is better suited than a binary correctness label, as has been done by Morrison et al. (2023). Suppose our annotators deem the AI advice (e.g. “pneumonia”) to be correct (“Definitely present” or “Maybe present”). In that case, we ask them “How correctly does the NLE (or heatmap) explain the AI advice pneumonia in this image?” and record their response on a 7-point Likert scale. We also ask them “If you consider the heatmap and the NLE as a joint explanation, how correctly do they explain the AI advice pneumonia in this image?” to obtain a correctness score for the combined explanation. In case they think the AI prediction is incorrect, we still want to get a measure of how much correct information an explanation contains and ask them the following: “How correctly does the heatmap (or NLE) highlight radiographic findings that would be relevant for the AI advice pneumonia in this image?”. An illustration of the annotator interface can be found in Figure 26.

We obtain our consensus by selecting the overall *advice correctness* as the majority vote of the three annotations, and the *explanation correctness* score of each explanation as the average of the three scores. We mean-center the *explanation correctness* scores for each type of explanation. Detailed outcomes of our annotation process can be found in the Appendix.

## J Study User Interface

Figure 25 shows an example test case from our screening survey and 26 shows a screenshot (bar the overlaying explanations) of our study user interface.

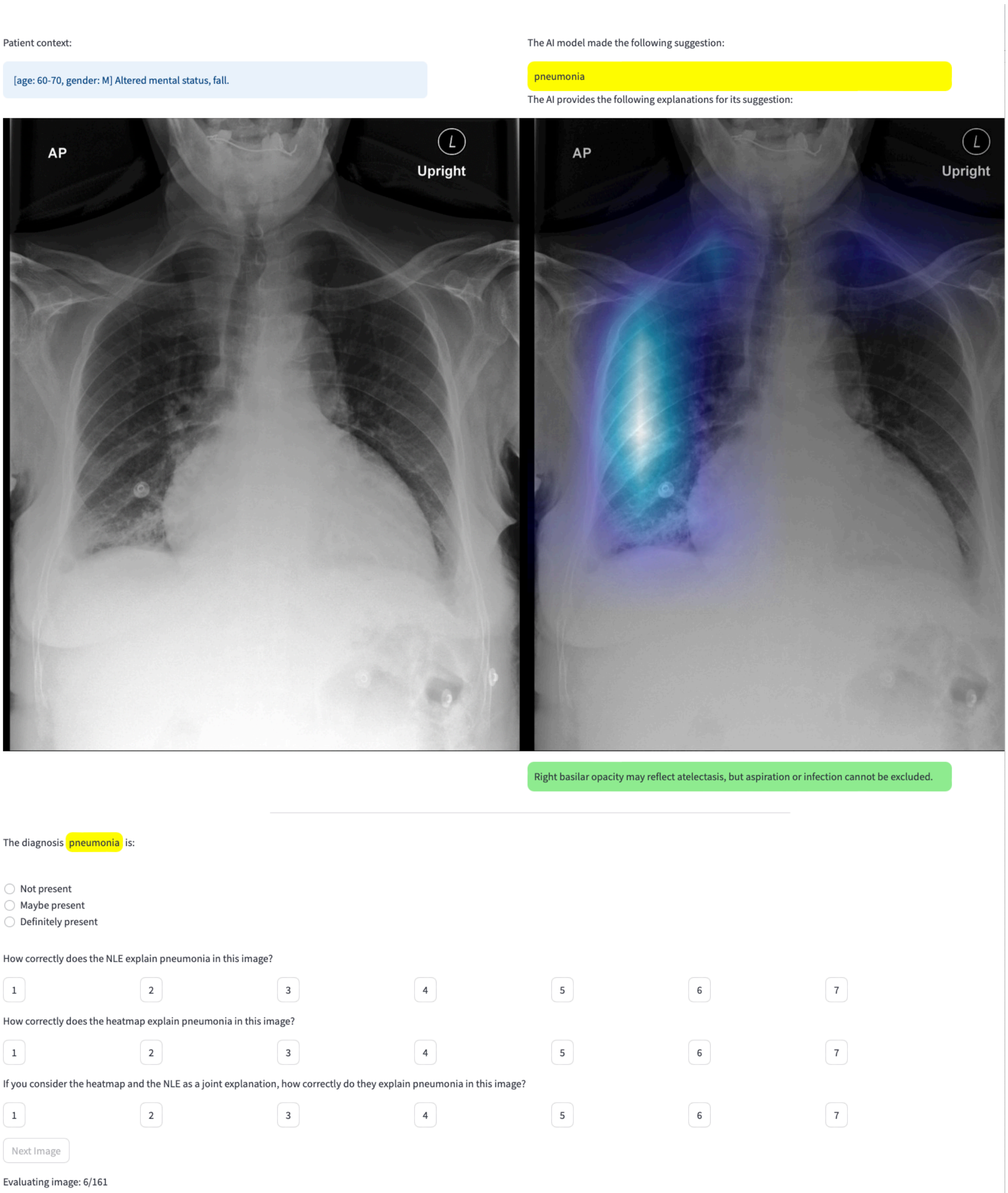


Figure 24: The platforms annotators used to annotate chest x-rays.


Figure 25: An example of one of the three test cases included in the screening survey.

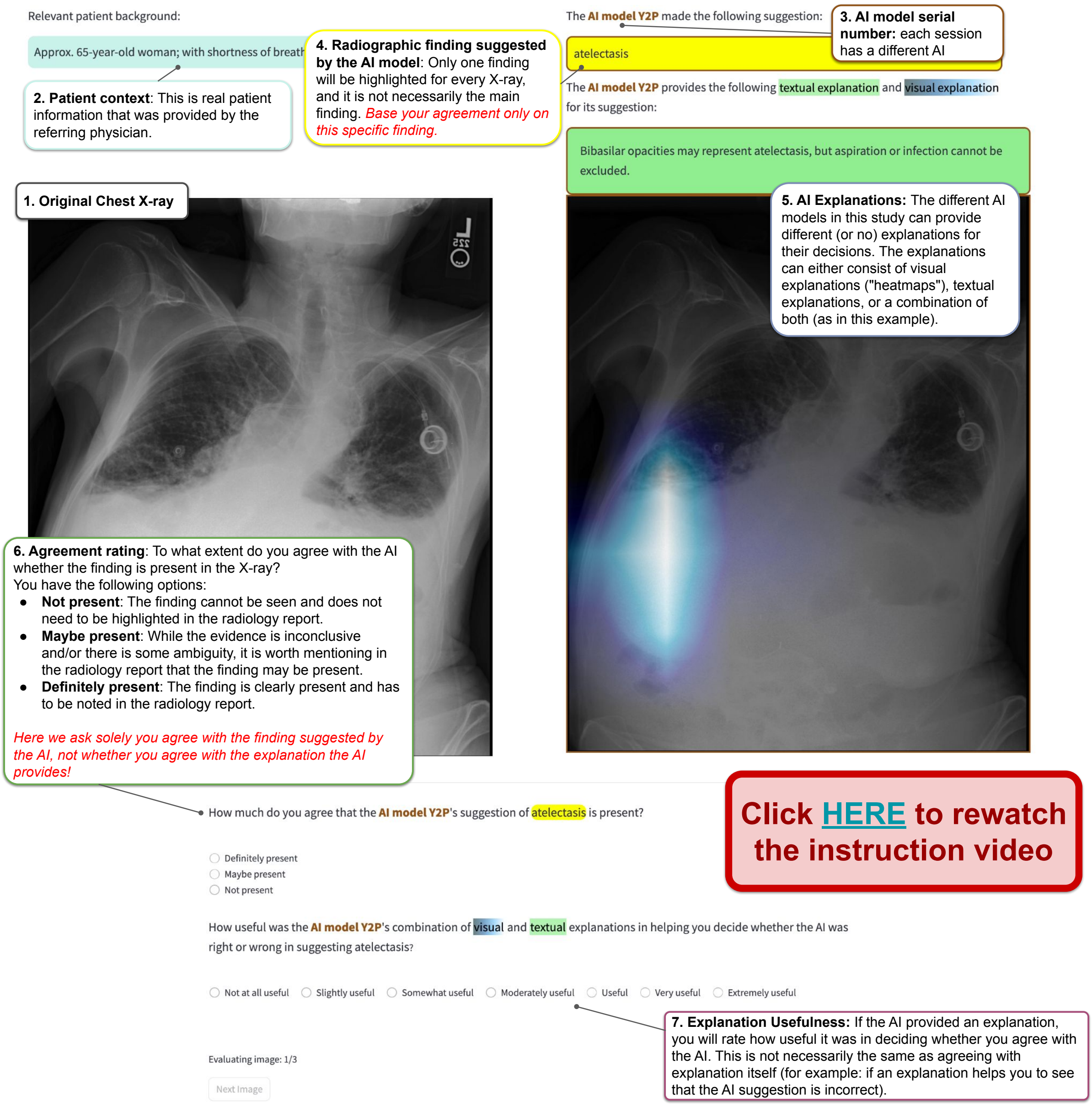


Figure 26: The instruction PDF that people have access to throughout the study. The 3-minute explanation video will be shared once the authors are no longer anonymized. This also shows cases of the UI that we used throughout the study (without the overlaying explanation boxes.