Imperfect Natural Language Explanations in Human-AI Decision-Making

KATELYN MORRISON[*], Carnegie Mellon University, USA
PHILIPP SPITZER[*], Karlsruhe Institute of Technology, Germany
VIOLET TURRI, Carnegie Mellon University, USA
MICHELLE FENG, Carnegie Mellon University, USA
NIKLAS KÜHL, University of Bayreuth, Germany
ADAM PERER, Carnegie Mellon University, USA

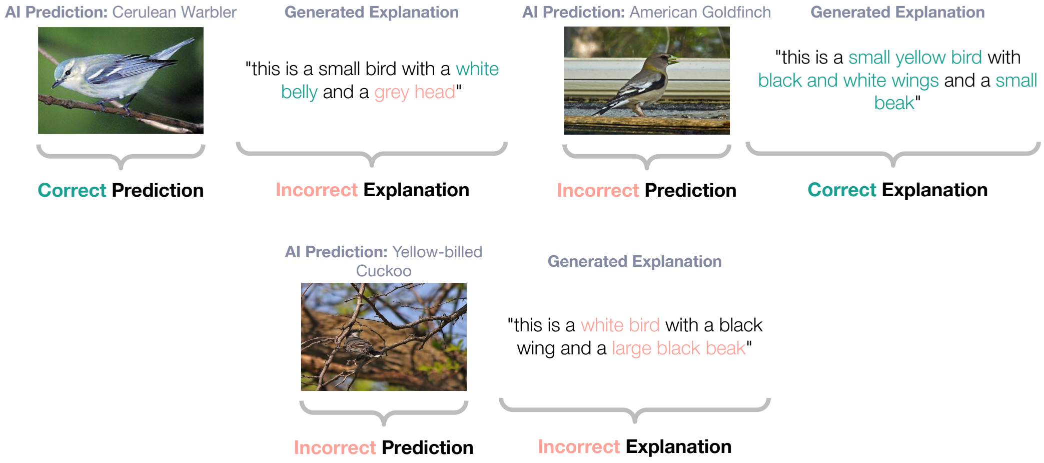

AI Prediction: Cerulean Warbler
Generated Explanation
"this is a small bird with a white belly and a grey head"
Correct Prediction
Incorrect Explanation
AI Prediction: American Goldfinch
Generated Explanation
"this is a small yellow bird with black and white wings and a small beak"
Incorrect Prediction
Correct Explanation
AI Prediction: Yellow-billed Cuckoo
Generated Explanation
"this is a white bird with a black wing and a large black beak"
Incorrect Prediction
Incorrect Explanation

Fig. 1. Different scenarios of imperfect natural language explanations based on model proposed by Hendricks et al. [19]: correct prediction, incorrect explanation; incorrect prediction, correct explanation; both incorrect prediction and explanation.

Explainability techniques are rapidly being developed to improve human-AI decision-making across various human-computer interaction settings. Consequently, previous research has evaluated how decision-makers collaborate with imperfect AI by investigating appropriate reliance and task performance to design more human-centered computer-supported collaborative tools. Several human-centered explainable AI (XAI) techniques have been proposed in hopes of improving decision-makers' collaboration with AI; however, these techniques are grounded in findings from previous studies that primarily focus on the impact of incorrect AI advice. Few studies acknowledge the possibility of the explanations being incorrect even if the AI advice is correct. With recent advancements in large language models (LLMs), post-hoc natural language explanations are becoming increasingly prevalent. However, LLMs are far from perfect, making it crucial to understand how

[*] Both authors contributed equally to this research.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

imperfect natural language explanations affect human-AI decision-making. In this encore paper, we summarize findings from a robust, mixed-methods user study to evaluate how incorrect natural language explanations influence humans' decision-making behavior in a bird species identification task, taking into account their level of expertise and an explanation's level of assertiveness[1]. Our findings reveal the influence of imperfect natural language explanations and humans' level of expertise on their reliance on AI and human-AI team performance. We also discuss how these explanations can be deceptive during human-AI collaboration. Hence, we shed light on the impacts of language models as post-hoc explainers in human-computer interaction and provide guidelines for designers of human-AI collaboration systems.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Artificial intelligence**; **Computer vision tasks**.

Additional Key Words and Phrases: Human-AI Collaboration, Explainable AI, Natural Language Explanation

ACM Reference Format:
Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2018. Imperfect Natural Language Explanations in Human-AI Decision-Making. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With the deployment of imperfect artificial intelligence (AI) in high-stakes decision-making scenarios, decision-makers struggle with knowing when they should and should not rely on AI advice, causing frustration and potentially harmful decisions. As a result, designing and developing human-centered explanations has become a core theme in human-computer interaction (HCI) research [14, 30, 54]. Tangentially, recent work has proposed new post-hoc explanation techniques that leverage machine learning models to explain the prediction of another machine learning model [7, 9, 19, 23, 29, 40, 44, 47, 50, 55, 60]. A subset of these studies propose to exploit language models to generate natural language explanations for image classifications with the rationalization that natural language is more "human-friendly" [19, 23, 29, 47]. These approaches to explainability introduce another level of uncertainty in the collaboration between the decision-maker and the AI, as language models can hallucinate [44].

HCI, and more specifically human-AI collaboration, is prevalent across high-stakes scenarios [11, 31, 48, 53]. For instance, radiologists collaborate with LLMs to identify abnormalities in medical imagery and generate impressions from image embeddings [20, 48]; conservationists use AI to help monitor biodiversity [8], and humanitarian aids use AI to help identify damaged buildings after natural disasters or armed conflicts from satellite imagery [34, 61]. While some of these human-AI collaboration scenarios require the human decision-maker to have several years of experience in the given domain, such as radiology, monitoring biodiversity with the help of AI doesn't necessarily require domain expertise [8, 39]. Platforms, such as iNaturalist [1], Merlin Bird App [2], and Wildbooks from WildMe.org [3], have allowed non-experts (*i.e.*, citizen scientists, hobbyists, or students) to partake in monitoring biodiversity alongside domain experts (*i.e.*, ornithologists and conservationists). While these platforms are valuable for non-experts to use, the AI models backing these platforms are imperfect: They do not always provide correct predictions [28].

Experts and non-experts interacting with the same imperfect AI and the same type of explanations in human-AI collaboration scenarios, such as decision-making [42] or learning systems [46], could result in some users misunderstanding or inappropriately relying on/overriding the AI advice. Experts may have more context outside of the AI's classification and confidence that a non-expert may not have. This might result in experts using their contextual information to appropriately rely

[1] The encore submission intentionally only focuses on the natural language component of the original study to align more closely with discussions during the workshop.

on the AI when advice is provided, such as correctly overriding when wrong AI advice is presented and correctly using AI advice when it is correct. Non-experts, on the other hand, might not be able to judge the correctness of the AI advice appropriately as they are missing this contextual information. For example, for bird species identification, ornithologists tend to be more aware of information related to the visual differences between the male and female birds for a given species, the bird’s habitat, and migration patterns, whereas non-experts may not know some or all of that information. This same situation can arise in radiology where residents (“non-experts”) may initially be less familiar with certain diseases than an attending radiologist (“experts”). However, both experts and non-experts can struggle to identify certain instances. In this case, collaborating with AI can result in complementary team performance (CTP), leveraging the unique knowledge of both humans and AI, resulting in the human-AI team’s task performance being better than the human or AI alone [18].

Inappropriate reliance can also occur in the presence of *imperfect XAI*, a term that we introduce to represent the phenomenon where an explanation reveals evidence that does not necessarily comply with the prediction. Papenmeier et al. [38] use the term ‘explanation fidelity’ while Kroeger et al. [29] use the term ‘faithfulness’ to measure how “truthful” an explanation is. We use imperfect XAI to align with existing terms, such as imperfect AI, in the HCI community. Specifically, we define imperfect XAI as explanation techniques that can potentially provide explanations that do not fit with the AI’s predictions. We view explanation fidelity or faithfulness as a term that can be used under the umbrella term of imperfect XAI; we view explanation fidelity as referring to the continuum of explanation correctness, such as when explanations are partially correct. Imperfect XAI can exist regardless of whether the AI’s advice is correct or not. AI explanations may oversimplify or improperly estimate complex models in order to make them more interpretable, deceiving and misleading the human decision-maker. As a result, non-experts may be more prone to under- or over-relying on AI advice in the presence of incorrect explanations. Within knowledge transfer scenarios, this could cause the non-expert to learn incorrect information about a given class. Furthermore, previous research shows that the language tone within natural language explanations can impact decision-makers [12]. However, the effect of language tone on humans’ appropriate reliance when interacting with imperfect explanations is underexplored. Therefore, it is necessary to investigate how the communication style of explanations (*e.g.*, the language tone and the content) impacts human-AI collaborations across levels of expertise [12, 27].

Collaborating with imperfect AI is not a new concept to the HCI community [18, 28, 31]. Despite numerous user studies over the years investigating human-AI collaborations and XAI, few have formally acknowledged the existence of imperfect XAI in their studies [18, 38]. By formally acknowledging the existence of imperfect XAI in human-AI collaboration, research has several new interesting dimensions to explore. Although numerous user studies seek to understand how humans align, perceive, and interact with different types of explanations in various human-AI collaboration scenarios (*e.g.*, [13, 25]), few studies explore the impact that incorrect or “noisy” explanations have on human-AI collaboration [18, 21, 38, 51]. Recent work has used technical approaches to mitigate “noisy” or incorrect natural language explanations [21] and Kroeger et al. [29] proposed metrics to algorithmically evaluate the effectiveness of the generated post-hoc natural language explanations. However, to our knowledge, no studies investigate how the interaction between the correctness of explanations and the decision-maker’s level of expertise impact appropriate reliance on AI, human-AI team performance, and the extent to which AI explanations deceive decision-makers.

With the rapid advancements of LLMs and their growing use in high-stakes decision-making scenarios, our encore submission echoes the argument from our full paper ([35]) that it is necessary to understand how humans interact with natural language explanations that are incorrect, even when the AI’s advice is correct. We also argue that it is important to understand the relationship that the level of expertise and the tone of explanations (*i.e.*, assertive, non-assertive, or neutral) have on a decision-maker's reliance on AI. Understanding these dimensions of human-AI collaboration will provide insight to XAI and HCI researchers. With these topics under-explored in current literature, we present the following research questions for this encore submission:

RQ1: How does the correctness of explanations affect appropriate reliance on AI, and to what extent do the decision-maker's level of expertise and the explanation's assertiveness moderate this effect?

RQ2: How is complementary team performance impacted by the correctness of explanations and the decision-maker's level of expertise?

RQ3: To what extent do incorrect and correct explanations deceive decision-makers with different levels of expertise?

2 METHODOLOGY

Human-computer interaction is becoming core to wildlife conservation efforts [10, 16, 49]. While identifying bird species from images may not be posed as a high-stakes task in our study, this task is imperative to conserving and managing species and biodiversity [4]. Furthermore, the task of fine-grained image classification, such as bird species identification, is comparable to higher-stakes tasks, such as identifying diseases from medical imagery [48]. For example, radiologists collaborating with an imperfect AI and imperfect XAI to diagnose diseases present in chest X-rays would go through a similar visual decision-making process as if they were trying to classify an image of a Bewick Wren in our study interface. Hou et al. [20], Kayser et al. [23] propose natural language explanations for chest x-rays, similar to the explanation we implement in our study, which helps bridge our findings between bird species identification and higher-stakes tasks.

2.1 Study Design

To answer our research questions, we design a mixed between- and within-subjects study to examine various effects of explanations on appropriate reliance. Our study was carefully reviewed by two experienced birders: one experienced birder is a migration counter for a bird sanctuary, and the other experienced birder holds a graduate degree in environmental science, conducted research at a nature center, and worked at a nature conservancy. The procedure of the study follows the design outlined in Figure 2 and is divided into four different parts ($A - D$), which we explain in further detail below.

The study begins with part A in Figure 2: In a bird identification test, we assess participants' expertise in classifying six different species of birds. We distinguish the six bird images based on their level of difficulty. This level of difficulty is derived from discussions with an experienced migration counter from a bird sanctuary. While previous work has collected participants' self-perception of expertise [25], this method is subjective, and participants may self-perceive their skills differently. Kazemitabaar et al. [24] measure participants "experience-level" through log data instead of subjective measures. Similarly, we try to avoid defining expertise subjectively. For the purpose of our analyses, we identify two different levels of expertise: *non-experts* and *experts*.

In the next section of the study (part B in Figure 2), participants are randomly assigned to one of two explanation types: *Natural language* explanations or *visual, example-based* explanations. We use natural language explanations because the AI model that we used for the study was specifically designed to generate natural language explanations based on fine-grained image classifications [19]. We will only focus on the findings for the natural language explanations for this encore submission.

The human-AI bird identification task (part C of Figure 2) consists of two phases. For each treatment condition, the participants are asked to identify the bird species from an image (phase 1). After submitting an initial identification, they are shown the AI's prediction along with the explanation, and again, they have to submit an identification for the bird species in the image (phase 2). The structure of phases one and two are corroborated with previous work [15]. Initially, participants must click on a button that shows “Show AI Explanation”. Without revealing the explanation, the participant cannot proceed to the next question. This is one way for us to ensure that the participant acknowledges the presence of an explanation. Overall, participants do this process for twelve different random bird images.

As the AI we are utilizing for identifying the bird species is not perfect [19], the predictions and explanations provided can be incorrect. To understand how this affects participants’ appropriate reliance, we ensure that each participant is shown three samples of the following four categories in random order: correct prediction and correct explanation (**CC**); correct prediction and incorrect explanation (**CI**); incorrect prediction and correct explanation (**IC**); incorrect prediction and incorrect explanation (**II**).

Overall, participants are shown twelve different bird species. For each of the four categories we identified, we show three samples where each explanation is framed with a different level of assertiveness: *assertive*, *non-assertive*, or *neutral*[2]. As a result, each participant is shown one *assertive*, one *non-assertive*, and one *neutral* explanation for each category. We ensure that the order is randomized for each participant. Moreover, we also vary the samples shown, meaning that not every participant sees the same bird images. This is to ensure that our results are not dependent on the difficulty of the bird species.

2.2 Data Selection

We create a dataset of bird images and explanations to show participants by manually curating bird images from the well-established CUB-200-2011 [52] dataset and explanations from the Generating

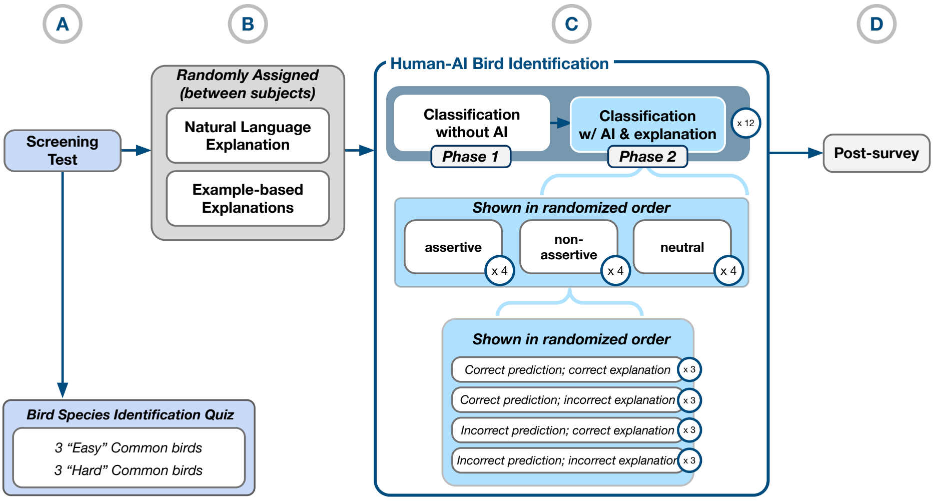
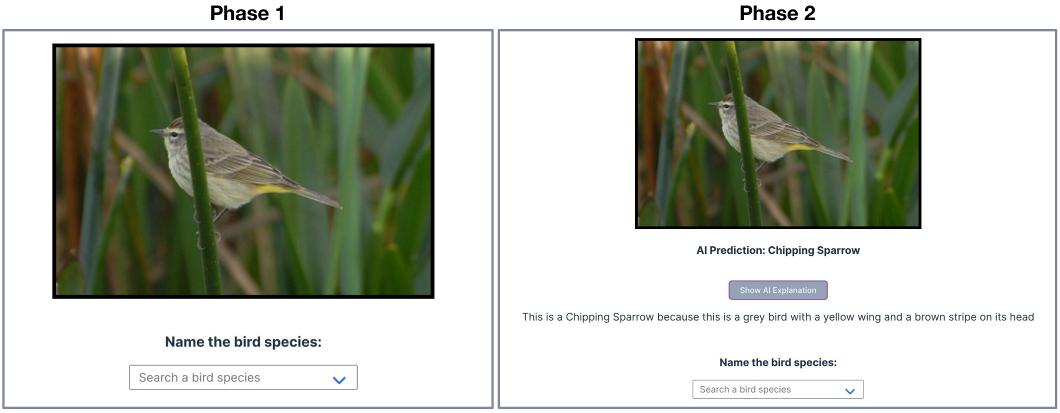

A
B
C
D
Screening Test
Randomly Assigned (between subjects)
Natural Language Explanation
Example-based Explanations
Human-AI Bird Identification
Classification without AI
Phase 1
Classification w/ AI & explanation
Phase 2
x 12
Shown in randomized order
assertive
x 4
non-assertive
x 4
neutral
x 4
Shown in randomized order
Correct prediction; correct explanation
x 3
Correct prediction; incorrect explanation
x 3
Incorrect prediction; correct explanation
x 3
Incorrect prediction; incorrect explanation
x 3
Post-survey
Bird Species Identification Quiz
3 “Easy” Common birds
3 “Hard” Common birds

Fig. 2. We conduct a rigorous mixed-methods study leveraging a mixed design. Before participants start the task, they are shown a screening test (A). For the human-AI bird identification task, participants are assigned an explanation modality (B). During the task, participants are shown explanations with different levels of assertiveness and different scenarios of correctness (C). Lastly, participants complete a post-survey (D). This encore submission only focuses on the findings for the natural language explanations.

[2] Figure 4 provides examples of *assertive*, *non-assertive*, and *neutral* explanations.



This is a Chipping Sparrow because this is a grey bird with a yellow wing and a brown stripe on its head

Fig. 3. Example of the two phases for a single bird image that a participant is shown in the study for the **II** case.

Visual Explanations model [19]. The original dataset consists of 11,788 images of 200 bird species and is split into 5,994 training and 5,794 test images. Each bird species is represented with around 60 images of the respective bird class. When curating birds, we first filtered for bird families with several species in the CUB dataset. We specifically filtered out every bird class that is not a part of the Warblers, Wrens, Swallows, Sparrows, or Finches/Grosbeaks families. After applying this filter, we had 1, 864 out of 5, 794 images from the test set of CUB-200-2011. Of those 1, 864 images, 1, 609 were predicted correctly by the AI, and 255 images were predicted incorrectly by the model.

After filtering the bird species, multiple researchers on our team separately classified the natural language explanations for a subset of the 1, 864 birds as incorrect or correct. Cases of doubt were discussed by a subset of the research team and excluded from consideration if an agreement was not met. In total, we identified ten examples for each category[3] and explanation type. As a result, the dataset represents 66 different images and 43 different bird species from the CUB-200-2011 dataset.

We define a correct natural language explanation to be when the explanation aligns with the description of the predicted bird class. We define an incorrect natural language explanation to misalign with all or part of the description of the predicted bird class. Thus, an incorrect natural language explanation contains a factual error. This type of incorrectness is present in different natural language techniques and is a focus of current research [32, 59]. We use descriptions from the Cornell Lab of Ornithology All About Birds Guide [**?**] to corroborate our classification for each explanation. Examples of incorrect and correct natural language explanations are provided in Figure 1.

2.3 Explanation Modalities

The natural language explanations were generated by the model proposed by Hendricks et al. [19]. We followed the PyTorch implementation [5] of Hendricks et al.'s model to obtain the natural language explanations since the original model from Hendricks et al. was unavailable. After running the test images through the model, a natural language explanation is generated for each classification. For example, the natural language explanation for the Magnolia Warbler in Figure 3 is: "`this is a grey bird with a yellow wing and a brown stripe on its head`".

[3] The four categories are identified in Section 2.1.

Following previous studies [12, 37, 58], we define assertive explanations to include words and adjectives such as “definitely” and “clearly”. We define the non-assertive explanations to include words and adjectives such as “might be” and “appears to be”. For the neutral condition, we omit the adjectives to maintain the same structure of the information being presented. For the natural language explanations, we append the assertiveness to the beginning of the explanation generated by the model to read like a sentence. For the non-assertive and assertive conditions, we removed the text “this is a” from the generated explanation in order to incorporate it into the sentence structure we designed. The three versions of assertiveness for both explanation modalities are shown in Figure 4.

This might be **a** {predicted class}**,**
since it appears to be **a** {explanation}
non-assertive

This is definitely **a** {predicted class}
because it clearly **is a** {explanation}
assertive

This is a {predicted class}
because {explanation}
neutral

Fig. 4. The three different language tones that an explanation could have regarding assertiveness for the natural language explanations. Non-assertive explanations included the words “might be” and “appears to be”. Assertive explanations included the words “definitely” and “clearly”. Neutral explanations did not include any additional adjectives.

2.4 Recruitment

We recruit the participants through several communication channels that are related to the environment and conservation, such as the AI for Conservation Slack, Birding International Discord, Climate Change AI community forum, WildLabs.net community forum, and Audubon Society mailing lists. Additionally, we use Prolific with a custom filter to target individuals who currently work in environment-related fields. Participants receive compensation that is above minimum wage. After excluding participants who provide incomplete and fake responses (i.e., lorem ipsum response to our survey question), we have 69 people complete the study for natural language explanations.

2.5 Quantitative Metrics

We quantitatively calculate appropriate reliance across the four dimensions defined by Schemmer et al. [43]: correct AI reliance, correct self-reliance, under-reliance, and over-reliance. Following these metrics, we calculate relative self-reliance (RSR) and relative AI reliance (RAIR) to account for the appropriateness of reliance. We separately measure RAIR and RSR for correct and incorrect explanations and derive their impact on appropriate reliance. In order to measure this impact, we propose the Deception of Reliance (DoR) caused by imperfect XAI. In order to measure the overall deception impact of explanations on humans' decision-making behavior, we compute the deception on appropriate reliance by calculating the difference between the Gaussian distance in the RAIR-RSR space for incorrect and correct explanations. This difference represents the deception between the correct and incorrect explanations. If the deception is a positive value, then incorrect explanations are more deceptive; if the difference is a negative value, then correct explanations are more deceptive. Lastly, as defined by previous work (*e.g.*, [6, 15, 18]), we can calculate the human-AI team performance to determine if CTP exists. We utilize accuracy as a performance metric. Since every participant is shown six birds that the AI correctly classifies and six that the AI incorrectly classifies, the model performance is 50%.

3 RESULTS

3.1 Participant Statistics

On average, the study takes 24 minutes to complete. In order to distinguish experts from non-experts, we perform K-means clustering ($k = 2$) based on a principal component analysis with two components for four features from the bird species identification test (part A of Figure 2). These four features represent participants' scores in correctly identifying the family and species of the easy and the difficult bird images. By clustering all of the participants into the expert and non-expert groups, we end up with 41 experts and 28 non-experts for the natural language explanations.

3.2 Moderation Analyses

In order to test whether humans' level of expertise and the explanations' assertiveness moderate the relation of the correctness of explanations on humans' appropriate reliance, we conduct several moderation analyses.

3.2.1 Participants' level of expertise moderates the effect of the correctness of explanations on RAIR for natural-language explanations. We model the correctness of explanations as an independent variable. Accordingly, we model RAIR as the dependent variable. To account for the moderation effect of the level of expertise and assertiveness, we examine each variable as a moderator and report the interaction effects with the correctness of explanations. The moderation analysis shows that the interaction of the level of expertise with the correctness of explanations is significant (coeff = −1.00, p-value = .05). We observe a negative coefficient. Accordingly, the moderation effect on the relation of correctness on RAIR is higher for non-experts than for experts. In other words, non-experts change their initially incorrect decision to align with the correct AI advice more often than experts do when the natural language explanation is correct. However, there is no significant effect in the interaction of assertiveness and the correctness of explanations. Thus, we conduct a regression analysis with the moderators as independent variables to evaluate for a direct effect of assertiveness as recommended by Hayes [17] and Warner [56]. The results of the regression analysis show that there is no direct effect between assertiveness and RAIR (coeff = .04, p-value = .77).

3.2.2 Participant's level of expertise has a direct effect on RSR for natural language explanations. In addition to analyzing whether the level of expertise and assertiveness moderate the effect of explanations' correctness on RAIR, we conduct the same analyses for the effect of explanations' correctness on RSR. For RSR, we look at all cases in which the AI prediction is giving incorrect advice (i.e., the prediction is wrong) and the initial human decision is correct [43]. The moderation analysis for the natural language explanation shows that there is no significant effect of correctness on RSR moderated by level of expertise (coeff = −13.33, p-value = .98) and assertiveness (Z1 x corr.: coeff = −.28, p-value = .71; Z2 x corr.: coeff = .90, p-value = .23). Thus, we perform a regression analysis with the level of expertise and assertiveness as independent variables and drop the interaction terms. We observe that there is no significant effect of assertiveness on RSR (coeff = .10, p-value = .58). However, the level of expertise (coeff = 3.18, p-value = .00) has a significant effect on RSR. With a positive coefficient, this means that experts dismiss incorrect AI advice more than non-experts when shown natural language explanations. This can also be seen in ??, which tells us that experts have a higher RSR than non-experts.

3.3 Human-AI Team Performance

The AI's performance is always 50% because the study was designed to show participants six birds that the model correctly classified and six that the model incorrectly classified. We see that when experts are paired with the AI, their performance improves by 8.74%, performing 6.91% better than the AI alone. While experts reach CTP, we do not see this for non-experts. However, we see that the non-experts greatly improve their performance and nearly match the AI's performance when paired with the AI. When we only consider cases with correct explanations, the non-experts' task accuracy is approximately the same as the AI alone, 48.81%, while the experts maintain complementary team performance. When only considering incorrect explanations, we still see complementary team performance for the experts. However, the non-experts' task accuracy suffers more when shown incorrect explanations. Non-experts' task accuracy for natural language explanations is 42.86%.

3.4 Deception caused by Imperfect XAI

Lastly, we compare RAIR to RSR for both levels of expertise and the correctness of explanations. By measuring RAIR and RSR for incorrect and correct explanations separately, we can calculate the deception caused by imperfect XAI. We do not explore assertiveness since we do not see any significant direct or moderation effects. We calculate the deception of reliance separately for RSR and RAIR: DoR_{RSR} and DoR_{RAIR}.

We observed that experts have a higher RSR than non-experts for both incorrect and correct explanations for natural language explanations, validating that experts rely more on their own initial decisions when AI advice is given. However, we do not see this trend for natural language explanations. For the natural language explanations, the DoR_{RAIR} is positive, meaning that experts rightly follow correct AI advice more often when provided with correct explanations than with incorrect explanations. Non-experts have a similar DoR_{RSR}, indicating no significant difference in their RSR between correct and incorrect explanations. Interestingly, for natural language explanations, the incorrect explanations are not as misleading ($DoR_{RAIR} = 0.03$, not significant, with a p-value = 0.68). In general, non-experts have a higher RAIR than experts.

4 DISCUSSION AND CONCLUSION

We investigate how imperfect XAI impacts humans' decision-making when collaborating with AI. More precisely, we assess how imperfect explanations affect humans' reliance behavior and investigate the effects on human-AI team performance. To answer RQ 1 and RQ 2, we assess the validity of our research model for natural language explanations. Previous research emphasizes the need to consider imperfect AI when designing for human-AI collaboration [28]. With recent research looking into how humans and AI can achieve complementary team performance [6], Schemmer et al. [43] conceptualize the role of appropriate reliance in human-AI collaboration. We extend Schemmer et al. [43]'s framework by adding another dimension: XAI advice. Given that an explanation can be incorrect even if the AI advice is correct, it is crucial to understand the impact of incorrect XAI advice on decision-making. Furthermore, it is necessary to understand the impact of imperfect XAI for different types of explanations. Below, we discuss how our contributions are situated in current literature and the implications for HCI.

In our study, we observe a significant moderation of humans' level of expertise on the effect of explanations' correctness on RAIR for natural language explanations. However, we do not see this moderation for RSR. When humans are being provided wrong AI advice, their level of expertise does not moderate the impact of imperfect explanations on humans' RSR. We identify a direct effect of the level of expertise on RSR. Overall, our work synthesizes how humans' level of expertise impacts their reliance on AI when provided with imperfect explanations. Non-experts rely more on AI than experts, whereas experts rely more on their initial decisions. Thus, this study sets a starting point to investigate the effect of imperfect XAI for natural language explanations.

Our findings show that imperfect explanations impact human-AI decision-making. We observe a difference between reaching complementary team performance when the correctness, or fidelity, of the explanation changes. Previous research discusses the impact of explanations’ fidelity on humans’ reliance on AI and hypothesizes that fidelity has a positive impact on humans’ reliance behavior on AI [18]. With our results, we confirm this hypothesis. Furthermore, Papenmeier et al. [38] observe that low-fidelity explanations (or incorrect explanations) impact user trust in AI when the global model performance is around 75% accurate, which helps validate our findings. We also observe that the lack of expertise among non-experts impacts their task performance when shown incorrect explanations regardless of the AI advice being correct. Similar to our findings, Nourani et al. [36] observe that non-experts tend to over-rely on AI advice, attributing this to their inability to identify when the AI is incorrect because of their lack of expertise. These findings contribute to a more integrated understanding of the impact of human-AI decision-making on different user groups in the presence of imperfect XAI.

The language tone of explanations does not impact humans’ decision-making behavior. Calisto et al. conclude that the level of expertise influences whether the framing of the explanation should be assertive or non-assertive [12]. Based on their observations, they specifically suggest that natural language explanations should be designed such that the tone of the explanation is appropriate for the end user’s level of expertise. Despite the numerous previous studies finding that the framing of the explanations has a significant impact on human-AI collaboration [12, 26, 27], our findings do not show an impact on appropriate reliance. However, given the potential for natural language explanations to present irrelevant or incorrect information [44], we encourage future work to explore various ways to alter the presentation of an explanation based on its likelihood to contain hallucinations.

Our findings can guide researchers and practitioners on assessing and designing for imperfect XAI in human-AI collaborations. It is important to understand how humans interact with imperfect XAI. Visual explanations, such as example-based explanations and saliency maps, have been shown in the past to be of high educational value to the end-user (*e.g.*, [25, 33]), making it even more important to understand how to design for and mitigate imperfect XAI. This need is intensified with the role AI takes in organizational learning [46]. Especially in the workplace, AI can facilitate knowledge transfer and support organizations in retaining and distributing expert knowledge [22, 45, 57]. Similarly, it is also crucial to understand how imperfect XAI affects the learning of novices through AI-based learning systems [46] or through collaboration with AI [41]. Our findings can guide knowledge managers within organizations on how to make use of explanations for employees with different levels of domain knowledge. More precisely, knowledge managers should be aware of the impact of exposing humans with different levels of expertise to imperfect XAI. In addition, our findings contribute to a more integrated understanding of the impact that incorrect explanations can have on human-AI decision-making and inform different stakeholders in organizations. As non-experts are more affected by imperfect XAI, their performance drops more than experts’ performance when incorrect explanations are provided in comparison to correct explanations. With this finding, knowledge managers can adjust their knowledge retention activities when training new employees; designers can adjust the development of human-AI collaboration systems to successfully facilitate explanations in decision-making and support humans in their work setting. Thus, we encourage practitioners designing human-AI collaboration systems to apply our findings to structure their design approach. This can aid organizations in laying out the strategic direction of human resource development by matching the use of explanations to humans’ prior knowledge. Overall, these findings shed light on the ongoing discussion in CSCW on how to make use of explanations within work settings.

5 CONCLUSION

This encore submission echoes the argument for understanding the impact of imperfect natural language explanations in human-AI collaboration. Through a mixed-methods study, we empirically analyze humans' decision-making and specifically assess whether their level of expertise and explanations' assertiveness moderate the effect of imperfect XAI on appropriate reliance. With the rapid development of large language models and HCI's increasing adoption of them in human-AI collaboration settings, we encourage future work and discussions to uncover insights beyond our findings.

REFERENCES

[1] 2023. https://www.inaturalist.org/. Accessed: July 2, 2023.

[2] 2023. https://merlin.allaboutbirds.org/. Accessed: July 2, 2023.

[3] 2023. https://www.wildme.org/. Accessed: July 2, 2023.

[4] Hüseyin Gökhan Akçay, Bekir Kabasakal, Duygugül Aksu, Nusret Demir, Melih Öz, and Ali Erdoğan. 2020. Automated bird counting with deep learning for regional bird distribution mapping. *Animals* 10, 7 (2020), 1207. https://doi.org/10.3390/ani10071207

[5] Stephan Alaniz. 2018. pytorch-gve-lrcn: PyTorch implementation of Visual Generation and Execution for Long-term Predictions. https://github.com/salaniz/pytorch-gve-lrcn.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16. https://doi.org/10.1145/3411764.3445717

[7] Catarina Barata and Carlos Santiago. 2021. Improving the explainability of skin cancer diagnosis using CBIR. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III*. Springer, 550–559. https://doi.org/10.1007/978-3-030-87199-4_52

[8] Tanya Y Berger-Wolf, Daniel I Rubenstein, Charles V Stewart, Jason A Holmberg, Jason Parham, Sreejith Menon, Jonathan Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. 2017. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arXiv preprint arXiv:1710.08880* (2017). https://arxiv.org/abs/1710.08880

[9] Thales Bertaglia, Stefan Huber, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2023. Closing the Loop: Testing ChatGPT to Generate Model Explanations to Improve Human Labelling of Sponsored Content on Social Media. *arXiv preprint arXiv:2306.05115* (2023). https://doi.org/10.1007/978-3-031-44067-0_11

[10] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. 2022. Role of human-AI interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5286–5294. https://doi.org/10.1609/aaai.v36i5.20465

[11] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24. https://doi.org/10.1145/3359206

[12] 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*. 1–20. https://doi.org/10.1145/3544548.3580682

[13] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *arXiv preprint arXiv:2301.07255* (2023). https://doi.org/10.1145/3610219

[14] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): beyond opening the black-box of AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7. https://doi.org/10.1145/3491101.3503727

[15] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24. https://doi.org/10.1145/3359152

[16] Siân E Green, Jonathan P Rees, Philip A Stephens, Russell A Hill, and Anthony J Giordano. 2020. Innovations in camera trapping technology and approaches: The integration of citizen science and artificial intelligence. *Animals* 10, 1 (2020), 132. https://doi.org/10.3390/ani10010132

[17] Andrew F Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications. https://doi.org/10.1111/jedm.12050

[18] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS* (2021), 78. https://www.researchgate.net/profile/Patrick-Hemmer-3/publication/352882174_Human-AI_Complementarity_in_Hybrid_Intelligence_Systems_A_Structured_Literature_Review/links/60dddc9d299bf1ea9ed5c5a8/Human-AI-Complementarity-in-Hybrid-Intelligence-Systems-A-Structured-Literature-Review.pdf

[19] Lisa Anne Hendricks, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Zeynep Akata. 2021. Generating visual explanations with natural language. *Applied AI Letters* 2, 4 (2021), e55. https://doi.org/10.1002/ail2.55

[20] Benjamin Hou, Georgios Kaissis, Ronald M Summers, and Bernhard Kainz. 2021. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*. Springer, 293–303. https://doi.org/10.1007/978-3-030-87234-2_28

[21] Myeongjun Jang, Bodhisattwa Prasad Majumder, Julian McAuley, Thomas Lukasiewicz, and Oana-Maria Camburu. 2023. KNOW How to Make Up Your Mind! Adversarially Detecting and Alleviating Inconsistencies in Natural Language Explanations. *arXiv preprint arXiv:2306.02980* (2023). https://doi.org/10.18653/v1/2023.acl-short.47

[22] Mohammad Hossein Jarrahi, Sarah Kenyon, Ashley Brown, Chelsea Donahue, and Chris Wicher. 2023. Artificial intelligence: A strategy to harness its power through organizational learning. *Journal of Business Strategy* 44, 3 (2023), 126–135. https://doi.org/10.1108/jbs-11-2021-0182

[23] Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. 2022. Explaining Chest X-Ray Pathologies in Natural Language. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*. Springer, 701–713. https://doi.org/10.1007/978-3-031-16443-9_67

[24] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara J Ericson, David Weintrop, and Tovi Grossman. 2023. How Novices Use LLM-Based Code Generators to Solve CS1 Coding Tasks in a Self-Paced Learning Environment. *arXiv preprint arXiv:2309.14049* (2023). https://arxiv.org/abs/2309.14049

[25] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3544548.3581001

[26] Taenyun Kim and Hayeon Song. 2020. The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence's Suggestion. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8. https://doi.org/10.1145/3334480.3383038

[27] Taenyun Kim and Hayeon Song. 2023. Communicating the limitations of AI: the effect of message framing and ownership on trust in artificial intelligence. *International Journal of Human–Computer Interaction* 39, 4 (2023), 790–800. https://doi.org/10.1080/10447318.2022.2049134

[28] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3290605.3300641

[29] Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Are Large Language Models Post Hoc Explainers? *arXiv preprint arXiv:2310.05797* (2023).

[30] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021). https://arxiv.org/abs/2110.10790

[31] Martin Lindvall, Claes Lundström, and Jonas Löwgren. 2021. Rapid assisted visual search: Supporting digital pathologists with imperfect AI. In *26th International Conference on Intelligent User Interfaces*. 504–513. https://doi.org/10.1145/3397481.3450681

[32] Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed H Awadallah. 2022. On improving summarization factual consistency from natural language feedback. *arXiv preprint arXiv:2212.09968* (2022). https://doi.org/10.18653/v1/2023.acl-long.844

[33] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. 2018. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3820–3828. https://doi.org/10.1109/cvpr.2018.00402

[34] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. 2023. Evaluating the Impact of Human Explanation Strategies on Human-AI Visual Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW (Apr 2023). https://doi.org/10.1145/3579481

[35] 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* (2023).

[36] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121. https://doi.org/10.1609/hcomp.v8i1.7469

[37] António C Pacheco and Carlos Martinho. 2019. Alignment of player and non-player character assertiveness levels. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 15. 181–187. https://doi.org/10.1609/aiide.v15i1.5242

[38] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019). https://arxiv.org/abs/1907.12652

[39] Avery B Paxton, Erica Blair, Camryn Blawas, Michael H Fatzinger, Madeline Marens, Jason Holmberg, Colin Kingen, Tanya Houppermans, Mark Keusenkothen, John McCord, et al. 2019. Citizen science reveals female sand tiger sharks (Carcharias taurus) exhibit signs of site fidelity on shipwrecks. *Ecology* 100, 8 (2019), 1–4. https://doi.org/10.1002/ecy.2687

[40] Mahya Sadeghi, Parmit K Chilana, and M Stella Atkins. 2018. How users perceive content-based image retrieval for identifying skin images. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications:*

First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1. Springer, 141–148. https://doi.org/10.1007/978-3-030-02628-8_16

[41] Max Schemmer, Andrea Bartos, Philipp Spitzer, Patrick Hemmer, Niklas Kühl, Jonas Liebschner, and Gerhard Satzger. 2023. Towards effective human-ai decision-making: The role of human learning in appropriate reliance on ai advice. *arXiv preprint arXiv:2310.02108* (2023). https://arxiv.org/abs/2310.02108

[42] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.* 617–626. https://doi.org/10.1145/3514094.3534128

[43] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces.* 410–422. https://doi.org/10.1145/3581641.3584066

[44] Francesco Sovrano, Kevin Ashley, and Alberto Bacchelli. 2023. Toward Eliminating Hallucinations: GPT-based Explanatory AI for Intelligent Textbooks and Documentation. (2023). https://ceur-ws.org/Vol-3444/itb23_s3p2.pdf

[45] Philipp Spitzer, Niklas Kühl, and Marc Goutier. 2022. Training novices: The role of human-ai collaboration and knowledge transfer. *arXiv preprint arXiv:2207.00497* (2022). https://arxiv.org/abs/2207.00497

[46] Philipp Spitzer, Niklas Kühl, Daniel Heinz, and Gerhard Satzger. 2023. ML-Based Teaching Systems: A Conceptual Framework. *arXiv preprint arXiv:2305.07681* (2023). https://doi.org/10.1145/3610197

[47] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7433–7442. https://doi.org/10.1109/cvpr52729.2023.00718

[48] 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 (2020), 1229–1234. https://www.nature.com/articles/s41591-020-0942-0

[49] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. 2022. Perspectives in machine learning for wildlife conservation. *Nature communications* 13, 1 (2022), 792. https://www.nature.com/articles/s41467-022-27980-y

[50] Osman Tursun, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2023. Towards Self-Explainability of Deep Neural Networks with Heatmap Captioning and Large-Language Models. *arXiv preprint arXiv:2304.02202* (2023). https://arxiv.org/abs/2304.02202

[51] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q Vera Liao, and Jennifer Wortman Vaughan. 2023. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. *arXiv preprint arXiv:2302.07248* (2023). https://arxiv.org/abs/2302.07248

[52] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *Caltech-UCSD Birds-200-2011 (CUB-200-2011).* Technical Report CNS-TR-2011-001. California Institute of Technology. https://www.vision.caltech.edu/datasets/cub_200_2011/

[53] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–24. https://doi.org/10.1145/3359313

[54] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–15. https://doi.org/10.1145/3290605.3300831

[55] Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. 2021. Bayesian triplet loss: Uncertainty quantification in image retrieval. In *Proceedings of the IEEE/CVF International conference on Computer Vision.* 12158–12168. https://doi.org/10.1109/iccv48922.2021.01194

[56] Rebecca M Warner. 2012. *Applied statistics: From bivariate through multivariate techniques.* Sage publications.

[57] Uta Wilkens. 2020. Artificial intelligence in the workplace–A double-edged sword. *The International Journal of Information and Learning Technology* 37, 5 (2020), 253–265. https://doi.org/10.1108/ijilt-02-2020-0022

[58] Stephan Winter, Nicole C Krämer, Leonie Rösner, and German Neubaum. 2015. Don't keep it (too) simple: How textual representations of scientific uncertainty affect laypersons' attitudes. *Journal of Language and Social Psychology* 34, 3 (2015), 251–272. https://doi.org/10.1177/0261927x14555872

[59] Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond. *medRxiv* (2023). https://doi.org/10.21203/rs.3.rs-3661764/v1

[60] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–14. https://doi.org/10.1145/3544548.3581393

[61] Daniel Zhang, Yang Zhang, Qi Li, Thomas Plummer, and Dong Wang. 2019. Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1221–1232.