The Impact of Imperfect XAI on 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

Explainability techniques are rapidly being developed to improve human-AI decision-making across various cooperative work settings. Consequently, previous research has evaluated how decision-makers collaborate with imperfect AI by investigating appropriate reliance and task performance with the aim of designing 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. Thus, it is crucial to understand how imperfect XAI affects human-AI decision-making. In this work, we contribute a robust, mixed-methods user study with 136 participants to evaluate how incorrect 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. Our findings reveal the influence of imperfect XAI and humans' level of expertise on their reliance on AI and human-AI team performance. We also discuss how explanations can deceive decision-makers during human-AI collaboration. Hence, we shed light on the impacts of imperfect XAI in the field of computer-supported cooperative work 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, Explainable AI for Computer Vision

ACM Reference Format:
Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 183 (April 2024), 39 pages. https://doi.org/10.1145/3641022

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,

*Both authors contributed equally to this research.

Authors' addresses: Katelyn Morrison, kcmorris@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Philipp Spitzer, philipp.spitzer@kit.edu, Karlsruhe Institute of Technology, Germany; Violet Turri, vturri@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Michelle Feng, msfeng@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Niklas Kühl, kuehl@uni-bayreuth.de, University of Bayreuth, Germany; Adam Perer, adamperer@cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

ACM 2573-0142/2024/4-ART183
https://doi.org/10.1145/3641022

causing frustration and resulting in potentially harmful decisions. As a result, designing and developing human-centered explanations has become a core theme in computer-supported cooperative work (CSCW) and human-AI collaboration research [25, 53, 89]. Tangentially, recent work has proposed new explanation techniques that leverage machine learning models to explain the prediction of another machine learning model [8, 13, 34, 41, 48, 71, 77, 81, 85, 90, 97]. 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" [34, 41, 81]. Aside from natural language explanations, another subset of recent work proposes advanced content-based image recognition techniques to generate example-based explanations [8, 71, 90]. These types of approaches to explainability introduce another level of uncertainty in the collaboration between the decision-maker and the AI, as explanation models are imperfect.

CSCW, and more specifically human-AI collaboration, is prevalent across high-stakes scenarios [17, 54, 83, 88]. For instance, radiologists collaborate with AI to identify abnormalities in medical imagery [83]; conservationists use AI to help monitor biodiversity [12], and humanitarian aids use AI to help identify damaged buildings after natural disasters or armed conflicts from satellite imagery [57, 98]. 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 [12, 65]. 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 [47].

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 [73] or learning systems [79], 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 context information to appropriately rely 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 context 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 [33].

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. [64] use the term 'explanation fidelity' while Kroeger et al. [48] 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 CSCW and human-computer interaction (HCI) communities. 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 [18]. 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, the information provided, and their representations) impacts human-AI collaborations across levels of expertise [18, 46].

Collaborating with imperfect AI is not a new concept to the CSCW and HCI communities [33, 47, 54]. 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 [33, 64]. 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.*, [20, 44]), few studies explore the impact that incorrect or "noisy" explanations have on human-AI collaboration [33, 39, 64, 86]. Recent work has used technical approaches to mitigate "noisy" or incorrect natural language explanations [39], and Kroeger et al. [48] propose 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 growing use of machine learning models to explain other machine learning models in high-stakes decision-making scenarios, we argue that it is necessary to understand how humans interact with 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 CSCW researchers. With these topics under-explored in current literature, we present the following research questions:

- **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**: How do different types of explanations change the effect that the correctness of explanations has on appropriate reliance and complementary team performance?
- **RQ4**: To what extent do incorrect and correct explanations deceive decision-makers with different levels of expertise?

To address our research questions, we employ an imperfect AI model for a bird species identification task [34]. We focus on bird species identification because the use of AI for wildlife conservation efforts among experts and non-experts is a rapidly growing field in research and practice [84]. Furthermore, it is less difficult to find people with varying levels of expertise in birding than in radiology who will have time to participate in our study.

Through a mixed-methods study, we answer our research questions by asking participants to classify bird images in two phases: without any advice from AI (phase 1) and then showing the AI's advice and explanation (phase 2). To answer **RQ** 1, we design a research model based on phenomena from relevant research in CSCW and conduct rigorous moderation analyses based on [32]. Our analyses leverage the appropriate reliance metrics defined by Schemmer et al. [74]. We design our study to be within-subjects for the correctness of the explanation and the assertiveness of explanations allowing us to answer **RQ** 2 and between-subjects for the explanation modality allowing us to answer **RQ** 3. Moreover, we calculate the magnitude of deception caused by incorrect explanations compared to correct explanations across both explanation modalities and levels of expertise to account for the impact of imperfect XAI on humans' decision-making behavior. This measurement gives us insight into **RQ** 4. Lastly, we conduct an inductive content analysis based on Gioia et al. [29] to assess the open-ended responses from participants to gain insight into designing for imperfect XAI in human-AI collaborations.

As a result of our study, we contribute the following to the CSCW community:

- *Research Model for Human-AI Collaboration with Imperfect XAI*: We propose a research model for the moderating roles of the decision-maker's level of expertise and the explanation's assertiveness on the effect of the correctness of explanations on appropriate reliance.
- *Novel Empirical Study*: To validate our proposed research model, we conduct the first empirical investigation that explores the moderation of assertiveness of explanations and the level of expertise on the impact that the correctness of explanations has on appropriate reliance. We do this on a human-AI collaboration scenario across two different modalities of explanations: natural language explanations and visual, example-based explanations. We also investigate the impact on complementary team performance. Our findings inform designers of human-AI collaboration systems on how to deploy imperfect XAI from a user-centric perspective.
- *Novel Metric for Impact of Imperfect XAI:* We contribute a novel metric to the human-AI decision-making field accounting for the impact of incorrect explanations on humans' decision-making behavior when collaborating with AI. Specifically, we propose the Deception of Reliance (DoR) caused by imperfect XAI. With DoR, we investigate to what extent imperfect XAI deceives humans.
- *Qualitative Insights*: We provide insights on how decision-makers, regardless of expertise, prefer the tone of explanations to align with factors related to the AI's behavior and the impact on decision-makers. These insights can inform future works to conduct new evaluations.

The remainder of this article is structured as follows: We present related literature on imperfect AI systems, human-AI collaboration, and explainability (Section 2, p. 4) before outlining our theoretical development (Section 3, p. 6) and methodology for conducting an empirical study (Section 4, p. 9). We then present the results of our work for two different types of explanations (Section 5, p. 16). After that, we discuss the implications (Section 6, p. 24) and limitations (Section 7, p. 27) of our findings. Finally, we end our article with a brief conclusion (Section 8, p. 28).

2 RELATED WORK

We situate our contributions in relation to past work about decision-making with imperfect AI/XAI, the impacts of end-user expertise on human-AI collaboration, and the impact of explanation type on human-AI collaboration.

2.1 Decision-Making with Imperfect AI/XAI

Numerous studies in CSCW and HCI have investigated the impact that imperfect AI has on human-AI collaboration (*e.g.*, [7, 47, 86]). Kocielnik et al. [47] offer three techniques for setting user

expectations about the performance of an imperfect AI system, including an accuracy indicator, example-based explanations, and performance control. Through a user study with an AI-powered scheduling assistant, the authors demonstrate the efficacy of their techniques in maintaining user satisfaction and acceptance. The authors also demonstrate that the nature of system errors can impact user perception.

Several recent studies investigate how programmers collaborate with Copilot, an imperfect AI programming assistant (*e.g.*, [10, 21, 86]). One of those studies specifically looks at how to convey the uncertainty of outputs from Copilot [86]. By highlighting code that is most likely going to be edited by the programmer instead of highlighting based on the probability of the code being generated, they observe that programmers arrive at solutions faster. Furthermore, Dakhel et al. [21] conclude that GitHub Copilot is valuable for expert programmers but something non-expert programmers should be cautious about.

Previous studies explore the impact that revealing the confidence of the model's prediction has on the human-AI team (*e.g.*, [7, 45, 82]). For example, Kim and Song [45] investigate the effect of various framings and timings for presenting the performance of an AI system on user acceptance. Through their user study, the authors reveal that users find AI advice to be more reasonable when it is not accompanied by information about AI system performance than when it is. In the case that AI system performance is shown, users consider AI advice to be more reasonable when system performance is displayed before they make a decision rather than afterward. However, communicating uncertainty for image classification in a visual format is under-explored. Recent work conducts a user study to see how showing the confidence of an AI prediction through a green hue on an image impacts reliance on AI [82].

Fewer studies investigate the impact that imperfect XAI has on human-AI collaboration [64]. Similar to our contributions, Papenmeier et al. [64] investigate the impact that explanation fidelity has on user trust. They present a user study where participants collaborate with AI of different accuracies and XAI with different levels of correctness to determine if a Tweet should be published or not based on its content. While Papenmeier et al. [64] investigate how an explanation's level of correctness impacts trust, they do not explore the role that a user's level of expertise plays.

2.2 Domain Expertise & Human-AI Complementarity

There has been a growing interest in understanding the impact that the decision-maker's domain expertise has on human-AI collaborations [11, 18, 23, 28, 60, 61, 80, 83, 99]. One recent study investigates the impact of decision-makers' domain expertise on task accuracy in a high-stakes human-AI collaboration task [18]. Calisto et al. [18] also look at the impact that the assertiveness of natural language explanations has on human-AI collaboration. In their study, they present natural language explanations with varying levels of assertiveness to radiologists with different years of experience on a mammogram classification task. Their main analysis consists of the radiologists' task performance. Unlike Calisto et al. [18], our experiment collects the human's initial decision before showing the AI's advice to the human, allowing us to measure appropriate reliance and assess for complementary team performance.

One study investigates how the level of expertise for Arabic or Indian Numerals from various versions of MNIST impacts task accuracy and model perception [28]. A similar study shows clinicians with various levels of expertise four different types of explanations, including visual, example-based explanations [83]. Similar to our study design, they show participants the three most similar example images for the example-base explanations. Another study investigates how practitioners with different levels of expertise perceive explanations that were implemented in a manufacturing industry context [99]. They observe that practitioners with a higher level of expertise are more accepting of the explanations. Recent work proposes a research model to identify the

impact decision-maker's level of expertise has on trust in XAI [11]. Their research model does not consider the correctness or tone of explanations. Through their online, AI-supported chess experiment, they observe that expertise negatively affects trust.

While numerous previous works investigate the impact of domain expertise, to the best of our knowledge, none explore the impact of the correctness of explanations and the level of expertise on the decision-maker's reliance behavior together.

2.3 Explanation Modality

In human-AI collaboration scenarios, the human decision-making behavior depends on the type of explanation provided (*e.g.*, [38]). To validate why we evaluate our research model for two different types of explanations (*i.e.*, natural language and visual, example-based), we synthesize previous work that compares multiple different modes of XAI.

Several studies investigate the use of example-based explanations in human-AI decision-making [16, 20, 24, 38, 96]. Cai et al. [16] propose normative and comparative explanations, different types of example-based explanations. They evaluate how these explanations impact end-users' understandability and perception of the AI model in a drawing guessing game. The authors find the normative explanations to help users better understand how the AI makes decisions when the model prediction is incorrect. Another paper investigates example-based explanations in a slightly different format from Cai [96]. They similarly found the example-based explanations improve the users' appropriate trust in the classifier.

In a different study, Du et al. [24] examines the effect of different explanation modalities on clinical practitioners' reliance behavior. The authors show no significant differences between example-based explanations and feature-based explanations. However, they find that different types of practitioners prefer different modalities from a user-centric perspective. More recent work compares example-based explanations to feature importance through a think-aloud study [20]. From their mixed-methods study, Chen et al. [20] outline three types of intuition that are employed when decision-makers reason about AI predictions and explanations, including task outcomes, features, and AI limitations. The authors use these three intuition types to explain study results in which feature-based explanations lead to overreliance on AI while example-based explanations improve human-AI performance.

Several recent works have compared text-based explanations to visual explanations (*e.g.*. [43, 69, 80]). For example, Kim et al. [43] analyze a unified explanation technique from a human-centric point of view. In their work, Kim et al. [43] explore visual and text explanations in a user study. They investigate users' preferences for different AI interfaces. The authors conclude that users prefer local visual explanations in such interfaces over text-based ones. Another study compares six different types of explanation modalities in an extensive user study [69]. Robbemond et al. [69] look into the impact of text, audio, graphics, and combinations of the previous modalities on decision-makers' reliance on decision support systems. Their results show that combinations of different explanation modalities lead to higher user performance. Szymanski et al. [80] conduct a similar study, only evaluating visual and textual explanations.

Based on findings from previous studies that evaluate the impact of various explanation modalities on human-AI collaboration, we choose to explore visual, example-based explanations and natural language explanations.

3 THEORETICAL DEVELOPMENT

The increasing use of explanations to reveal the rationale behind AI predictions has led to a rise in research examining the impact of explanations on decision-makers' behavior [22, 51, 73]. As imperfect AI is utilized more within high-stakes contexts, such as decision-making in the medical

sector, research has focused on the impacts of potentially inaccurate AI advice on humans' decision-making [47, 50, 69]. However, there are very few works investigating how **imperfect XAI** impacts humans' appropriate reliance on AI advice [64].

Thus, in this work, we draw from the conceptualization on appropriate reliance previous research has established [7, 73, 74, 76]. We specifically build on the conceptualization presented by Schemmer et al. [74] in Figure 1 by adding a new dimension to consider when investigating appropriate reliance in human-AI collaboration: The correctness of XAI advice. We simplified the correctness of XAI advice to be a binary case of correct or incorrect in Figure 1.

The introduction of this new dimension unveils previously unexplored avenues within the realm of human-AI collaboration in CSCW, thereby offering a conceptual framework to delve into a more profound comprehension of human decision-making behavior with an AI collaborator. As a result, researchers can calculate more specific metrics regarding human decisions after receiving the AI and XAI advice. For example, the ratio of under-reliance based on a correct prediction and incorrect explanation could be different than when based on a correct prediction and correct explanation. These types of scenarios should not be overlooked when investigating human-AI collaborations in work settings. In Appendix A.1, we provide additional details of the newly introduced metrics.

With this new dimension for analyzing human-AI collaborations, we investigate the effect that imperfect explanations have on humans' decision-making; we investigate this relation in a sequential decision-making scenario. Based on the constructs of relative AI reliance (RAIR) and

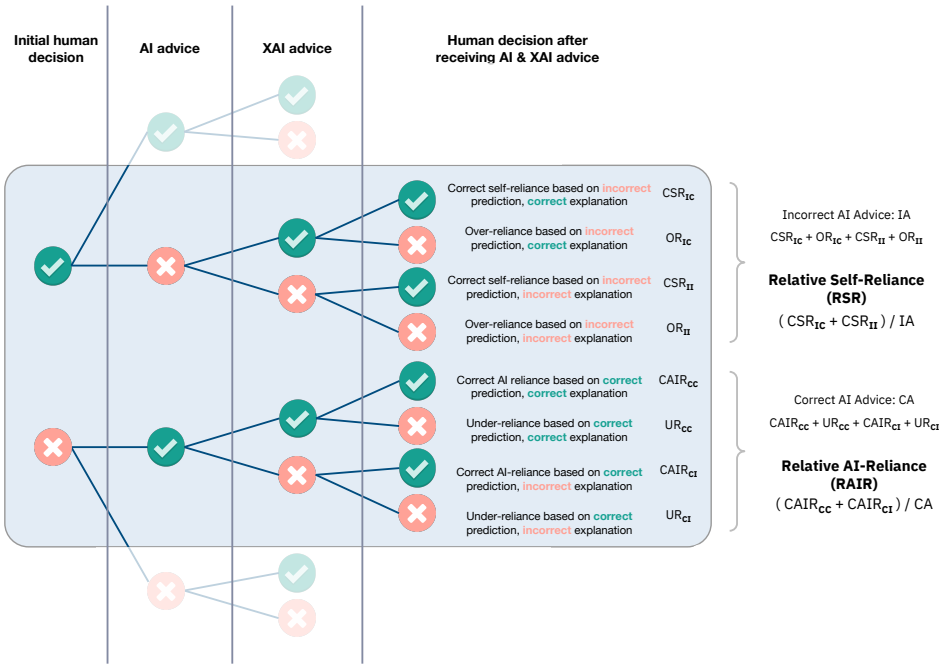

Fig. 1. Different paths that human decision-makers could follow based on receiving AI and XAI advice. This figure expands that presented by Schemmer et al. [74] by contributing the XAI advice dimension. The XAI advice is simplified into correct and incorrect explanations. The green checkmarks represent correct advice/decisions, while the red 'x' represents incorrect advice/decisions.

relative self-reliance (RSR), we account for the appropriateness of reliance [74][1]. RAIR comprises the cases in which the human corrects their initially incorrect decision by overriding it with the correct AI advice. On the other hand, RSR comprises all cases in which the human initially makes a correct decision, the AI system gives incorrect advice, and the human rightly dismisses this advice. Thus, we use appropriateness of reliance as the dependent variable in our research model (see Figure 2). In the recent work of Schoeffer et al. [75] the authors investigate how explanations affect distributed fairness in AI-assisted decision-making. Their study shows that task-relevant explanations impact humans' reliance behavior into increasing stereotype-based errors. We apply these findings to our research model and assume that for the cases in which the AI provides correct advice, explanations will affect RAIR. Accordingly, we hypothesize:

Hypothesis 1: The correctness of explanations impacts humans' relative AI reliance in human-AI decision-making.

Hypothesis 2: The correctness of explanations impacts humans' relative self-reliance in human-AI decision-making.

One crucial factor in this interrelation between imperfect explanations and humans' appropriate reliance is the level of domain knowledge that humans possess. Previous work shows that humans' level of expertise can influence their decision-making [11, 18]. Related research in information systems investigates the role of domain knowledge in decision-making. Erjavec et al. [26] show in their behavioral experiment in online supply chains that domain knowledge positively impacts humans' confidence in decision-making. Similarly, Dikmen and Burns [23] analyze humans' reliance on AI when possessing different levels of domain knowledge. In their study, the authors provide an imperfect AI and argue that higher domain knowledge leads to less trust in AI. With this impact of domain knowledge on human-AI decision-making, we intend to examine how the effect of imperfect explanations on appropriate reliance is influenced by humans' level of expertise. Humans with high domain-specific knowledge demonstrate an enhanced ability to discriminate between erroneous explanations and accurate ones with greater rigor [52]. This discernment is facilitated by their extensive expertise, which empowers them to readily identify and discern false information [9]. Thus, in our study, we hypothesize:

Hypothesis 3: Humans' level of expertise moderates the effect of the correctness of explanations on RAIR.

Hypothesis 4: Humans' level of expertise moderates the effect of the correctness of explanations on RSR.

Humans' information processing is not only influenced by what they are provided but also by the way this information is provided. Previous research demonstrates that language style can impact humans' decision-making behavior [18, 37, 49]. Huang et al. [37] show that the level of assertiveness in reviews affects humans' online review persuasion. Similarly, Kronrod et al. [49] investigates the effect of the level of assertiveness on the tone of language. They find that there is a relationship between the tone of language and its level of assertiveness. Next to the psychological research field, recent research in HCI examines how assertiveness in explanations affects humans' performance when interacting with AI [18]. They show that the level of assertiveness does impact humans' trust when collaborating with AI. In their work, they reveal that humans are more likely to follow AI advice when the explanation is presented in their own communication style. With those findings and related research on the impact of assertiveness on humans' decision-making, it

[1]Appropriateness of reliance is the quantitative measurement for appropriate reliance. These terms will be used interchangeably throughout the article.

is reasonable to hypothesize that assertive explanations will impact humans' reliance on AI. Hence, we hypothesize:

Hypothesis 5: Explanations' level of assertiveness moderates the effect of the correctness of explanations on RAIR.

Hypothesis 6: Explanations' level of assertiveness moderates the effect of the correctness of explanations on RSR.

Our research model shown in Figure 2 summarizes the hypotheses that we test in our mixed-methods study. Overall, with this research model, we test for moderating effects of the level of expertise and level of assertiveness on the effect of the correctness of explanations on appropriate reliance.

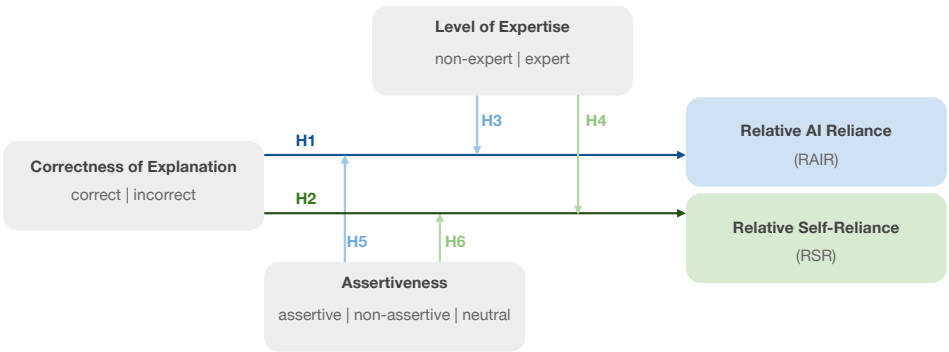

Fig. 2. Research model for collaborating with imperfect XAI systems. We analyze the moderation of the level of expertise and assertiveness on the effect of the correctness of explanation on RAIR and RSR.

4 METHODOLOGY

In this section, we describe the task of bird species identification, the experiment design, the recruitment process of participants, and the development of the explanations we use in the study. Finally, we end this section by outlining the data we select to use for the study and the metrics we use to analyze the results.

4.1 Task Domain: Bird Species Identification

Computer-supported cooperative work is becoming core to wildlife conservation efforts [14, 31, 84]. With mobile devices becoming increasingly powerful, non-experts and experts alike can use AI-powered applications like the Merlin Bird ID app [2] to identify bird species for monitoring biodiversity and learning about birds. The popularity of both birding and AI-based image classification techniques suggests that bird species identification would be a sensible domain to investigate our research techniques. Furthermore, this is a task for people with a wide range of expertise.

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 [5]. 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 [83]. 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. [36], Kayser et al. [41] propose an imperfect natural language explanation 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.

Previous studies that focused on human-centered XAI and human-AI collaboration also use the domain of bird species identification to understand better human-AI collaboration (*e.g.*, [15, 44, 59]). However, few previous works focus on human-AI collaboration for decision-making in the wildlife conservation domain overall. Yet, the field of AI for wildlife conservation is rapidly growing [84] and could benefit from research related to CSCW and HCI.

4.2 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 3 and is divided into four different parts ($A - D$), which we explain in further detail below.

The study begins with part A in Figure 3: In a bird identification test, we assess participants' expertise in classifying six different species of birds[2]. 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 [44], this method is subjective, and participants may self-perceive their skills differently. Kazemitabaar et al. [42] measure participants "experience-level" through log data

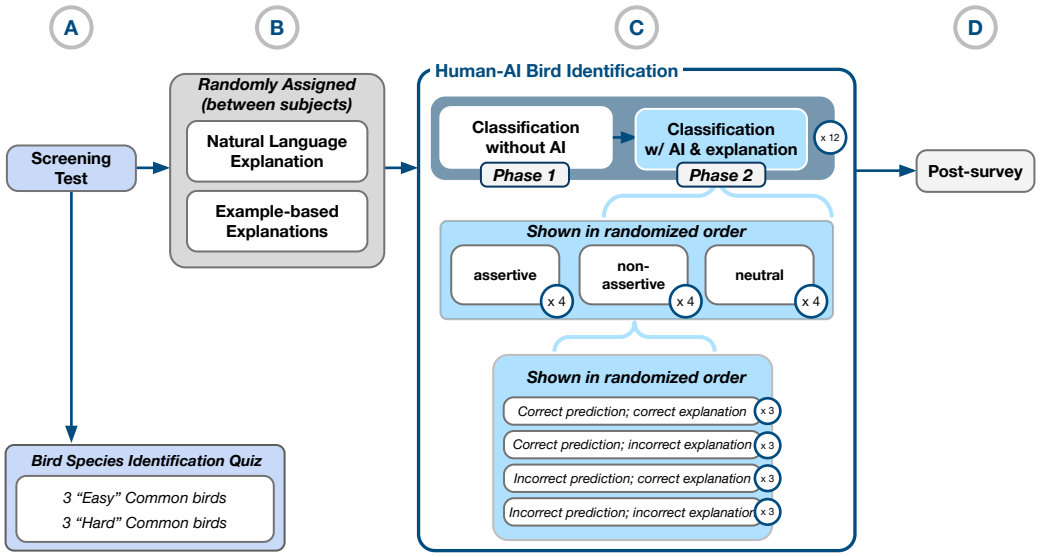

Fig. 3. We conduct a rigorous mixed-methods study leveraging a mixed design. Before participants start the task, they are shown a screening test (A). For the human-AI bird identification task, participants are assigned an explanation modality (B). During the task, participants are shown explanations with different levels of assertiveness and different scenarios of correctness (C). Lastly, participants complete a post-survey (D).

[2]Specific birds used for the bird identification test are reported in Appendix A.2, p. 36.

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 3), participants are randomly assigned to one of two explanation types. Similar to previous studies [19, 68, 69, 80], the treatments differ in the explanation modality participants receive: *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 [34]. We choose to also look at example-based explanations as recent studies focus on this modality in human-AI collaboration [16, 20, 38, 44]. However, recent studies show that example-based explanations have potential benefits. Chen et al. [20] show that example-based explanations improve humans' performance, so much so that it leads to complimentary team performance. With promising results from previous research and numerous clinical decision-support tools proposing to incorporate example-based explanations (*e.g.*, [8, 71]), we find it necessary to investigate the effect of example-based explanations on humans' appropriate reliance in the context of imperfect XAI.

The human-AI bird identification task (part C of Figure 3) consists of two phases. For each treatment condition, the participants are asked to initially identify the bird species from an image (phase 1 in Figure 4). 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 in Figure 4). The structure of phases one and two are corroborated with previous work [30]. 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.

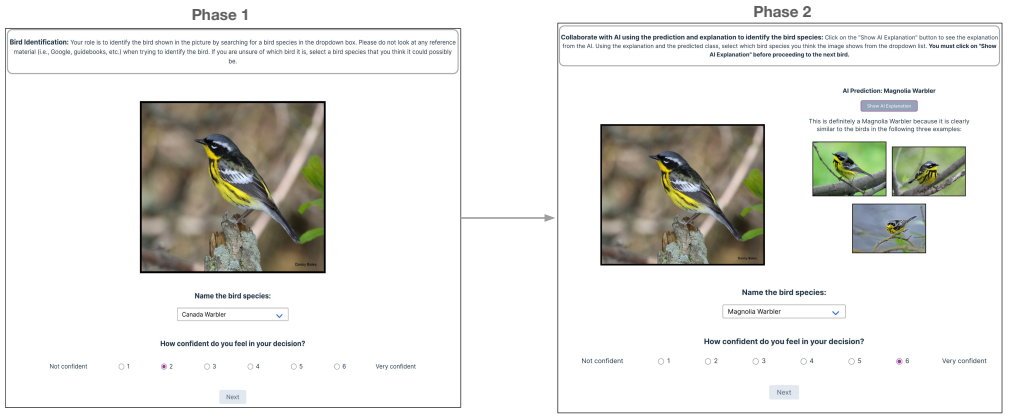

Fig. 4. Example of the two phases for a single bird image that a participant is shown in the study. This specifically shows a Magnolia Warbler (correct prediction, correct explanation), and this participant is assigned to the example-based explanations. For this bird, the participant is shown an assertive explanation.

As the AI that we are utilizing for identifying the bird species is not perfect [34], the predictions and explanations provided can be incorrect. In order 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:

- **CC**: correct prediction and correct explanation
- **CI**: correct prediction and incorrect explanation
- **IC**: incorrect prediction and correct explanation
- **II**: incorrect prediction and incorrect explanation

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*[3]. 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.

After finishing the task, participants must fill out an additional questionnaire (part D of Figure 3). Here, we qualitatively assess participants' ability to properly rely on the AI based on the explanations that they were shown. Thus we ask them: "*Under what circumstances would you prefer assertive (e.g., "definitely", "clearly") versus non-assertive (e.g., "might be", "appears to be") versus neutral explanations and why?*". Aside from this open-text question, we also ask participants about their occupations and the regions in North America that they are most familiar with in terms of bird species.

4.3 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 [87] dataset and explanations from the Generating Visual Explanations model [34]. The original dataset consists of 11,788 images of 200 different 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 and the visual, example-based 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[4] and explanation type. In some cases, the example-based and natural language explanations for a single bird are used. 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 [55, 95]. We use descriptions from the Cornell Lab of Ornithology All About Birds Guide [4] to corroborate our classification for each explanation. Examples of incorrect and correct natural language explanations are provided in Figure 5.

[3]Figure 6 provides examples of *assertive*, *non-assertive*, and *neutral* explanations.
[4]The four categories are identified in Section 4.2.

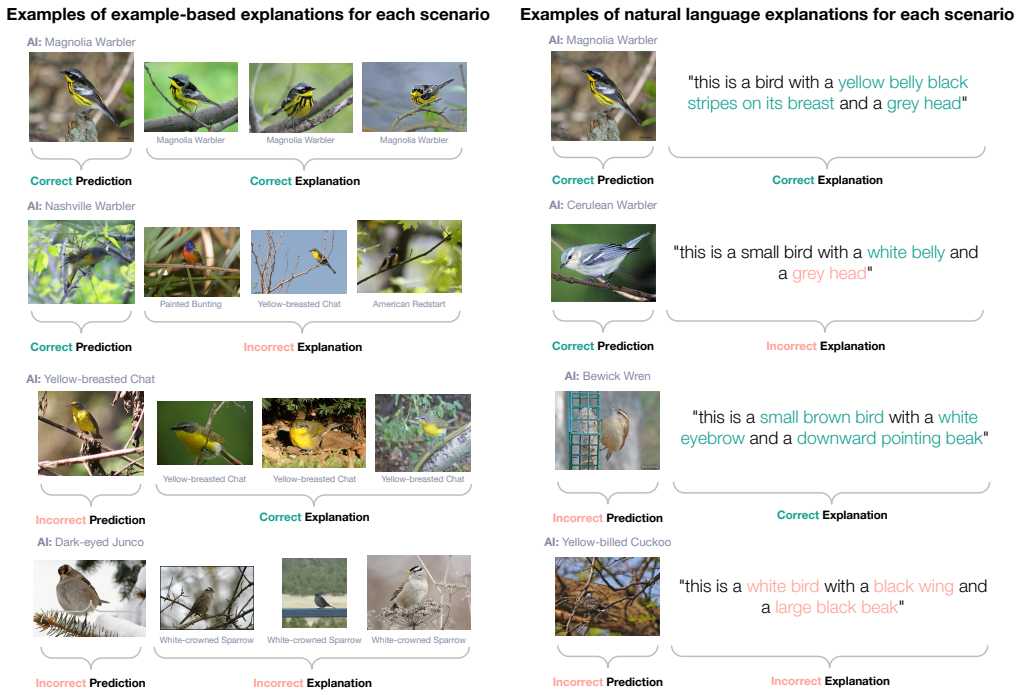

Fig. 5. Representative examples of the example-based and natural language explanations for each scenario: CC, CI, IC, and II. The class of the example-based images in the explanation is not shown to participants during the study. The red and green coloring on the natural language explanations was not shown during the study. This is only provided in the figure to guide the reader. The natural language explanation for the Cerulean Warbler is incorrect because this bird species does not have a grey head. The natural language explanation for the Yellow-billed Cuckoo is incorrect because this bird species is brown with a white belly, has a gold and black beak, and does not have a black wing.

For the example-based explanations, we define a correct explanation as the three most similar images belonging to the predicted class (as shown in phase 2 of Figure 4). We define an incorrect explanation to be most similar to at least one image that is not of the predicted class. This means that incorrect example-based explanations incorporate logical errors as the examples shown are dissimilar from the predicted class. Moreover, such incorrect explanations can have an inconsistency as the examples shown might differ in the classes shown. However, we only choose to show participants incorrect explanations that have at least two images that are not of the predicted class. For example, in Figure 5, the AI correctly predicts a Nashville Warbler; however, the three most similar examples are a Painted Bunting, a Yellow-Breasted Chat, and an American Redstart. In some cases where the advice is correct and the explanation is incorrect[5], the explanation may align with the ground truth class. For example, for a Tennessee Warbler, the AI predicts an Orange-crowned Warbler (incorrect advice since the wrong bird species is predicted), but the three most similar examples are all of Tennessee Warblers (incorrect explanation since the examples' bird species do not align with prediction). It's possible that a model could be relying on spurious patterns to make

[5]The explanation does not align with the predicted class.

classifications [67]. Since we are dealing with an imperfect AI, we do not choose to exclude such cases from our dataset.

4.4 Explanation Modalities

Natural Language Explanations. The natural language explanations were generated by the model proposed by Hendricks et al. [34]. We followed the PyTorch implementation [6] 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 5 is: "`this is a bird with a yellow belly black stripes on its breast and a grey head`".

Example-Based Explanations. Previous work creates example-based explanations, specifically normative explanations, by calculating the Euclidean distance between the given image and the images in the dataset [16]. Another study generates the example-based explanation by calculating the L_2 distance of the embedded features [58]. We generated the example-based explanations by following methods used in previous works [8, 83]. As done by Tschandl et al. [83] and Barata and Santiago [8], we calculate the cosine similarity between the extracted feature vector of the given image and the rest of the extracted feature vectors of the images in the training set. Unlike Barata and Santiago [8], we choose not to take the example's ground truth class into consideration. The extracted features from the images were provided by Hendricks et al. [34]. Because the model is not perfect, the example-based explanations are also not perfect. For example, even though the model correctly predicted an image of a Nashville Warbler in Figure 5, the three most similar images are of three different birds. For this study, we consider an example-based explanation to be incorrect if two of the three examples are of a different class than the predicted class. Inspired by Ford et al., we choose to show participants the three most similar examples [28].

Assertiveness of Explanations. Following previous studies [18, 62, 93], 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

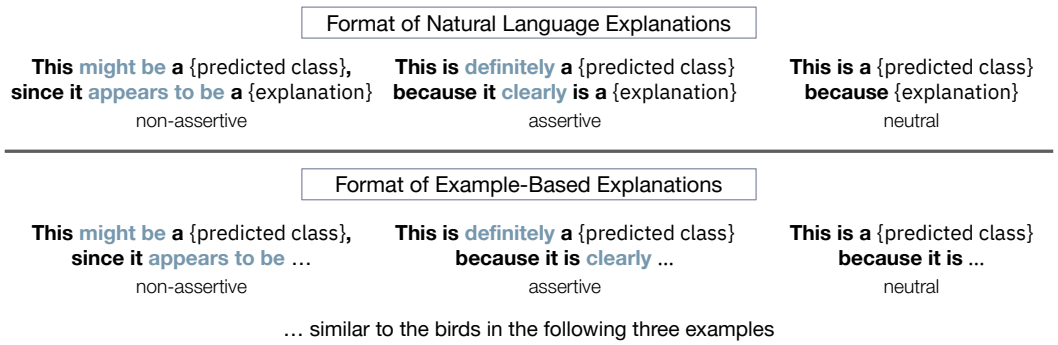

Fig. 6. The three different language tones that an explanation could have in terms of assertiveness for the natural language and example-based explanation modality. Non-assertive explanations included the words "might be" and "appears to be". Assertive explanations included the words "definitely" and "clearly". Neutral explanations did not include any additional adjectives.

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 6. To our knowledge, there is no literature to rationalize how to appropriately present assertiveness visually for example-based explanations, so we opt to use natural language in combination with the example-based explanations.

4.5 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 as previous research has indicated that this platform is a reliable source of research data [63, 66]. We apply a custom filter on Prolific to target individuals who currently work in a field related to nature, science, the environment, or animals. Participants receive compensation that is above minimum wage. Overall, we try to limit recruitment to only address people with prior knowledge of birding to minimize the prevalence of novices' randomly guessing bird species identification. After excluding participants who provide incomplete and fake responses (i.e., lorem ipsum response to our survey question), we have 136 people complete our study. In order to determine if a participant is familiar with birding, participants take a bird identification test (phase A in Figure 3). The participant's score on the bird identification test is used to determine whether they are considered a non-expert or an expert. Details related to clustering participants based on their test scores are provided in Section 5.

4.6 Quantitative and Qualitative Metrics

Quantitative Metrics. We quantitatively calculate appropriate reliance across the four dimensions defined by Schemmer et al. [74]: correct AI reliance, correct self-reliance, under-reliance, and over-reliance. Accordingly, correct AI reliance measures the number of correct decisions when the human's initial decision is incorrect, and the human is rightly taking over the correct AI advice. Correct self-reliance is when the human initially makes the correct decision and does not overwrite their decision with the incorrect AI advice. On the other hand, under-reliance reflects the case in which the human initially makes an incorrect decision and does not adhere to the correct AI advice. On the other hand, over-reliance represents the scenario in which the human makes an initial correct decision but overrides her own decision with incorrect AI advice. Following the appropriate reliance metrics defined by Schemmer et al. [74], we calculate RSR and RAIR to account for the appropriateness of reliance.

With the new dimension for XAI advice, we can separately measure RAIR and RSR for correct and incorrect explanations and derive its impact on appropriate reliance. In order to measure this impact, we look at the Deception of Reliance (DoR) caused by imperfect XAI. For RAIR, we can apply the following:

$$DoR_{RAIR} = RAIR_C - RAIR_I. \tag{1}$$

In this equation, the subscript I represents incorrect explanations, whereas the subscript C represents correct explanations. We can compute the same for RSR:

$$DoR_{RSR} = RSR_C - RSR_I. \tag{2}$$

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 Gaussian distance

in the RAIR-RSR space between incorrect and correct explanations:

$$DoR(RAIR, RSR) = \sqrt{(RAIR_C - RAIR_I)^2 + (RSR_C - RSR_I)^2}. \quad (3)$$

According to Schemmer et al.'s conceptualization of Appropriateness of Reliance [74], this results in the following:

$$DoR_{AoR} = AoR_C(RAIR, RSR) - AoR_I(RAIR, RSR). \quad (4)$$

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.*, [7, 30, 33]), we can calculate the human-AI team performance to determine if CTP exists. Following the constructs defined in those previous works, we determine if CTP exists by calculating the participants' performance in identifying the bird species **before** and **after** they see the AI advice and compare this to the performance of the model on the twelve birds images shown to the participant. 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%.

Qualitative Metrics. We also conduct an inductive content analysis of the open-ended responses to better understand the participants' preferences regarding assertiveness. As a reminder, in the end-survey, we ask participants specifically "*Under what circumstances would you prefer assertive (e.g., "definitely", "clearly") versus non-assertive (e.g., "might be", "appears to be") versus neutral explanations and why?*". To analyze these responses, we follow the established procedure of Gioia et al. [29] and screen the answers in three coding workshops. In the first workshop, the first and second authors of this article initially screen the answers and applied open coding [35] to extract and aggregate core constructs of participants' answers. In this procedure, both authors discuss their findings and align their understanding of relevant constructs. Through a second coding workshop, we apply axial coding to derive subcategories of these constructs and align those with the data. In a final workshop, we distill those emerging themes and derive aggregated dimensions [94].

5 RESULTS

To answer our research questions and confirm or reject our hypotheses, we conduct rigorous statistical and qualitative analyses. We measure appropriate reliance based on metrics defined in previous work: Relative AI reliance (RAIR) and relative self-reliance (RSR) [74]. By doing so, we answer **RQ** 1 and **RQ** 2 in Section 5.2. We also calculate the participant's task accuracy before and after receiving AI and XAI advice to answer **RQ** 3 (see Section 5.3). This way, we can determine whether complementary team performance exists [33]. Additionally, based on the new metric, Deception of Reliance (DoR), we measure to what extent explanations deceive humans, answering **RQ** 4 in Section 5.4. Lastly, we qualitatively analyze open-ended responses through rigorous inductive content analysis in Section 5.5. For all of our research questions, we look at two different types of explanations: natural language explanations that are focused on specific features present in the image and visual, example-based explanations showing the top three most similar example images from the training set. While our analyses look at both modalities, we do not intend to compare them directly. Therefore, we do not conclude one modality is better or worse than the other.

5.1 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 3). These four features represent participants' scores in correctly identifying the family and species of the easy and the difficult bird images. By clustering the 136 participants into the expert and non-expert group, we end up with 83 experts and 53 non-experts. With this clustering, the average bird identification test score (summing up all four scores in the identification test) for non-experts is 38.99%(STD = 11.42%) while the average test score for experts is 83.84%(STD = 12.30%)[6]. Of the 83 experts, 42 see example-based explanations, and 41 see natural language explanations. Of the 53 non-experts, 25 see example-based explanations, and 28 see natural language explanations. In terms of the fields that the 136 participants represent, 45 participants have an occupation primarily related to biology, conservation, and/or the environment. 26 have an occupation primarily related to engineering and/or technology; 30 are either researchers, students, or affiliated with education in some other way; 24 have occupations in miscellaneous industries; and 11 are retired.

5.2 Moderating Effects in Imperfect XAI Research Model

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 utilizing the process macro model of Hayes [32]. An overview of the regression analyses is presented in Table 1.

Table 1. Moderation analyses of the correctness of natural language and example-based explanations on RAIR and RSR with the level of expertise and assertiveness as moderators. The coding of assertiveness used for the moderation analyses is provided.

Coding of assertiveness

assertiveness	Z1	Z2
neutral	0	0
non-assertive	1	0
assertive	0	1

	RAIR				RSR			
	Natural Language		Example-Based		Natural Language		Example-Based	
	coeff	p	coeff	p	coeff	p	coeff	p
const	1.26	.00	.43	.17	-17.16	.98	-3.60	.00
corr	.57	.29	1.02	.04	13.25	.98	-.4.36	.74
exp	2.12	.00	-1.25	.00	15.89	.98	3.25	.00
Z1	-.46	.24	.26	.47	.14	.26	-.09	.83
Z2	.00	1.00	.00	1.00	-.31	.58	.09	.83
exp x corr	-1.00	**.05**	-1.04	**.03**	-13.33	.98	-.73	.57
Z1 x corr	.46	.44	-.03	.95	-.28	.71	-.45	.55
Z2 x corr	.19	.74	-.16	.77	.90	.23	-.43	.55

[1] *corr — correctness*; *exp — level of expertise*

[6]Participants performance on the bird identification test is shown in Figure 10, Appendix Section A.2.

5.2.1 Participants' level of expertise moderates the effect of the correctness of explanations on RAIR for natural-language explanations. As theoretically developed in Section 3, 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 results of this moderation analysis are shown in Table 1 (a detailed view is shown in Table 3 on p. 37 in the Appendix A.3).

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 [32] and Warner [91]. The results of the regression analysis show that there is no direct effect between assertiveness and RAIR (coeff = .04, p-value = .77). Thus, we confirm hypotheses 1 and 3 and reject hypothesis 5 for natural language explanations.

5.2.2 Participants' level of expertise moderates the effect of the correctness of explanations on RAIR for example-based explanations. Next, we present the moderation analysis of example-based explanations on RAIR. We set up the analysis for example-based explanations the same as the analyses for natural language explanations (see Section 5.2.1). As seen in Table 1, there is a significant moderation effect of the level of expertise (coeff = −1.04, p-value = .03). The negative coefficient signals that this moderation is higher for non-experts than for experts. The correctness of the example-based explanations has a positive coefficient (coeff = 1.02, p-value = .04), and thus, correct explanations have a positive impact on RAIR. Thus, if participants are provided with a correct explanation, they more often correctly follow the AI advice. Overall, correct explanations result in humans changing their initially incorrect decisions to align with the correct AI advice more often, and this is especially prevalent among non-experts.

Furthermore, the analysis reveals that assertiveness does not moderate the effect between correctness and RAIR. According to Hayes [32] and Warner [91], we drop the interaction term and conduct a regression analysis with assertiveness set as the independent variable. The result shows that there is no significant direct effect of assertiveness on RAIR (coeff = −.04, p-value = .79). Hence, we confirm hypotheses 1 and 3 and reject hypothesis 5 for example-based explanations.

5.2.3 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 [74].

The moderation analysis in Table 1 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 Figure 8, which tells us that experts have a higher RSR than non-experts. Therefore, we reject hypothesis 2, and additionally, hypothesis 4 as the level of expertise does not have a moderating role but has a direct effect on RSR for natural language explanations. On top of that, we reject hypothesis 6.

5.2.4 The correctness of explanations and participants' level of expertise have a direct effect on RSR for example-based explanations. Lastly, we report the results of the moderation analysis for example-based explanations on RSR. The analysis is set up the same as it is in Section 5.2.3 but for the example-based explanations.

We can see in Table 1 that the level of expertise does not significantly moderate the effect of correctness on RSR (coeff = -.73, p-value = .57). Additionally, there is no significant moderation of assertiveness (Z1 x corr.: coeff = −.45, p-value = .55; Z2 x corr.: coeff = −.43, p-value = .55). Thus, we conduct a regression analysis without the interaction terms. The results of this analysis show no direct effect of assertiveness on RSR (coeff = −.03, p-value = .86). However, there is a significant direct effect of explanations' correctness on RSR (coeff = −1.40, p-value = .00) and a direct effect of level of expertise on RSR (coeff = 3.05, p-value = 0.00). This means that experts more often correctly override the wrong AI advice and stick to their correct initial decision compared to non-experts for example-based explanations. Additionally, when the explanations are correct, participants more often correctly override wrong AI advice and stick to their correct initial decision. Hence, experts have a higher RSR than non-experts, which can also be seen in Figure 8. Moreover, as incorrect explanations have a higher impact on RSR, experts are able to identify false AI advice for incorrect explanations better. This also shows that experts are able to identify incorrect AI advice to a greater extent than non-experts; experts, in this case, rely more heavily on their own judgment.

Thus, we confirm hypothesis 2 and reject hypothesis 4 as the level of expertise does not take in a moderating role but has a direct effect on RSR for example-based explanations. On top of that, we reject hypothesis 6.

Overall, the moderation analyses reveal that the level of expertise moderates the effect of the correctness of explanations on RAIR for both explanation modalities. Additionally, the analyses show that the level of expertise has a direct effect on RSR for both explanation modalities, and the correctness of explanations has a direct effect on RSR for example-based explanations.

5.3 Human-AI Team Performance

Hemmer et al. argue that interpretability is a key component of human-AI complementarity [33]. Several previous user studies have failed to show that incorporating XAI into AI systems can lead to CTP [27]. However, with a new dimension of XAI advice in Figure 1, we can contribute to the current literature by investigating how the correctness of explanations affects CTP. By calculating the participants' performance before and after seeing the AI and XAI advice, we can determine whether CTP exists in the presence of imperfect XAI. As the analyses in Section 5.2 reveal, the level of expertise impacts appropriate reliance in terms of RSR and RAIR. Thus, in comparing the human-AI team performance, we distinguish by participants' level of expertise. We use accuracy as the performance metric. Figure 7 presents the performance of AI and humans for each treatment.

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. In Figure 7, we see that when experts are paired with the AI, their performance improves by 8.74% for the natural language modality and 9.53% for the example-based modality. When experts are paired

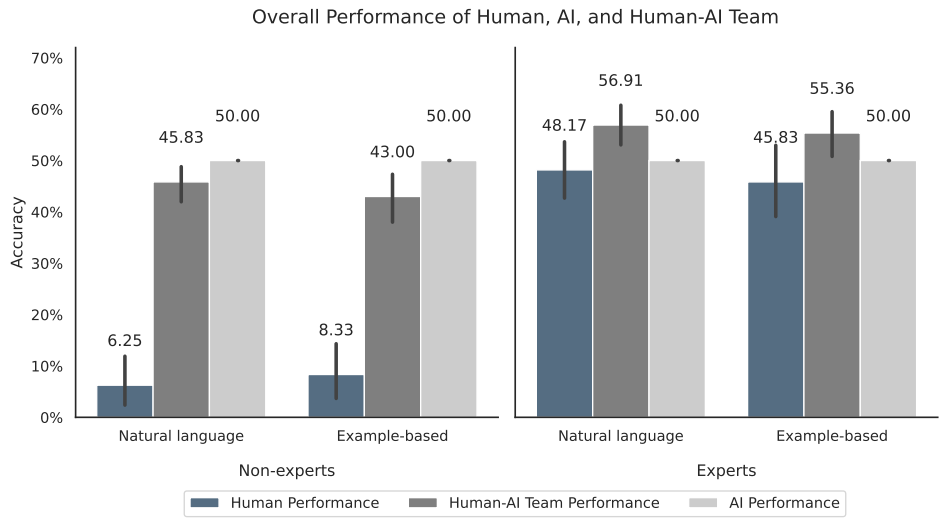

Fig. 7. The average overall performance of the human, AI, and human-AI teams for identifying 12 birds. The bar chart on the left shows the performance of the non-experts, while the bar chart on the right shows the performance of the experts.

with AI, they perform 6.91% better than the AI alone for natural language explanations and 5.36% for example-based explanations.

While experts reach CTP, we do not see this for non-experts. However, we do see that the non-experts greatly improve their performance and nearly match the AI's performance when paired with the AI. Specifically, non-expert participants who see the natural language explanations improve their performance by 39.58% (task accuracy of 45.83%), while non-expert participants who see the example-based explanations improve their performance by 34.67% (task accuracy of 43.00%) when paired with the AI.

We can separate Figure 7 into correct and incorrect explanations. When we only consider cases with correct explanations (Figure 11 in Appendix A.4), the non-experts' task accuracy is approximately the same as the AI alone: 48.81% for natural language explanations and 49.33% for example-based explanations. Experts reach CTP in both modalities. When only considering incorrect explanations (Figure 12 in Appendix A.4), 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% and 36.67% for example-based explanations.

Additionally, we calculate two-sample t-tests to assess whether the trends in Figure 7 are significant. The team performance of experts and AI is significantly higher than the team performance of non-experts and AI in both explanation modalities (natural language: p-value = 0.00, example-based: p-value = 0.00). Furthermore, we also compare the performance for correct and incorrect explanations. Here, we see the same results: experts achieve a significantly higher team performance than non-experts (correct explanations — natural language: p-value = 0.00, example-based: p-value = 0.01; incorrect explanations — natural language: p-value = 0.00, example-based: p-value = 0.00).

5.4 Deception caused by Imperfect XAI

In Figure 8, we compare RAIR to RSR for both levels of expertise and the correctness of explanations. We show this comparison for example-based explanations (the graph on the left side of Figure 8) and natural language explanations (the graph on the right side of Figure 8). By measuring RAIR and RSR for incorrect and correct explanations separately, we can calculate the deception caused by imperfect XAI (refer to Equation (3) on p. 16). We do not visualize assertiveness since we do not see any significant direct or moderation effects.

The figure shows that experts have a higher RSR than non-experts for both incorrect and correct explanations across both explanation modalities, validating that experts rely more on their own initial decisions when AI advice is given. The most striking result that emerges from the data is that for example-based explanations, we observe that experts have a significantly higher RSR for incorrect explanations (RSR = 0.57) than correct explanations (RSR = 0.29), resulting in a negative DoR_{RSR} of −0.28 (p-value = 0.00). As a result, experts are falsely relying on the AI advice when provided with correct example-based explanations[7]. This means that experts are prone to being misled by correct explanations when the AI advice is incorrect. However, we do not see this trend for natural language explanations. Here, there is a positive DoR_{RSR} of 0.09, which is not significant (p-value = 0.41). For both modalities, the DoR_{RAIR} is positive, meaning that experts rightly follow correct AI advice more often when provided with correct explanations than with incorrect explanations. The data shows a weak, significant positive deception of reliance for example-based explanations (DoR_{RAIR} = 0.16, p-value = 0.10). Similarly to the RSR cases, for

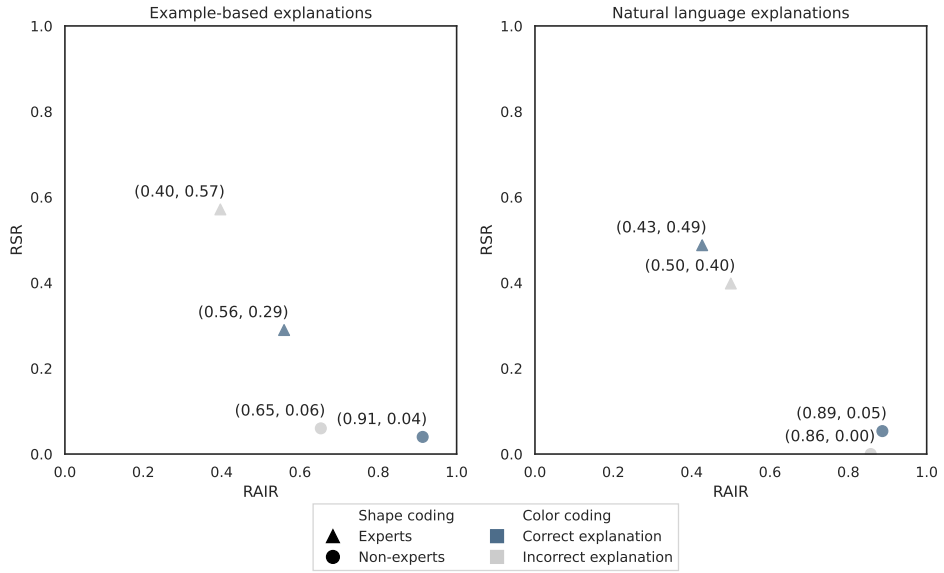

Fig. 8. Average observed RAIR (correct AI advice) and RSR (incorrect AI advice) for example-based explanations (on the left side) and natural language explanations (on the right side). We show the average RAIR and RSR for both levels of expertise as well as correct and incorrect explanations.

[7]Note that correct example-based explanations are consistent in showing three images of the predicted class. Incorrect example-based explanations represent three images that do not correspond to the predicted class of the AI. Moreover, the examples shown are not consistent with the bird species displayed in 90% of the *correct advice, incorrect explanation* cases and in 40% of the *incorrect advice, incorrect explanation* cases in our study.

the RAIR cases, the experts are provided with three consistent examples for correct explanations that represent the AI's correctly predicted bird species. The incorrectly provided explanations represent three images that can be inconsistent in the bird species. Thus, experts are deceived by such incorrect explanations even though the AI advice is correct.

Non-experts have, in both modalities, a similar DoR_{RSR} indicating no significant difference in their RSR between correct and incorrect explanations. However, non-experts follow the correct AI advice for correct example-based explanations more often than for incorrect example-based explanations (significant with p-value = 0.01). For the latter, the three examples can show inconsistent bird specie(s) that are different from the ground truth of the shown image. Thus, the DoR_{RAIR} for non-experts is at 0.26. Interestingly, for natural language explanations, the incorrect explanations are not misleading as much (DoR_{RAIR} = 0.03, not significant, with a p-value = 0.68). This means that non-experts are not misled by incorrect explanations in natural language as much as by visual, example-based explanations. In general, non-experts have a higher RAIR than experts.

Overall, participants have a higher $DoR(RAIR, RSR)$ for example-based explanations (experts: $DoR(RAIR, RSR)$ = 0.32; non-experts: $DoR(RAIR, RSR)$ = 0.26) than for natural language explanations(experts: $DoR(RAIR, RSR)$ = 0.11; non-experts: $DoR(RAIR, RSR)$ = 0.06). This means that especially the correctness of example-based explanations has an impact on humans' decision-making behavior.

5.5 Designing for Imperfect XAI

At the end of the bird identification task, we ask participants: "*Under what circumstances would you prefer assertive (e.g., "definitely", "clearly") versus non-assertive (e.g., "might be", "appears to be") versus neutral explanations and why?*". With imperfect XAI existing in human-AI collaborations, it is necessary not only to understand quantitatively how it impacts decision-makers but also qualitatively. Even though we do not observe the level of assertiveness to have a direct effect or a moderation effect on appropriate reliance, it is still valuable to analyze participants' preferences when it comes to the tone of explanations.

Through inductive content analysis of participants' responses to this question, we derive two dimensions that researchers and designers should consider when developing and evaluating imperfect XAI in human-AI collaborations: *AI Behaviors* and the *Impact on Human-AI Teams*. Each dimension is made up of four themes that are derived from concepts that emerge in the responses, shown in Figure 9. We highlight those themes in bold. 25% (34 participants) of the responses either do not provide reasoning for their opinion or do not answer the question such that it could be grouped into one of the eight themes we identify. We provide quotes from participants for each theme to structure the aggregated dimensions and shape our insights on designing for imperfect XAI.

5.5.1 Aggregated Dimension: AI Behaviors. 54 out of the 136 participants answer the survey question with comments relating to the first aggregated dimension: AI Behavior. Participants rationalize that the AI's behavior determines when they prefer assertive, non-assertive, and neutral explanations. Within the AI Behavior dimension, four themes emerge from the participants' comments, such as the model's overall performance, whether the model's prediction is correct or not, the model's confidence in individual predictions, and the correctness/quality of the explanation for a given prediction.

While only 6 out of the 136 participants make comments about the **model's performance**, it still provides interesting insight that should be considered. Instead of looking at the individual prediction level, these participants focus on the global performance of the model. One participant states that if developers find their model "*... to be 90% accurate in your testing, use more definite*

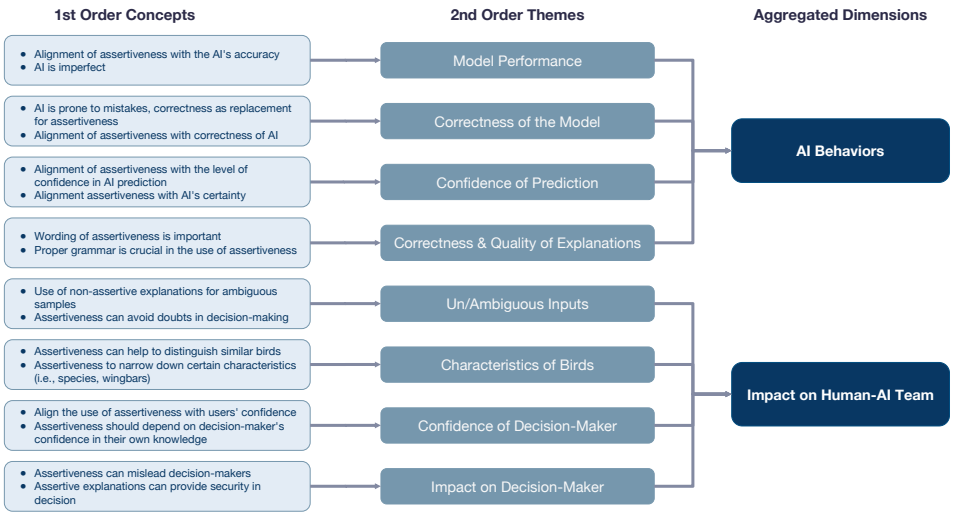

Fig. 9. Findings from the qualitative analysis of the survey question: "Under what circumstances would you prefer assertive (*e.g.*, "definitely", "clearly") versus non-assertive (*e.g.*, "might be", "appears to be") versus neutral explanations and why?". Factors that should be taken into consideration when designing explanations for imperfect XAI systems.

language, but if it's not there yet, consider making the tone more neutral and put more responsibility with the end user to interpret the field marks ...".

Looking at the individual prediction level, 8 out of 136 participants comment on the **correctness of the AI's prediction**. For example, one participant says that "*... a non-assertive response would be preferred since the AI selections were incorrect ...*" while another participant says, "*I would prefer the assertive language to be accompanied by correct identifications*".

Considering individual predictions on a more granular level, many participants (31 out of 136) make comments related to **the confidence of the AI's prediction**. These participants comment on how this factor could be used to determine the tone of the explanation. Specifically, participants, "*... would prefer the level of assertiveness to depend on the level of confidence of the answer given by the AI*". One participant expands upon that sentiment by specifying when non-assertive versus assertive tones should be used: "*I would prefer assertive sentences when the probability of the AI model is very high, while I would prefer non-assertive when the probability is very close to other classes of the model*".

9 out of 136 participants make comments related to the **correctness and quality of the AI explanations**. One participant who sees example-based explanations comments on how some of the examples are incorrect and do not show the correct species. They use this specific situation to rationalize when they would prefer the tone of explanations to be assertive versus non-assertive: "*I would prefer assertive explanations if the 'similar' photos were actually of the correct species and if the explanation confirmed this. Otherwise, non-assertive explanations are more helpful*". Another participant who makes a comment related to this theme agrees that assertive language should be used, "*when all of the reference pictures match up and there are no other similar-looking species*".

One participant who sees the natural language explanations comments on the quality and detail of the explanation being a factor to use when determining the tone of the explanation, "*If the bird description [natural language explanation] is very generic, i.e., brown wings, gray body, or yellow beak*

(traits that correspond to many birds), I'd rather the AI appear more cautious in its judgment and use non-assertive explanations. However, if the bird has some standout characteristic that the AI correctly identifies [through the natural language explanation], i.e., bright yellow body or red-tipped wings, etc., then assertive explanations seem more convincing".

5.5.2 Aggregated Dimension: Impact on Human-AI Team. 48 out of the 136 participants answer this survey question with comments related to the human-AI team, such as the confidence and knowledge of the decision-maker, impact on the decision-maker, unambiguous input data, and characteristics of the input data.

10 of 136 prefer the level of assertiveness to align with their **own confidence level** and knowledge of the domain. For example, one participant says, "*I would prefer more assertive explanations when I don't feel very sure about my choice*". Another participant adds that they would prefer assertive explanations if they "*... didn't know anything about the topic in question ...*". However, one participant says, "*I would prefer neutral explanations if I'm unsure of the species and assertive if I'm confident in my identification*".

15 out of the 136 make comments related to the **impacts on the decision-maker**. Some comments consist of concerns related to being misled and over- or under-relying on the AI, while other comments motivate the benefits of having assertive explanations. For example, one participant states that "*Assertive words create more security while changing your opinion or trying to gain knowledge*". Another participant who shares the same sentiment said, "*I would prefer assertive explanations because it would make me feel more secure about the answer*". However, some participants do not share the same sentiment about assertive explanations and pointed out the potential consequences of them: "*Assertive AI explanations were given for incorrect identifications, which would mislead users*". Given that potential to be misled, another participant rationalizes they "*... would prefer non-assertive explanations because I [they] do not fully trust AI with bird ID just yet*".

17 out of the 136 participants rationalize that the assertiveness of explanations should be based on how **ambiguous the input** is for a given prediction. For example, one participant brings up the quality of the input image and the difficulty of the bird ID as a way to determine whether explanations should be assertive or not: "*When it comes to less distinctive IDs like most sparrows, or harder to ID circumstances like winter or females or juveniles, or situations with weird lighting or harder angles it makes sense to use non-assertive.*". On a similar line of thought, another participant says, "*I would prefer assertive explanations for birds that have distinctive traits over similar bird species, and non-assertive or neutral explanations for birds that are similar with characteristics that are more difficult to tell apart*".

6 out of the 136 participants commented on how the use of assertive explanations helped them realize various **characteristics of the birds**. For example, one participant values the assertive tone because it is "*... helpful for pointing out distinctive features that would help ID the bird.*". They also think that "*... the non-assertive language was helpful for species that share similar characteristics (aka clear, non-streaked breast) with other species that share that characteristic*".

6 DISCUSSION

We investigate how imperfect XAI impacts humans' decision-making when collaborating with an AI. More precisely, we assess how imperfect explanations affect humans' reliance behavior and investigate the effects on the human-AI team performance. To answer RQ 1 and RQ 2, we assess the validity of our research model for two different types of explanations: natural language explanations and example-based explanations. Previous research emphasizes the need to consider imperfect AI when designing for human-AI collaboration [47]. With recent research looking into how humans and AI can achieve complementary team performance [7], Schemmer et al. [74]

conceptualize the role of appropriate reliance in human-AI collaboration. We extend Schemmer et al. [74]'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 CSCW.

In our study, we observe a significant moderation of humans' level of expertise on the effect of explanations' correctness on RAIR for both explanation modalities. 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 in both explanation modalities. Additionally, the correctness of explanations impacts RSR negatively for example-based explanations. 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. Especially for example-based explanations, imperfect XAI deceives experts' self-reliance and experts'/non-experts' relative AI reliance, fostering inappropriate reliance on the AI. Thus, this study sets a starting point to investigate the effect of imperfect XAI for different explanation modalities.

Our findings show that imperfect explanations impact human-AI decision-making. We observe that experts reach complementary team performance when imperfect explanations are provided. This holds true for natural language and example-based explanations. While non-experts do not reach complementary team performance, our analyses reveal that their performance can be improved to be similar to that of the AI performance. Moreover, there is a difference between reaching complementary team performance when the correctness, or fidelity, of the explanation changes (see Figure 11 and Figure 12 in Appendix A.4). 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 [33]. With our results, we confirm this hypothesis. Furthermore, Papenmeier et al. [64] 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 (Figure 12 in Appendix A.4). Similar to our findings, Nourani et al. [60] 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. For example, this can inform managers on how to assign tasks to humans with different levels of expertise and provide them with explanations in different modalities. It could also lead to organizations modifying their human-AI collaboration workflows. From informal conversations with the product team of an AI decision-support tool[8] for biologists and conservationists to classify species and identify individuals from camera trap imagery, we learned that organizations using their tool have modified their workflow to incorporate "checks-and-balances". For example, intro-level biologists will collaborate with the AI to match individuals and then request a review of their "human-AI team" decision from a higher-up. In this unique human-human-AI collaboration scenario, the expert biologist could potentially correct situations when an intro-level biologist over-relies on AI advice because of an incorrect explanation.

[8]WildMe.org

Visual, example-based explanations are more deceptive than natural language explanations. To account for the impact of imperfect XAI on humans' appropriate reliance, we establish a novel metric **DoR**, (Deception of Reliance), to measure the difference in RAIR and RSR for correct and incorrect explanations. Our results indicate that people are more deceived by example-based explanations than by natural language explanations. In terms of RAIR (note that in RAIR cases, the AI advice is correct), experts and non-experts are deceived by incorrect explanations. In terms of RSR (note that in RSR cases, the AI advice is incorrect), experts are deceived by correct explanations. This is an interesting observation that may be explained by the consistency of shown visual examples. For *correct advice, incorrect explanations* cases, the XAI is providing three visual examples that show a different bird species than the bird species on the image to be classified (see Figure 6). Moreover, these three visual examples can belong to different bird species since the XAI is choosing the top three most similar bird images (in our study, this is in 90% of all *correct advice, incorrect explanation* cases). This inconsistency in example-based explanations might deceive experts and non-experts to not rely on the AI anymore when they identify visual differences in the images provided as explanations, disregarding the correct AI advice. We discover the same behavior for experts for *incorrect advice, correct explanation* cases. In those cases, the explanations consist of three images of the same bird species as the AI predicted. The incorrect explanations consist of three images that can be inconsistent in the bird species shown (in our study, this in 40% of all *incorrect advice, incorrect explanation* cases). Thus, this inconsistency in examples might deceive experts into no longer relying on themselves anymore when they identify three consistent examples shown, disregarding the incorrect AI advice. Hence, the DoR of experts and non-experts is positive for RAIR cases as they are deceived by incorrect explanations, while experts additionally have a negative DoR and are deceived by correct explanations. Note that the overall RSR for experts is still higher than non-experts' RSR; the impact on deception caused by imperfect XAI is higher. As participants mentioned in the survey, it is more convincing that there is less uncertainty in the AI advice when three images that are similar to each other are shown than when three different images are shown. This corroborates our findings. This trend is not present in natural language explanations.

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 [18]. 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 [18, 45, 46], our findings do not show an impact on appropriate reliance. Quantitatively, we do not find a direct effect of assertiveness on appropriate reliance. Qualitatively, we observe that the tone of the explanation should depend on certain situations, such as the AI's behaviors, instead of the human's level of domain expertise. We observe that 31% of the participants prefer assertiveness to align with the confidence of the model's prediction, while only 10% of participants prefer assertiveness to align with their confidence in the domain. This finding could be attributed to the fact that participants are collaborating with imperfect AI and are able to acknowledge that the AI advice was occasionally incorrect. However, given these qualitative suggestions and the potential for natural language explanations to present irrelevant or incorrect information [77], we encourage future work to explore various ways to alter the tone of an explanation based on the themes we identified in Figure 9.

Our findings can guide researchers and practitioners on how to assess and design for imperfect XAI in human-AI collaborations. Regardless of the explanation modality, 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.*, [44, 56]), 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 [79]. Especially in the workplace, AI can facilitate knowledge transfer and support organizations in retaining and distributing expert knowledge [40, 78, 92]. Similarly, it is also crucial to understand how imperfect XAI affects the learning of novices through AI-based learning systems [79] or through collaboration with AI [72]. 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.

7 LIMITATIONS & FUTURE WORK

We elaborate on various limitations of our study, how they could impact the interpretation of our results, and identify opportunities for future work.

Lack of Information to Properly Identify Birds. Expert birders usually rely on more information than just the visual characteristics of a bird when determining the bird species, especially for ambiguous cases. For example, the location and habitat in which the bird was spotted can be imperative to determine the exact bird species within a family. It's unclear to what extent the lack of this information influences our results. Future technical work could consider using this information to help build more transparent bird classification models.

Correctness of Explanations versus Explanation Fidelity. Throughout our study, we consider explanations to either be incorrect or correct. However, as we mention, some explanations that we classify as incorrect can contain evidence that is correct, making it difficult to only have a binary categorization for the correctness of explanations. While we use a binary scale for our analyses, explanation correctness, or fidelity, can be quantitatively measured on a continuous scale and categorized as low fidelity and high fidelity [64]. We encourage future work to explore explanation fidelity using multiple categories instead of two to understand the differences between low- and medium-fidelity explanations when it comes to task performance and appropriate reliance. This will provide insight into the impact that noisy explanations have on decision-making, such as when an explanation reveals some information that is aligned with the ground truth class and some information that is aligned with the predicted class.

Simplistic Natural Language Explanations. The natural language explanations are very limited and simplistic, although this is a fault of the model that we are using to generate those natural language explanations [34]. Compared to a field guide such as All About Birds [4], these natural language explanations could be correct for several species given their lack of detail and their short length. This is potentially a side effect of the text descriptions in the CUB-200-2011

dataset being sourced from crowd workers instead of experts. We encourage technical researchers who are developing explanation methods that explain image classifications using natural language to pay careful attention to the text descriptions used for training the language model. For example, researchers could consider using text data from a field guide as their training data. We also encourage CSCW researchers to conduct more studies on how the explanation's length and level of detail impact experts' and non-experts' appropriate reliance.

Different Explanations Convey Different Information. While we do not intend to directly compare the two explanation modalities throughout our analyses, they are discussed in terms of similarities and differences throughout the article. Our main intention is to investigate how our research model holds across different modalities of explanations. While several previous studies compare multiple different types of explanation modalities qualitatively and quantitatively (*e.g.*, [20, 24, 43, 44, 80]), we encourage readers to avoid directly comparing the two explanations because they present different information. Previous research reveals that different explanation techniques can result in disagreements for the same dataset [70]. For example, the natural language explanations from Hendricks et al. [34] are feature-based, providing descriptions of features present in the image [34]. However, the example-based explanations present three similar images, which is very different information from the natural language description of features. On top of that, the incorrectness in both explanation modalities is represented in different ways. While there are factual errors in natural language explanations that previous research addresses (e.g., hallucination effects of natural language models [77]), there are logical errors (e.g., inconsistencies) within example-based explanations. This opens avenues for future research to investigate how different human cognitive abilities (i.e., cognitive styles) impact the perception of these imperfect explanations in AI-assisted decision-making scenarios.

Visualizing Assertiveness for Example-Based Explanations. While there is previous research on how to visualize the confidence of a prediction for image classification [82], to the best of our knowledge, there is no work visualizing the assertiveness of example-based explanations. Given this, our example-based explanations use natural language to convey assertiveness. Visualizing the assertiveness for each example in an example-based explanation would provide the decision-maker with another queue about whether they should rely on the model.

AI Expertise versus Domain Expertise. Our analyses are based on the participant's expertise in bird species identification. We do not ask participants about their knowledge of AI. Thus, we do not analyze their expertise related to AI. It is unclear to what extent the participant's expertise with AI would impact the effect of imperfect XAI on appropriate reliance and task accuracy for bird species classification. Future work should consider looking at the effect that AI knowledge combined with domain expertise has on appropriate reliance, taking into account an imperfect XAI.

8 CONCLUSION

This article sets out a research model to investigate the effect of imperfect XAI on human-AI decision-making. Thus far, human-computer interaction and CSCW literature fail to thoroughly scrutinize how explanations' correctness affects humans' decision-making and their reliance behavior on AI. Hence, through a human study with 136 participants, 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. Furthermore, we explore to what extent incorrect explanations deceive decision-makers' reliance on AI. With our findings, we make several contributions: First, we propose a research model to investigate the moderation of assertiveness and humans' level of expertise on imperfect XAI in decision-making tasks. We thereby extend the existing conceptualization of appropriate reliance by a new dimension of XAI

advice. Second, through an empirical study, we reveal that imperfect explanations and participants' level of expertise affect human-AI decision-making for two different explanation modalities. In addition, we show the effect on complementary team performance and provide guidance for future studies on how to investigate imperfect XAI in the context of human-AI decision-making. Third, we propose a novel metric called Deception of Reliance (DoR), which allows us to measure the impact of incorrect explanations on decision-makers' reliance. Our results inform designers of human-AI collaboration systems and provide guidelines for their development. Fourth, we reveal which role the language tone in explanations plays and outline important dimensions that should be considered when designing for XAI advice.

Overall, with this work, we reveal the impact of imperfect XAI on human-AI decision-making by taking into account humans' level of expertise and explanations' assertiveness. Extensive and rigorous research is needed to fully understand and exploit imperfect XAI in decision-making. We invite researchers to take part in this debate and hope to inspire scientists to actively participate in this endeavor.

ACKNOWLEDGMENTS

We would like to thank Youwei Jiang for helping with the development of portions of the user interface used for the study. We would like to thank Hao-Fei Cheng and Haiyi Zhu for providing feedback on the initial study design. We also thank Max Schemmer and other lab members of the KSRI for discussing our study design and findings with us. We would like to thank Nari Johnson, Hayden Stec, Adel Sharif, Donny Bertucci, Alex Cabrera, Will Epperson, Venkat Sivaraman, and other members of the DIG Lab at CMU for participating in our initial pilot studies. We would like to thank all of the experienced birders who provided us with feedback on our study design and discussed the limitations of our study with us. Lastly, we would like to thank all of our connections that helped us recruit birders. Research reported in this publication was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL164906. ChatGPT was utilized to generate LaTeX code for some of the tables in this paper.

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] 2023. https://www.allaboutbirds.org/guide/. Accessed: July 2, 2023.
[5] 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
[6] Stephan Alaniz. 2018. pytorch-gve-lrcn: PyTorch implementation of Visual Generation and Execution for Long-term Predictions. https://github.com/salaniz/pytorch-gve-lrcn.
[7] 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
[8] 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
[9] Woodrow Barfield. 1986. Expert-novice differences for software: Implications for problem-solving and knowledge acquisition. *Behaviour & Information Technology* 5, 1 (1986), 15–29. https://doi.org/10.1080/01449298608914495
[10] Shraddha Barke, Michael B James, and Nadia Polikarpova. 2023. Grounded copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages* 7, OOPSLA1 (2023), 85–111. https://doi.org/10.1145/3586030

[11] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. 2021. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems* (2021), 1–29. https://doi.org/10.1080/12460125.2021.1958505

[12] 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

[13] 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

[14] 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

[15] Ángel Alexander Cabrera, Adam Perer, and Jason I Hong. 2023. Improving human-AI collaboration with descriptions of AI behavior. *arXiv preprint arXiv:2301.06937* (2023). https://doi.org/10.1145/3579612

[16] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262. https://doi.org/10.1145/3301275.3302289

[17] 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

[18] 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

[19] Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to tango: Towards theory of AI's mind. *arXiv preprint arXiv:1704.00717* (2017). https://arxiv.org/abs/1704.00717

[20] 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

[21] Arghavan Moradi Dakhel, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C Desmarais, and Zhen Ming Jack Jiang. 2023. Github copilot ai pair programmer: Asset or liability? *Journal of Systems and Software* 203 (2023), 111734. https://doi.org/10.1016/j.jss.2023.111734

[22] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. 2022. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly* 39, 2 (2022), 101666. https://doi.org/10.1016/j.giq.2021.101666

[23] Murat Dikmen and Catherine Burns. 2022. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies* 162 (2022), 102792. https://doi.org/10.1016/j.ijhcs.2022.102792

[24] 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 (2022), 10323. https://doi.org/10.3390/app122010323

[25] 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

[26] Jure Erjavec, Nadia Zaheer Khan, and Peter Trkman. 2016. The impact of personality traits and domain knowledge on decision making–a behavioral experiment. (2016). https://core.ac.uk/download/pdf/301369906.pdf

[27] Raymond Fok and Daniel S Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv preprint arXiv:2305.07722* (2023). https://arxiv.org/abs/2305.07722

[28] Courtney Ford and Mark T Keane. 2022. Explaining Classifications to Non Experts: An XAI User Study of Post Hoc Explanations for a Classifier When People Lack Expertise. *arXiv preprint arXiv:2212.09342* (2022). https://doi.org/10.1007/978-3-031-37731-0_15

[29] Dennis A Gioia, Kevin G Corley, and Aimee L Hamilton. 2013. Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational research methods* 16, 1 (2013), 15–31. https://doi.org/10.1177/1094428112452151

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] Judith A Holton. 2007. The coding process and its challenges. *The Sage handbook of grounded theory* 3 (2007), 265–289. https://doi.org/10.4135/9781848607941.n13

[36] 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

[37] Huiling Huang, Stephanie Q Liu, and Zhi Lu. 2022. When and why language assertiveness affects online review persuasion. *Journal of Hospitality & Tourism Research* (2022), 10963480221074280. https://doi.org/10.1177/10963480221074280

[38] Christina Humer, Andreas Hinterreiter, Benedikt Leichtmann, Martina Mara, and Marc Streit. 2022. Comparing Effects of Attribution-based, Example-based, and Feature-based Explanation Methods on AI-Assisted Decision-Making. (2022). https://osf.io/h6dwz/download

[39] 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

[40] 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

[41] 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

[42] 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

[43] Doha Kim, Yeosol Song, Songyie Kim, Sewang Lee, Yanqin Wu, Jungwoo Shin, and Daeho Lee. 2023. How should the results of artificial intelligence be explained to users?-Research on consumer preferences in user-centered explainable artificial intelligence. *Technological Forecasting and Social Change* 188 (2023), 122343. https://doi.org/10.1016/j.techfore.2023.122343

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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).

[49] Ann Kronrod, Amir Grinstein, and Kerem Shuval. 2022. Think positive! Emotional response to assertiveness in positive and negative language promoting preventive health behaviors. *Psychology & health* 37, 11 (2022), 1309–1326. https://doi.org/10.1080/08870446.2021.1942876

[50] Yeonjoo Lee, Miyeon Ha, Sujeong Kwon, Yealin Shim, and Jinwoo Kim. 2019. Egoistic and altruistic motivation: How to induce users' willingness to help for imperfect AI. *Computers in Human Behavior* 101 (2019), 180–196. https://doi.org/10.1016/j.chb.2019.06.009

[51] Benedikt Leichtmann, Andreas Hinterreiter, Christina Humer, Marc Streit, and Martina Mara. 2023. Explainable Artificial Intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. *International Journal of Human–Computer Interaction* (2023), 1–18. https://doi.org/10.31219/osf.io/68emr

[52] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13. https://dl.acm.org/doi/abs/10.1145/3411764.3445522

[53] 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

[54] 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

[55] 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

[56] 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

[57] 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

[58] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems* 34 (2021), 26422–26436. https://proceedings.neurips.cc/paper/2021/hash/de043a5e421240eb846da8effe472ff1-Abstract.html

[59] Giang Nguyen, Mohammad Reza Taesiri, and Anh Nguyen. 2022. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. *Neural Information Processing Systems (NeurIPS)* (2022). https://arxiv.org/abs/2208.00780

[60] 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

[61] Jeroen Ooge and Katrien Verbert. 2021. Trust in Prediction Models: a Mixed-Methods Pilot Study on the Impact of Domain Expertise. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE, 8–13. https://doi.org/10.1109/trex53765.2021.00007

[62] 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

[63] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

[64] 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

[65] 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

[66] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

[67] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. 2021. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112* (2021). https://arxiv.org/abs/2106.02112

[68] Lara Riefle, Patrick Hemmer, Carina Benz, Michael Vössing, and Jannik Pries. 2022. On the Influence of Cognitive Styles on Users' Understanding of Explanations. In *Proceedings of the Forty-Third International Conference on Information Systems (ICIS)*. https://arxiv.org/abs/2210.02123

[69] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 223–233. https://doi.org/10.1145/3503252.3531311

[70] Saumendu Roy, Gabriel Laberge, Banani Roy, Foutse Khomh, Amin Nikanjam, and Saikat Mondal. 2022. Why Don't XAI Techniques Agree? Characterizing the Disagreements Between Post-hoc Explanations of Defect Predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 444–448. https://doi.org/10.1109/icsme55016.2022.00056

[71] 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

[72] 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

[73] 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

[74] 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

[75] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2024. On explanations, fairness, and appropriate reliance in human-AI decision-making. *ACM Conference on Human Factors in Computing Systems (CHI)* (2024). https://arxiv.org/abs/2209.11812

[76] Jakob Schoeffer, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. 2023. On the Interdependence of Reliance Behavior and Accuracy in AI-Assisted Decision-Making. In *HHAI 2023: Augmenting Human Intellect*. IOS Press, 46–59. https://doi.org/10.3233/faia230074

[77] 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

[78] 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

[79] 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

[80] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119. https://doi.org/10.1145/3397481.3450662

[81] 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

[82] Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. 2023. How Displaying AI Confidence Affects Reliance and Hybrid Human-AI Performance. In *HHAI 2023: Augmenting Human Intellect*. IOS Press, 234–242. https://doi.org/10.3233/faia230087

[83] 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

[84] 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

[85] 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

[86] 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

[87] 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/

[88] 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

[89] 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

[90] 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

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

[92] 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

[93] 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

[94] Joost F Wolfswinkel, Elfi Furtmueller, and Celeste PM Wilderom. 2013. Using grounded theory as a method for rigorously reviewing literature. *European journal of information systems* 22, 1 (2013), 45–55. https://doi.org/10.1057/ejis.2011.51

[95] 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

[96] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201. https://doi.org/10.1145/3377325.3377480

[97] 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

[98] 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. https://doi.org/10.1109/icdcs.2019.00123

[99] Zibin Zhao and Cagatay Turkay. 2023. Exploring how expertise impacts acceptability of AI explanations: a case study from manufacturing. In *Proceedings of the ACM CHI Workshop Human-Centered Perspectives in Explainable AI*. ACM. https://wrap.warwick.ac.uk/175429/

A APPENDIX

A.1 Appropriate Reliance with Imperfect XAI

Table 2. Overview of the newly introduced metrics when considering imperfect explanations.

CSR_{IC}	Correct self-reliance for the case where the AI gives incorrect advice and a correct explanation is one when the initial human decision is correct and the final decision is correct.
OR_{IC}	Over-reliance for the case where the AI gives incorrect advice and a correct explanation is one when the initial human decision is correct and the final decision is correct.
CSR_{II}	Correct self-reliance for the case where the AI gives incorrect advice and an incorrect explanation is one when the initial human decision is correct and the final decision is correct.
OR_{II}	Over-reliance for the case where the AI gives incorrect advice and an incorrect explanation is one when the initial human decision is correct and the final decision is correct.
$CAIR_{CC}$	Correct AI reliance for the case where the AI gives correct advice and a correct explanation is one when the initial human decision is incorrect and the final decision is correct.
UR_{CC}	Under-reliance for the case where the AI gives correct advice and a correct explanation is one when the initial human decision is incorrect and the final decision is correct.
$CAIR_{CI}$	Correct AI reliance for the case where the AI gives correct advice and an incorrect explanation is one when the initial human decision is incorrect and the final decision is correct.
UR_{CI}	Under-reliance for the case where the AI gives correct advice and an incorrect explanation is one when the initial human decision is incorrect and the final decision is correct.

[1] A correct AI explanation correpsonds with the AI's advice, no matter if the advice is correct or incorrect. For a classification task this means the following: If the AI gives incorrect advice and the explanation is correct, the explanation aligns with the incorrectly predicted class.

Following the newly introduced dimension, the calculation for RAIR and RSR are adjusted to the following:

$$RSR\ (Relative\ Self-Reliance) = \frac{\sum_{i=0}^{N}\left(CSR_{IC,i} + CSR_{II,i}\right)}{\sum_{i=0}^{N} IA_i} \tag{5}$$

$$RAIR\ (Relative\ AI\ Reliance) = \frac{\sum_{i=0}^{N}\left(CAIR_{CC,i} + CAIR_{CI,i}\right)}{\sum_{i=0}^{N} CA_i} \tag{6}$$

A.2 Bird Identification Test

The bird identification test consists of images of six images: three "easy" common bird species and three "hard" bird species. The three "easy" common bird species were selected with the intention that most beginning birders would be familiar with them. For the "easy" common bird species, participants have to identify a *Downy Woodpecker*, a *Herring Gull*, and a *Ruby-Throated Hummingbird*.

For the "hard" bird species, participants have to identify a *female Hooded Warbler*, a *Blue-headed Vireo*, and a *Chestnut-sided Warbler*. The *female hooded warbler* is chosen because it looks significantly different than a male Hooded Warbler and requires a higher level of expertise to be able to correctly identify that. The *Blue-headed Vireo* is chosen because it visually looks very similar to the *Philadelphia Vireo*, again requiring a higher level of expertise to correctly identify that. Lastly, the *Chestnut-sided Warbler* is chosen because there are several different species in the Warbler family, and they are easy for non-experts to mix up.

Below are figures showing the performance on the bird test based on the experts and non-experts grouping we do.

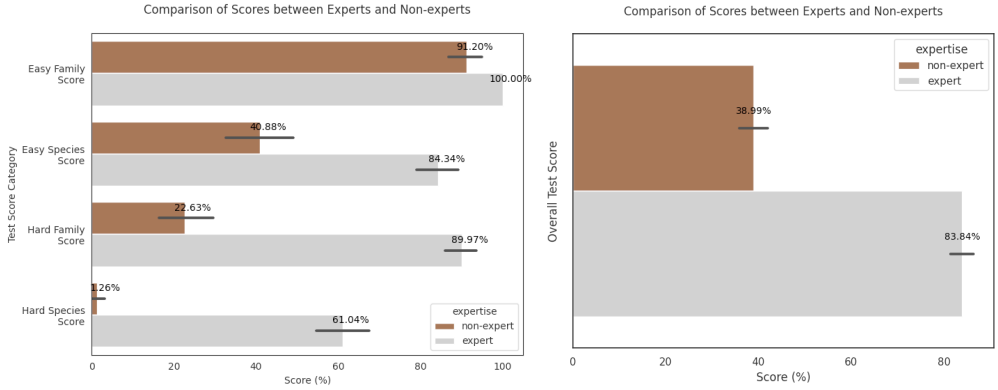

Fig. 10. The left half of this figure shows how participants (experts and non-experts) perform on average for the easy birds and three hard birds. We calculate their family accuracy as well as species accuracy. The right half of this figure shows the average overall score by combining the four scores from the left.

A.3 Moderation Analyses

Table 3. Moderation analysis of correctness of natural language explanations on RAIR with the level of expertise and assertiveness as moderators (Since the level of expertise is a three-level categorical moderator, it is split up into Z1 — non-assertive explanations in relation to all other values — and Z2 — assertive explanations in relation to all other values.

	coeff	ce	Z	p	LLCI	UCLI
const	1.26	.34	3.67	.00	.59	1.94
corr	.57	.53	1.07	.29	-.48	1.62
exp	-2.12	.33	-6.50	.00	-2.77	-1.48
Z1	-.46	.39	-1.16	.24	-1.23	.31
Z2	.00	.39	.00	1.00	-.76	.76
exp x corr	-1.00	.51	-1.96	.05	-1.99	.00
Z1 x corr	.46	.59	.77	.44	-.70	1.61
Z2 x corr	.19	.58	.33	.74	-.95	1.34

[1] *corr — correctness*; *exp — level of expertise*

Table 4. Moderation analysis of correctness of example-based explanations on RAIR with level of expertise and assertiveness as moderators (Since level of expertise is a three-level categorical moderator, it is split up into Z1 — non-assertive explanations in relation to all other values — and Z2 — assertive explanations in relation to all other values.

	coeff	ce	Z	p	LLCI	UCLI
const	.43	.32	1.35	.17	-.19	1.05
corr	1.02	.49	2.08	.04	.06	1.99
exp	-1.25	.31	-4.09	.00	-1.85	-.65
Z1	.26	.36	.73	.47	-.45	.98
Z2	.00	.37	.00	1.00	-.72	.72
exp x corr	-1.04	.47	-2.21	.03	-1.95	-.12
Z1 x corr	-.03	.54	-.06	.95	-1.08	1.02
Z2 x corr	-.16	.54	-.29	.77	-1.22	.90

[1] *corr — correctness*; *exp — level of expertise*

Table 5. Moderation analysis of correctness of natural language explanations on RSR with level of expertise and assertiveness as moderators (Since level of expertise is a three-level categorical moderator, it is split up into Z1 — non-assertive explanations in relation to all other values — and Z2 — assertive explanations in relation to all other values.

	coeff	ce	Z	p	LLCI	UCLI
const	-17.16	592.16	-.03	.98	-1177.77	1143.45
corr	13.25	592.16	.02	.98	-1147.37	1173.86
exp	15.89	592.16	.03	.98	-1144.72	1176.50
Z1	.14	.52	-.26	.79	-.89	1.16
Z2	-.31	.56	-.56	.58	-1.41	.79
exp x corr	-13.33	592.16	-.02	.98	-1173.94	1147.29
Z1 x corr	-.28	.75	-.37	.71	-1.75	1.19
Z2 x corr	.90	.74	1.20	.23	-.56	2.36

[1] *corr — correctness*; *exp — level of expertise*

Table 6. Moderation analysis of correctness of example-based explanations on RSR with the level of expertise and assertiveness as moderators (Since level of expertise is a three-level categorical moderator, it is split up into Z1 — non-assertive explanations in relation to all other values — and Z2 — assertive explanations in relation to all other values.

	coeff	ce	Z	p	LLCI	UCLI
const	-3.60	.76	-4.74	.00	-5.09	-2.11
corr	-.44	1.29	-.34	.74	-2.97	2.10
exp	3.25	.74	4.39	.00	1.80	4.70
Z1	-.09	.43	-.22	.83	-.94	.75
Z2	.09	.43	.21	.83	-.75	.93
exp x corr	-.73	1.28	-.57	.57	-3.23	1.77
Z1 x corr	-.45	.75	-.60	.55	-1.92	1.02
Z2 x corr	-.43	.72	-.59	.55	-1.85	.99

[1] *corr — correctness*; *exp — level of expertise*

A.4 Human-AI Team Performance

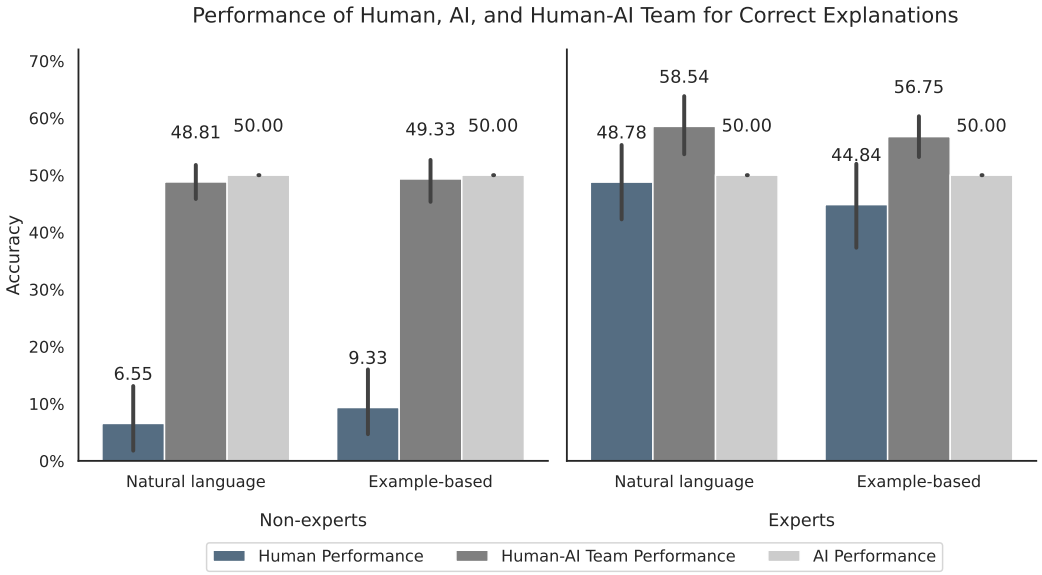

Fig. 11. Performance of the human, AI, and human-AI team specifically for correct explanations. This represents 6 birds from the 12 that participants saw.

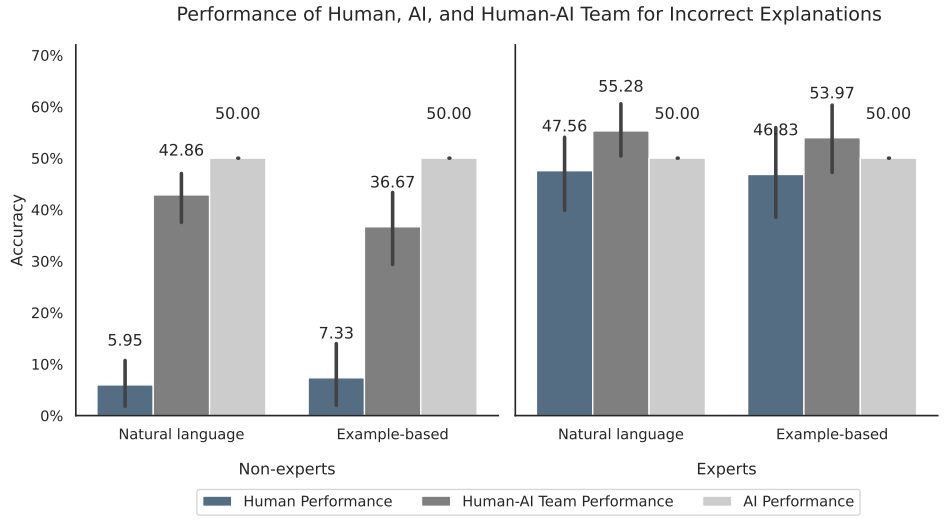

Fig. 12. Performance of the human, AI, and human-AI team specifically for incorrect explanations. This represents 6 birds from the 12 that participants saw.

Received July 2023; revised October 2023; accepted November 2023