Imperfections of XAI: Phenomena Influencing AI-Assisted Decision-Making

PHILIPP SPITZER, Karlsruhe Service Research Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany
KATELYN MORRISON, Human-Computer Interaction Institute, Carnegie Mellon University School of Computer Science, Pittsburgh, Pennsylvania, USA
VIOLET TURRI, MICHELLE FENG, and ADAM PERER, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
NIKLAS KÜHL, University of Bayreuth, Bayreuth, Germany

With the increasing use of AI, recent research in human–computer interaction explores Explainable AI (XAI) to make AI advice more interpretable. While research addresses the effects of incorrect AI advice on AI-assisted decision-making, the impact of incorrect explanations is neglected so far. Additionally, recent work shows that not only different explanation modalities impact decision-makers, but also human factors play a critical role. To analyze relevant phenomena influencing AI-assisted decision-making, this work explores the impacting factors by conceptualizing theories of appropriate reliance and taking the first steps toward empirical evidence. We show that humans' reliance on AI and the human–AI team performance are impacted by imperfect XAI in a study with 136 participants. Additionally, we find that cognitive styles affect decision-making in different explanation modalities. Hence, we shed light on diverse factors that impact human–AI collaboration and provide guidelines for designers to tailor such human–AI collaboration systems to individuals' needs.

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, Artificial Intelligence

ACM Reference format:
Philipp Spitzer, Katelyn Morrison, Violet Turri, Michelle Feng, Adam Perer, and Niklas Kühl. 2025. Imperfections of XAI: Phenomena Influencing AI-Assisted Decision-Making. *ACM Trans. Interact. Intell. Syst.* 15, 3, Article 17 (September 2025), 40 pages.
https://doi.org/10.1145/3750052

Philipp Spitzer and Katelyn Morrison contributed equally to this research.
This work was supported by the by the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL164906.
Authors' Contact Information: Philipp Spitzer (corresponding author), Karlsruhe Service Research Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany; e-mail: philipp.spitzer@kit.edu; Katelyn Morrison, Human-Computer Interaction Institute, Carnegie Mellon University School of Computer Science, Pittsburgh, Pennsylvania, USA; e-mail: kcmorris@andrew.cmu.edu; Violet Turri, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; e-mail: vturri@andrew.cmu.edu; Michelle Feng, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; e-mail: msfeng@andrew.cmu.edu; Adam Perer, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; e-mail: adamperer@cmu.edu; Niklas Kühl, University of Bayreuth, Bayreuth, Germany; e-mail: kuehl@uni-bayreuth.de.

ACM 2160-6463/2025/9-ART17
https://doi.org/10.1145/3750052

1 Introduction

With the deployment of imperfect **Artificial Intelligence (AI)** in high-stakes decision-making scenarios, decision-makers struggle with knowing when they should and should not rely on AI advice, causing frustration and resulting in potentially harmful decisions. As a result, designing and developing human-centered explanations has become a core theme in human–AI collaboration research [27, 56, 106]. Tangentially, recent work has proposed new explanation techniques that leverage machine learning models to explain the prediction of another machine learning model [8, 13, 38, 44, 52, 84, 92, 98, 102, 107, 113]. 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" [38, 44, 98]. Aside from natural language explanations, another subset of recent work proposes advanced content-based image recognition techniques to generate example-based explanations [8, 84, 107] for visual tasks. 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 [64]. In this article, we investigate the impact that this additional layer of uncertainty has on decision-makers with different levels of expertise and cognitive styles during human–AI collaboration.

Human–AI collaboration is prevalent across high-stakes scenarios [19, 57, 100, 105]. For instance, radiologists collaborate with AI to identify abnormalities in medical imagery [100]; conservationists use AI to help monitor biodiversity [12], and humanitarian aid uses AI to help identify damaged buildings after natural disasters or armed conflicts from satellite imagery [64, 114]. While some of these human–AI collaborations require the decision-maker to have several years of experience in the given domain, such as radiology, monitoring biodiversity with the help of AI does not necessarily require domain expertise [12, 72]. 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, the AI models backing them are imperfect: they do not always provide correct predictions [50].

Experts and non-experts interacting with such an imperfect AI and the same modality of explanations in interaction scenarios (e.g., decision-making [87] or learning systems [95]), 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 be unable to judge the correctness of the AI advice appropriately as they are missing this knowledge. 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 species. In this case, collaborating with AI can result in **Complementary Team Performance (CTP)**, leveraging the unique knowledge of both humans and AI. This results in the human–AI team's task performance being better than the human or AI alone [37].

Inappropriate reliance can also occur in the presence of *imperfect* ***Explainable AI (XAI)***, a term that we coined to represent the phenomenon where an explanation can be generally correct or incorrect. Papenmeier et al. [70] use the term "explanation fidelity" while Kroeger et al. [52] use the term "faithfulness" to measure how "truthful" an explanation is. We use the term imperfect XAI to align with existing terms, such as imperfect AI, in the **Human–Computer Interaction (HCI)**

community. Specifically, we define imperfect XAI as post-hoc explanation techniques that can potentially provide explanations that are misaligned with the predicted class.[1] Imperfect XAI can exist regardless of whether the AI's advice is correct or not [64]. AI explanations may oversimplify or improperly estimate complex models 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.

Collaborating with imperfect AI is not a new concept to the HCI community [37, 50, 57]. Despite numerous user studies over the years investigating human–AI collaborations and XAI, a smaller proportion has formally acknowledged the existence of imperfect XAI in their studies [16, 29, 37, 70, 85]. 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., [22, 47]), few studies explore the impact that incorrect or "noisy" explanations have on human–AI collaboration [37, 42, 70, 103]. Recent work has used technical approaches to mitigate "noisy" or incorrect natural language explanations [42], and Kroeger et al. [52] propose metrics to algorithmically evaluate the effectiveness of the generated post-hoc natural language explanations. We build on the limited literature looking at imperfect XAI by investigating how the interaction between the correctness of explanations and human factors (i.e., level of expertise and cognitive style) impacts appropriate reliance on AI, human–AI team performance, and the extent to which the explanations deceive decision-makers.

Findings from recent empirical studies tend to provide insight into which XAI techniques and designs are prone to being adversarial to human–AI collaborations. However, conceptual works from Srinivasan and Chander [96] and Miller [61] argue that these choices should be grounded in theories from psychology and cognitive science. Consequently, multiple studies in HCI research have empirically shown that human factors play an important role in the interaction of humans and AI [79, 80, 91, 111]. Riefle et al. [81] specifically show that the cognitive style (i.e., rational versus intuitive) plays a significant part in how decision-makers understand explanations [81]. However, there are other dimensions of cognitive styles (e.g., verbal versus visual [51, 77]) that are yet to be explored in this context. Prior research indicates that humans with a verbal cognitive style prefer textual information, while humans with a visual cognitive style learn better from images [78]. These definitions are core to our study. From a human-centered perspective, it is crucial to understand how humans with different traits, such as cognitive styles, may be impaired by imperfect AI advice.

The shift in AI research towards understanding human-centered challenges underscores a profound understanding that the effectiveness of AI hinges upon its alignment with human cognitive processes and human factors [56]. This shift represents a departure from conventional models that often prioritize technological prowess over the nuanced intricacies of human cognition. However, in this human-centered discourse on AI-assisted decision-making, research has neglected the impact of potentially incorrect explanations provided to decision-makers, no matter the correctness of the AI advice. Imagine a decision-maker who has a verbal cognitive style collaborating with an AI for an image classification task with visual-based explanations. Since the decision-maker has a verbal cognitive style, meaning they process textual information better than visual information, they may struggle to process the visual explanation, let alone determine the validity of the explanation. On the other hand, decision-makers with a visual cognitive style would likely prefer visual explanations and may do better at identifying if the explanation is misleading or incorrect. Ultimately, the

[1]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 the in-between cases of explanation correctness, such as when an explanation is partially correct.

decision-maker's cognitive style and the modality of explanation can prevent the human–AI team from reaching CTP.

With the growing use of AI models to facilitate decision-makers, we argue that it is necessary to understand phenomena impacting humans' decision-making. Such phenomena include the influence of incorrect explanations on decision-makers, even when the AI advice is correct. We also argue that it is important to understand the relationship of human factors for different types of explanations (i.e., visual or textual) on decision-makers' reliance on AI. Understanding these dimensions of human–AI collaboration will provide insight to XAI and HCI researchers. With limited literature exploring these topics, we present the following research questions:

RQ1: How do decision-makers' level of expertise and cognitive style moderate the effect of imperfect XAI and different explanation modalities on appropriate reliance on AI?
RQ2: To what extent does imperfect XAI deceive decision-makers with different levels of expertise?
RQ3: How do decision-makers' level of expertise and cognitive style moderate the effect of imperfect XAI and different explanation modalities on human–AI team performance?
RQ4: To what extent do decision-makers' visual and verbal cognitive abilities influence the human–AI team performance for different explanation modalities?

To address our research questions, we employ an imperfect AI model for a bird species identification task [38]. We focus on bird species identification because using AI for wildlife conservation efforts among experts and non-experts is a rapidly growing field in research and practice [101]. 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 and *RQ*3, we design a research model based on phenomena from relevant research in HCI and conduct rigorous mixed-effects regression analyses. Our analyses leverage the appropriate reliance metrics defined by Schemmer et al. [88]. We design our study to be within-subjects for the correctness of the explanation and between-subjects for the explanation modality to provide additional insights into *RQ*1, *RQ*3, and *RQ*4. 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*2. As a result of our study, we contribute the following insights to the HCI 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 their cognitive styles on the effect of the correctness of explanations and explanations' modalities on appropriate reliance and human–AI team performance.
— *Novel Empirical Study*: To validate our proposed research model, we conduct the first empirical investigation that explores human factors (i.e., level of expertise and cognitive style) on the impact that the correctness of explanations and explanations' modalities have on appropriate reliance and human–AI team performance. We do this on a human–AI collaboration scenario across two different modalities of explanations: Natural language explanations and visual, example-based explanations. 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 decision-makers.

— *Implications for the Design of AI Support*: We provide novel insights on factors influencing decision-makers when collaborating with an imperfect AI. These insights can help developers tailor AI and XAI systems to individuals' needs.

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 the field of **Computer-Supported Cooperative Work (CSCW)** and HCI have investigated the impact that imperfect AI has on human–AI collaboration (e.g., [7, 50, 103]). Kocielnik et al. [50] 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, 23, 103]). One of those studies specifically looks at how to convey the uncertainty of outputs from Copilot [103]. By highlighting code that will most likely 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. [23] 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, 48, 99]). For example, Kim and Song [48] 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. When 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 [99].

Fewer studies investigate imperfect XAI's impact on human–AI collaboration [70]. Similar to our contributions, Papenmeier et al. [70] investigate the impact of explanation fidelity 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 whether a Tweet should be published based on its content. While Papenmeier et al. [70] 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 and 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, 20, 25, 32, 67, 68, 97, 100, 115]. One recent study investigates the impact of decision-makers' domain expertise on task accuracy in a high-stakes human–AI

collaboration task [20]. Calisto et al. [20] also look at the impact that the assertiveness of natural language explanations has on human–AI collaboration. Their study presents 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. [20], 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 CTP.

One study investigates how the level of expertise for Arabic or Indian Numerals from various versions of MNIST impacts task accuracy and model perception [32]. A similar study shows clinicians with various levels of expertise four different types of explanations, including visual, example-based explanations [100]. Similar to our study design, they show participants the three most similar example images for the example-based explanations. Another study investigates how practitioners with different expertise levels perceive explanations implemented in a manufacturing industry context [115]. They observe that practitioners with higher 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., [41]). 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 [18, 22, 26, 41, 112]. Cai et al. [18] 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 article investigates example-based explanations in a slightly different format from Yang et al. [112]. They similarly found the example-based explanations improve the users' appropriate trust in the classifier.

In a different study, Du et al. [26] examine 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 [22]. From their mixed-methods study, Chen et al. [22] 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. [46, 82, 97]). For example, Kim et al. [46] analyze a unified explanation technique from a human-centric point of view. In their work, Kim et al. [46] 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 [82]. Robbemond et al. [82] 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. [97] 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.

2.4 Cognitive Styles

Recent research in human–AI interaction emphasizes human users' pivotal role in designing and refining XAI methods [56, 116]. Miller [61] posits that the processing of explanations by humans is inherently idiosyncratic, with their cognitive styles exerting a significant influence on the information assimilation process. This notion is consistent with the well-established understanding that humans tend to "process the same information in different ways" [78, p. 267]. Moreover, using computational techniques, Riding et al. [78] explore humans' different cognitive styles. Their taxonomy encompasses two dimensions: the Wholist-Analytic style, which concerns information organization, and the Verbal-Imagery style, which concerns the representation of information. The divergent preferences within these styles are pronounced: Individuals categorized as verbal prefer textual information, while their visual counterparts exhibit an enhanced ability to absorb knowledge through visual media [78]. This dichotomy is underscored by the observation that visuals tend to manifest a cognitive inclination toward mental imagery, while verbals rely predominantly on verbal constructs [77]. Empirical research, as exemplified in the field of recommendation systems, questions the perceptual dynamics that underlie the variance in human response to different styles of explanatory representation [39]. Riefle et al. [81] extend this trajectory by exploring the relationship between humans' cognitive styles and their ability to comprehend explanations generated by XAI.

Notwithstanding the progress made in the XAI landscape and the broader spectrum of AI-assisted decision-making, the existing body of research falls short in rigorously assessing human cognitive styles, particularly in high-stakes decision-making scenarios. Consequently, this study seeks to fill this gap by conducting a human-centered empirical investigation. Its core focus revolves around exploring the impact of imperfect XAI on humans with different characteristics (i.e., cognitive style).

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 [24, 54, 87]. 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 [50, 53, 82]. However, fewer works investigate how imperfect XAI impacts humans' appropriate reliance on AI advice [70].

Thus, in this work, we draw from the conceptualization on appropriate reliance, previous research has established [7, 87, 88]. We specifically build on the conceptualization presented by Schemmer et al. [88] 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. Our conceptualization can be expanded to non-binary cases of XAI correctness by considering the proportion of the explanation that presents information that is for or against the prediction, referred to as relevance by Cabitza et al. [16].

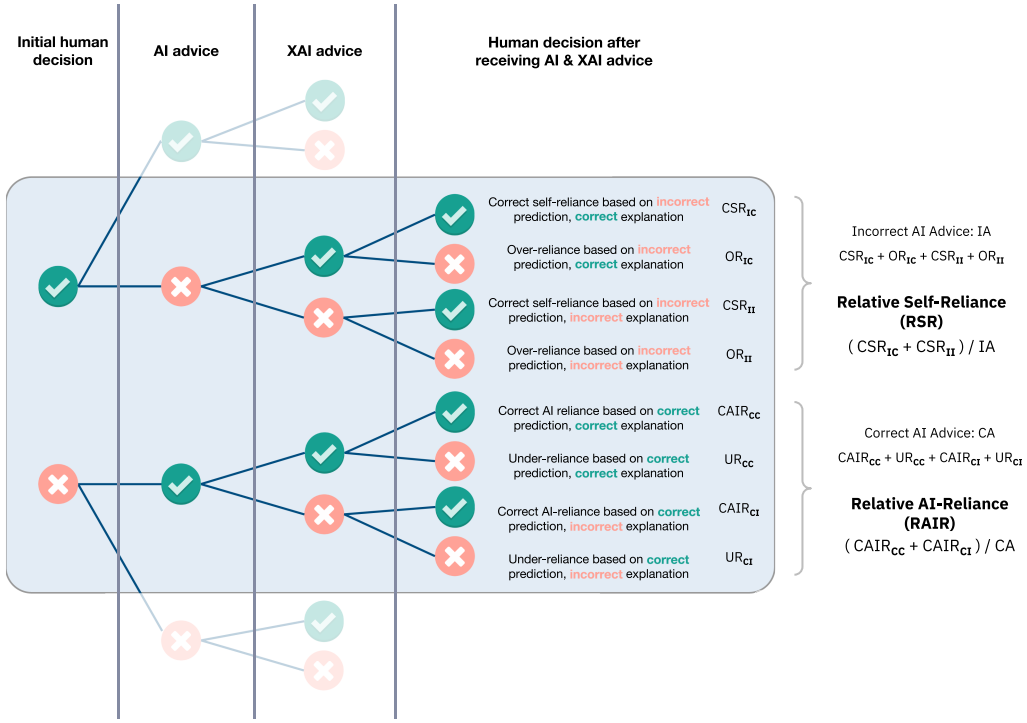


Fig. 1. Different paths that human decision-makers could follow based on receiving AI and XAI advice. This figure expands on that presented by Schemmer et al. [88] by contributing the XAI advice dimension. The XAI advice is simplified into correct and incorrect explanations. The green checkmarks represent correct advice/decisions, while the red "x" represents incorrect advice/decisions .

Introducing this new dimension unveils previously unexplored avenues within the realm of human–AI collaboration in HCI, offering a conceptual framework that allows for 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 Appendix A.1 and Table A1, 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 **Relative Self-Reliance (RSR)**, we account for the appropriateness of reliance [88].[2] RAIR comprises cases where humans correct their initial incorrect decision by overriding it with the correct AI advice. On the other hand, RSR comprises all cases in which the human makes an initially 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). Schemmer et al. [88] state that appropriate reliance is necessary to reach CTP. Thus,

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

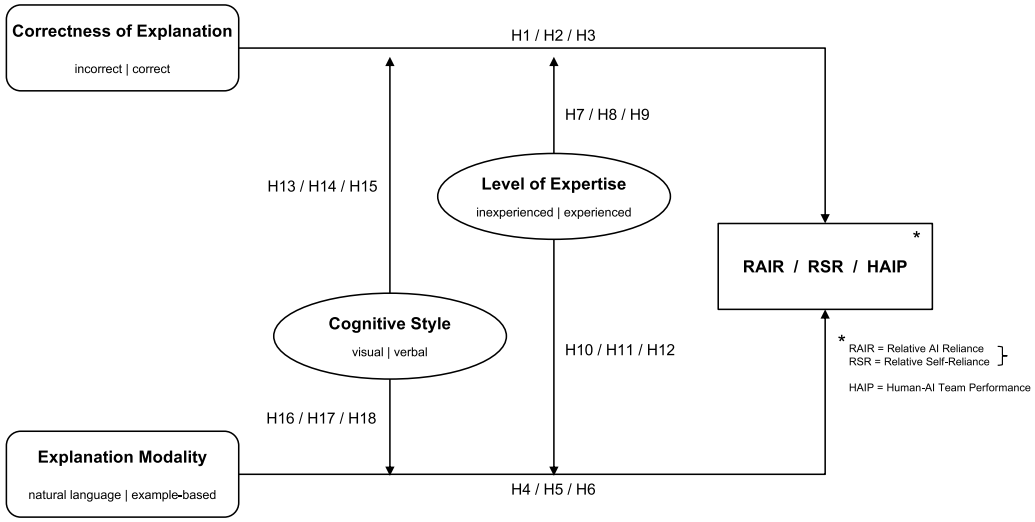


Fig. 2. Research model for collaborating with imperfect XAI systems. We analyze the moderation of the level of expertise and assertiveness on the effect of the correctness of explanation on RAIR and RSR.

to follow this line of research, we also explore how the correctness of explanations impacts the human–AI team performance. Similar to appropriate reliance, we also model the human–AI team performance as a dependent variable (see Figure 2). In the recent work of Schoeffer et al. [90], 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 not only affect the decision-making behavior but also the human–AI team performance. Accordingly, we hypothesize:

Hypothesis 1: The correctness of explanations impacts humans' RAIR in human–AI decision-making.
Hypothesis 2: The correctness of explanations impacts humans' RSR in human–AI decision-making.
Hypothesis 3: The correctness of explanations impacts the human–AI team performance in human–AI decision-making.

As explanations in human–AI decision-making are used to make AI advice more interpretable and understandable to humans, recent research investigates different explanation modalities to reveal the impact on humans' decision-making. Cai et al. [18] analyze how visual, example-based explanations affect end-users' understanding and perception of the AI model. Kim et al. [46] explore users' preferences for different interfaces and compare visual and textual explanations. Similarly, Robbemond et al. [82] shows that combining different modalities leads to higher performance when humans collaborate with an AI. These works show that different modalities impact decision-making behavior differently. Thus, we assume that in our study, the explanation modality will affect humans' reliance on AI and the human–AI team's performance. We hypothesize:

Hypothesis 4: The explanation modality impacts humans' RAIR in human–AI decision-making.
Hypothesis 5: The explanation modality impacts humans' RSR in human–AI decision-making.
Hypothesis 6: The explanation modality impacts the human–AI team performance 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, 20]. Related research in HCI investigates the role of domain knowledge in decision-making. Erjavec et al. [28] show in their behavioral experiment in online supply chains that domain knowledge positively impacts humans' confidence in decision-making. Similarly, Dikmen and Burns [25] 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. Humans with high domain-specific knowledge demonstrate an enhanced ability to discriminate between erroneous explanations and accurate ones with greater rigor [55]. This discernment is facilitated by their extensive expertise, which empowers them to readily identify and discern false information [9]. With this impact of domain knowledge on human–AI decision-making, we intend to examine how the effect of imperfect explanations on appropriate reliance and the human–AI team performance is influenced by humans' level of expertise. Similarly, we also intend to examine how the effect of different explanation modalities on appropriate reliance and the human–AI team performance is influenced by humans' level of expertise. Thus, in our study, we hypothesize:

Hypothesis 7: Humans' level of expertise moderates the effect of the correctness of explanations on humans' RAIR.
Hypothesis 8: Humans' level of expertise moderates the effect of the correctness of explanations on humans' RSR.
Hypothesis 9: Humans' level of expertise moderates the effect of the correctness of explanations on the human–AI team performance.
Hypothesis 10: Humans' level of expertise moderates the effect of the explanation modality on humans' RAIR.
Hypothesis 11: Humans' level of expertise moderates the effect of the explanation modality on humans' RSR.
Hypothesis 12: Humans' level of expertise moderates the effect of the explanation modality on the human–AI team performance.

Consistent with previous research emphasizing the importance of cognitive styles, existing evidence indicates a complex relationship between cognitive styles and individuals' understanding of explanations [81]. Cognitive style theories posit that individuals differ systematically in their preference for processing verbal versus visual information [78], which may directly affect how they comprehend and respond to explanations presented by AI systems.

This relationship is particularly relevant in domains involving visually complex decision-making tasks, such as identifying bird species from images. When explanations are presented in different modalities—for instance, natural language versus example-based visual cues—users may engage with or interpret them differently based on their cognitive style. Mismatches between the explanation modality and a user's preferred cognitive style may hinder comprehension or even lead to misinterpretation, especially when explanations are imperfect or misleading. Prior work in HCI has emphasized the importance of tailoring explanation strategies to user characteristics [40, 82], but limited attention has been paid to the moderating role of cognitive style.

Building on this assumption, we hypothesize that in visually complex task domains, such as identifying bird species on images, humans' cognitive styles may have a noticeable impact on their decision-making behavior.

Additionally, previous work in educational psychology and HCI has similarly shown that mismatches between a person's cognitive style and the modality of information presentation can hinder

comprehension and performance [33]. These findings suggest that modality-sensitive explanation design may be essential for effective human–AI interaction.

With prior research investigating not only one explanation modality but a combination of different ones (e.g., [40, 82]), it is crucial to understand how incorrect explanations might deceive humans with different cognitive styles. In our study, we thus explore two different explanation modalities (natural language explanations and example-based explanations) and how humans' cognitive styles moderate their effect on decision-making behavior. Thus, we hypothesize:

Hypothesis 13: Humans' cognitive styles moderate the effect of the correctness of explanations on humans' RAIR.
Hypothesis 14: Humans' cognitive styles moderate the effect of the correctness of explanations on humans' RSR.
Hypothesis 15: Humans' cognitive styles moderate the effect of the correctness of explanations on the human–AI team performance.
Hypothesis 16: Humans' cognitive styles moderate the effect of the explanation modality on humans' RAIR.
Hypothesis 17: Humans' cognitive styles moderate the effect of the explanation modality on humans' RSR.
Hypothesis 18: Humans' cognitive styles moderate the effect of the explanation modality on the human–AI team performance.

Our research model, shown in Figure 2, summarizes the hypotheses that we test in our mixed-methods study. With this research model, we test for factors that influence appropriate reliance and human–AI team performance.

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

Human–AI interaction is becoming core to wildlife conservation efforts [14, 35, 101]. 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 task is well-suited for people with a wide range of expertise.

Bird species identification 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 classifying diseases from medical imagery [100] or estimating building damage after disasters [64]. 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. Kayser et al. [44] 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 focus on human-centered XAI and human–AI collaboration also use the domain of bird species identification to better understand human–AI collaboration (e.g., [17, 47, 66]).

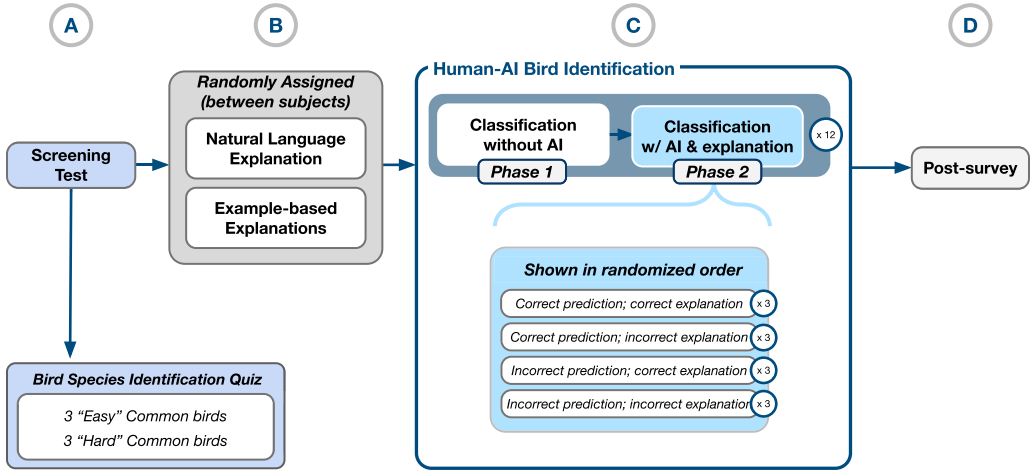


Fig. 3. We conduct a rigorous mixed-methods study leveraging a mixed design. Before participants start the task, they are shown a screening test (a). Participants are assigned an explanation modality (b) for the human–AI bird identification task. During the task, participants are shown explanations and different scenarios of correctness (c). Lastly, participants complete a post-survey (d) .

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 [101] 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. The institutional review board-approved study is 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, conducts research at a nature center, and works at a nature conservancy. The study procedure 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 Figure 3(a): In a bird identification test, we assess participants' expertise in classifying six different species of birds.[3] We distinguish the six bird images based on their level of difficulty. This difficulty level is derived from discussions with an experienced migration counter from a bird sanctuary. While previous work collects participants' self-perception of expertise [47], this method is subjective—and participants may self-perceive their skills differently. Kazemitabaar et al. [45] measure participants "experience-level" through log data instead of subjective measures. We, similarly, 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 (Figure 3(b)), participants are randomly assigned to one of two explanation types. Similar to previous studies [21, 81, 82, 97], 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 we use for the study is designed to generate natural language explanations based on fine-grained image classifications [38]. We also choose to look at example-based explanations as recent studies focus on this modality

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

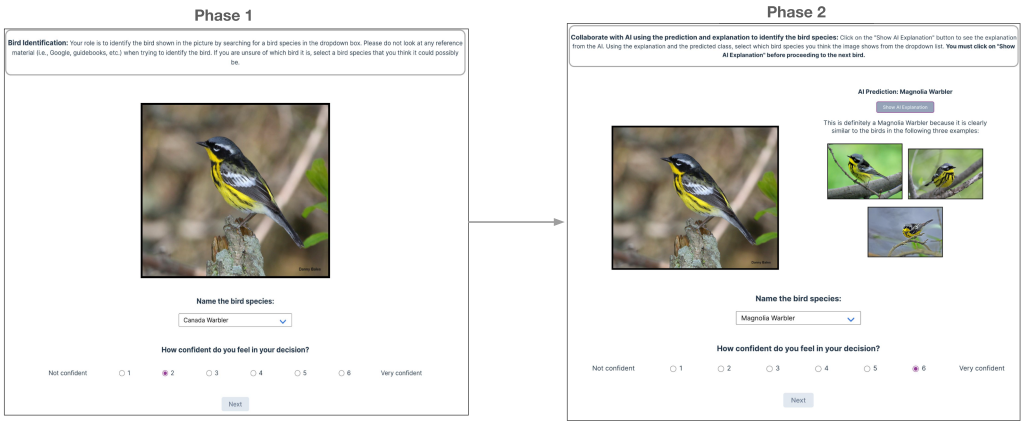


Fig. 4. Example of the two phases for a single bird image that a participant is shown in the study. This specifically shows a Magnolia Warbler (correct prediction, correct explanation), and this participant is assigned to the example-based explanations .

in human–AI collaboration [18, 22, 41, 47]. However, recent studies show that example-based explanations have potential benefits. Chen et al. [22] show that example-based explanations improve humans' performance, so much so that it leads to CTP. With promising results from previous research and numerous clinical decision-support tools proposing to incorporate example-based explanations (e.g., [8, 84]), we find it necessary to investigate the effect of example-based explanations on humans' appropriate reliance and the human–AI team performance in the context of imperfect XAI.

The human–AI bird identification task (Figure 3(c)) consists of two phases. For each treatment, 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 is corroborated with previous work [34]. Initially, participants must click on a button that shows "Show AI Explanation." The participant cannot proceed to the next question without revealing the explanation. 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, without receiving feedback on their task performance between questions.

Our study design does not include any training module because our recruitment strategy was designed to target individuals who had some level of familiarity with birding. Furthermore, our bird species identification quiz in Figure 3(a) simulates the task that participants would do in our interface.

As the AI that we are utilizing for identifying the bird species is not perfect [38], the predictions and explanations provided can be incorrect. To understand how this affects participants' appropriate reliance, we ensure that each participant is shown three samples of the following four categories in random order:

— *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. 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 complete an additional questionnaire (Figure 3(d)). Additionally, we assess participants' cognitive styles. We assess the cognitive styles by using the validated items of Kirby et al. [49] that research in HCI utilizes Riefle et al. [81]. All items are measured on a five-point Likert scale as recommended by Kirby et al. [49]. The questionnaire includes an embedded attention validation check to ensure participants are involved and attentive. Aside from this assessment, we also ask participants about their occupations and the regions in North America that they are most familiar with regarding bird species.

4.2.1 Recruitment. We recruit the participants through several communication channels 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 [69, 73]. We apply a custom filter on Prolific to target individuals who currently work in a field related to nature, science, the environment, or animals. Overall, we try to limit recruitment to only address people with prior birding knowledge to minimize the prevalence of novices' randomly guessing bird species identification. All participants are required to live in the United States, Canada, or Europe, be over 18 years of age, and be fluent in English. Based on informal conversations with birding experts from the Birding International Discord, we decided to include participants from Canada and Europe because bird experts can be familiar with bird species in North America, regardless of where they live. Eligible participants who complete the entire study are compensated with a 5 USD payment or an Amazon gift card, equivalent to a pay of 12.5 USD per hour on average.

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.

4.2.2 Participant Statistics. On average, the study takes participants 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 (Figure 3(a)). 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 groups, 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% ($STD = 12.30\%$).[4] 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.

[4]Participants performance on the bird identification test is shown in Figure A1, Appendix A.2.

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 [104] dataset and explanations from the Generating Visual Explanations model [38]. 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 filter for bird families with several species in the CUB dataset. We specifically filter out every bird class that is not a part of the Warblers, Wrens, Swallows, Sparrows, or Finches/Grosbeaks families. After applying this filter, we have 1,864 out of 5,794 images from the test set of CUB-200-2011. Of those 1,864 images, 1,609 are predicted correctly by the AI, and 255 images are predicted incorrectly by the model.

After filtering the bird species, multiple researchers on our team separately classify 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 are discussed by a subset of the research team and are excluded from consideration if an agreement is not met. In total, we identify 10 examples for each category[5] and explanation type. Sometimes, 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 as one that 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 [58, 110]. 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.

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 hold 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,[6] 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 false 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 the prediction). It is possible that a model could be relying on spurious patterns to make classifications [74]. Since we are dealing with an imperfect AI, we do not choose to exclude such cases from our dataset.

5 Results

We conduct rigorous statistical analyses to answer our research questions and validate our hypotheses. We measure appropriate reliance based on metrics defined in previous work: RAIR and

[5]The four categories are identified in Section 4.2.

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

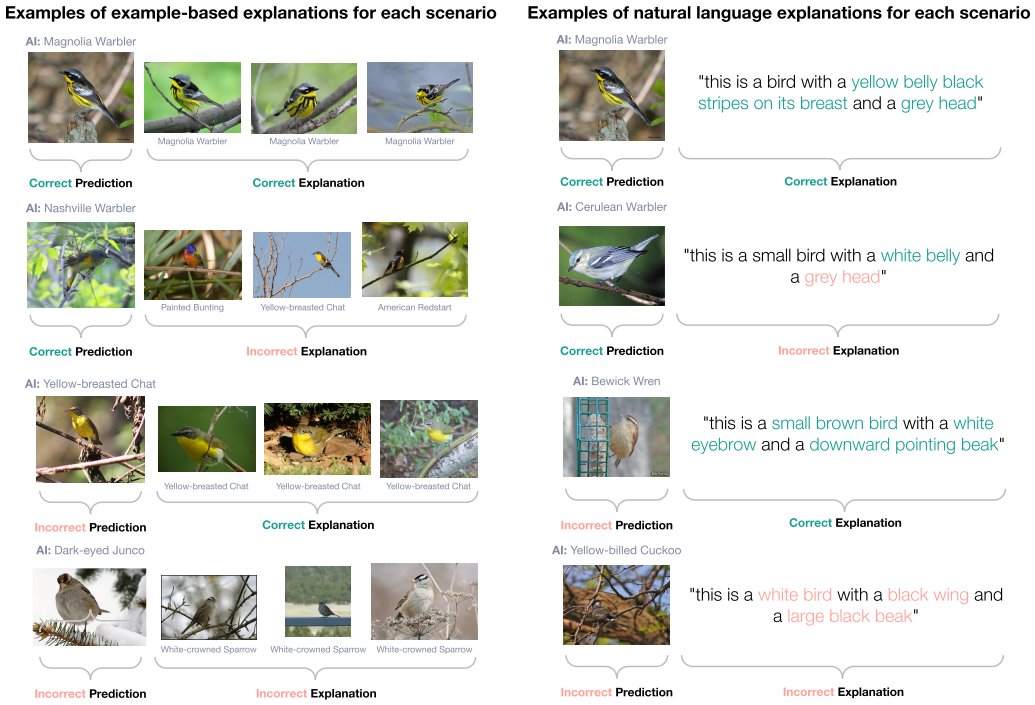


Fig. 5. Representative examples of the example-based and natural language explanations for each scenario: CC, CI, IC, and II. The class of the example-based images in the explanation is not shown to participants during the study. The study also does not show the red and green coloring on the natural language explanations. 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 .

RSR [88]. We put these constructs in relationship to explanation modalities, the correctness of explanations, and human factors (e.g., decision-makers' level of expertise and cognitive style)—and validate them through regression analyses. We align these analyses to our research model Figure 2. By doing so, we answer *RQ*1 in Section 5.1. Furthermore, we analyze the data in more detail and break it down by explanation modality, the correctness of explanations, and participants' level of expertise. By doing so, we also outline how imperfect explanations can deceive decision-makers. With this, we answer *RQ*2 in Section 5.2. In Section 5.3, we conduct mixed-effect regression analyses to explore how explanation modality and the correctness of explanation affect the human–AI team performance. Here, we answer *RQ*3. To get a deeper understanding of how participants' cognitive styles affect the human–AI team performance in different explanation modalities and to investigate which role the correctness of explanation plays, we perform subgroup analyses in Section 5.4 and answer *RQ*4. For all of our research questions, we look at two 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. Overall, we conduct regression analyses to compare the relationship between the corresponding variables in each subsection. We describe the variables used for each regression analysis in the respective subsection. Additionally, we use two-sample t-tests to compare means of groups within

the subgroup analyses to gain further insights into the study's data. For all analyses, we use the following threshold for significance: $*p < 0.1; * * p < 0.5; * * *p < 0.01$.

5.1 Effects on Appropriate Reliance

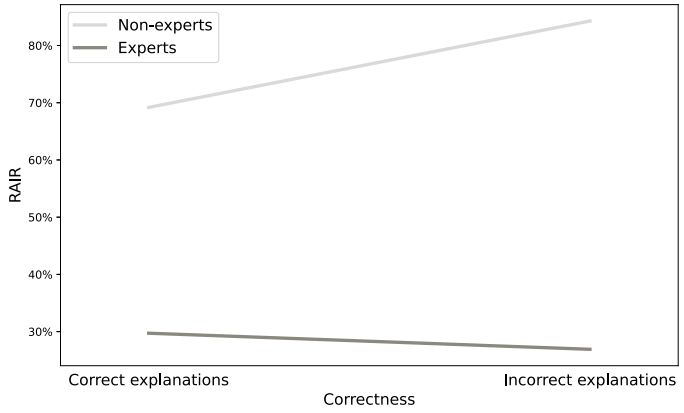
To test whether humans' level of expertise and their cognitive style moderate the relation between the correctness of explanations and the explanation modality on humans' appropriate reliance (see Figure 2), we conduct mixed-effects regression analyses where we model RAIR and RSR as dependent variables, the correctness of explanations and explanation modality as dependent variables, level of expertise and cognitive style as moderating variables, cognitive load as a control variable and the participants' ID as a random effect. An overview of the results of the regression analyses is presented in Table A2 in Appendix A.5. The categorical variables are coded as follows: Explanation modality—natural language explanations = 0, example-based explanations = 1; correctness—incorrect explanations = 0, correct explanations = 1; expertise—inexperienced participants = 0, experienced participants = 1.

5.1.1 Results for RAIR.

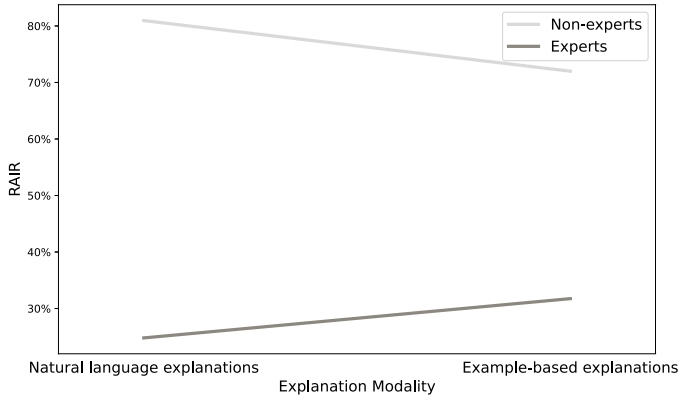
The Correctness of Explanations Has a Higher Effect on Non-Experts' RAIR. The regression analysis data for RAIR show that the interaction effect between the level of expertise and the correctness of explanations is significant (coeff = −1.428, p-value = .001). The negative coefficient indicates that the positive effect of correct explanations on RAIR is stronger for non-experts than for experts. This coefficient suggests that, for non-experts, the impact of incorrect explanations on RAIR increases by approximately 1.43 units compared to experts, indicating a substantial moderating effect. We plot the moderation effect in Figure 6(a). We can see that non-experts have a higher RAIR for incorrect explanations than for correct explanations. We can also see that non-experts rely more often correctly on AI than experts. In Section 5.2, we conduct a subgroup analysis to further analyze this interesting finding by splitting the data by the explanation modality. The data of the regression analysis also shows that there is no significant interaction effect between humans' cognitive styles and the correctness of explanations. Thus, our results support hypothesis 7, but they do not provide evidence for hypothesis 13.

The Effect of Explanation Modality on RAIR is Higher for Experts than Non-Experts. We also identify an interaction effect between explanation modality and participants' level of expertise (coeff = 1.372, p-value = .040). The interaction effect means that for participants with knowledge of the task, the effect of modality on their RAIR is greater. The effect of explanation modality on RAIR differs by 1.37 units between experts and non-experts, suggesting a meaningful amplification of modality effects based on expertise. We also note that the two explanation modalities provide different information, which affects their impact on RSR and RAIR. We plot the interaction effect in Figure 6(b). We can see that experts have a higher RAIR for example-based explanations than natural language explanations, whereas non-experts have a higher RAIR for natural language explanations than for example-based explanations. This trend could be because the task becomes more difficult when example-based images are shown, as now you need to classify the target bird and distinguish the example birds, which can be difficult for non-experts. The data does not show a moderation effect of humans' cognitive style on RAIR. Thus, our results support hypothesis 10, but they do not provide evidence for hypothesis 16.

Correct Explanations Lead to Participants More Often Correctly Relying on AI Advice than Incorrect Explanations, and Participants with a Higher Visual Cognitive Style More Often Correctly Rely on AI Advice. There are no further interaction effects in the data. Thus, we conduct a regression analysis with the moderators as independent variables to evaluate for direct effects as recommended by



(a) The moderating effect of level of expertise on the relationship of correctness of explanations on RAIR.



(b) The moderating effect of level of expertise on the relationship of explanation modality on RAIR.

Fig. 6. The moderating effects on the relationship of explanation modality on RAIR.

Hayes [36] and Warner [108]. The results of this regression analysis show that there is a direct effect of the correctness of explanations (coeff = 1.209, p-value < 0.001) and visual cognitive style (coeff = 6.279, p-value = .019) on RAIR. This could be explained by the fact that participants in the study have to conduct a visual classification task and, thus, can process the visual information better to judge the AI advice. Thus, our results support hypothesis 1, but they do not provide evidence for hypothesis 4.

5.1.2 *Results for RSR.*

The Effect of Explanation Modality on RSR is Higher for Participants with a Verbal Cognitive Style. The data shows an interaction effect between explanation modality and verbal cognitive style (coeff = -9.231, p-value = .081). The interaction effect has a negative coefficient, indicating that the effect of the explanation modality on RSR is higher for non-verbal participants. This relatively large negative coefficient indicates that as verbal cognitive style decreases (i.e., toward non-verbal preferences), the influence of explanation modality on RSR increases by over nine units, pointing to a strong differential effect even though it is only marginally significant. This effect might also stem from the aspect that example-based explanations provide different kinds of

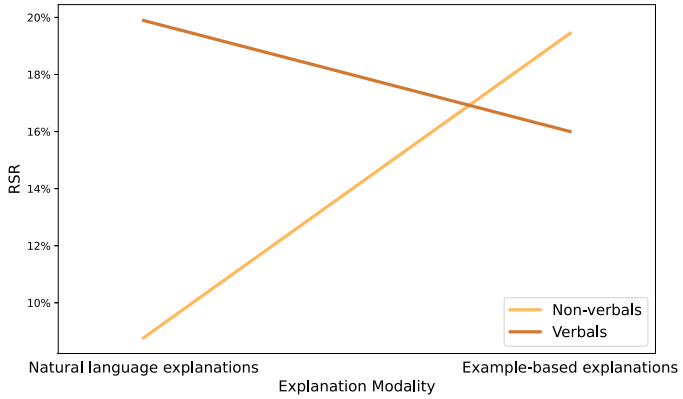


Fig. 7. The moderating effect of verbal cognitive style on the relationship of explanation modality on RSR.

information and patterns (e.g., consistency across the three examples) that might make it easier for humans to identify as incorrect. We illustrate the interaction effect in Figure 7. The figure shows that verbal participants have a higher RSR for natural language explanations, while non-verbals have a higher RSR for example-based explanations. This can be rationalized by the fact that those participants process textual information better and thus achieve higher correct self-reliance for these natural language explanations than when visual, example-based explanations are shown. Thus, the marginally significant results show a trend for hypothesis 17 and the relationship of the explanation modality on the cognitive style.

Participants Rely More Often Correctly on Themselves When the Explanation Is Incorrect. As no further interaction effects exist, we leave out the non-significant interaction terms and perform another regression analysis. We find that the correctness of explanations (coeff = −.719, p-value = .001), expertise (coeff = 3.097, p-value < .001) and visual cognitive style (coeff = −5.105, p-value = .049) have a direct effect on RSR. The coefficient of the correctness of explanations is negative, which might be because participants can determine incorrect AI output based on the explanation and thus rely more often on themselves in these cases. More expertise leads to a higher RSR, and participants with a visual cognitive style more often incorrectly rely on themselves. This is an interesting finding as it shows that inexperienced participants oftentimes do not correctly rely on themselves—potentially explained by their lack of knowledge—and that participants with a visual cognitive style seem to rely on themselves less appropriately. To disentangle these effects, we conduct subgroup analyses in the next subsection for the level of expertise, explanation modality, and correctness of explanations. Thus, our results support hypotheses 2 but they do not provide evidence for hypotheses 5, 8, 11, and 14.

5.2 Subgroup Analyses for Factors Impacting Human–AI Decision-Making Behavior

To account for different effects impacting RAIR and RSR, we conduct subgroup analyses to derive further insights from factors influencing appropriate reliance. In Figure 9, we compare RAIR and RSR for both levels of expertise and the correctness of explanations. We show this comparison for example-based explanations (the figure on the left side of Figure 9) and natural language explanations (the figure on the right side of Figure 9). By measuring RAIR and RSR for incorrect and correct explanations separately, we can calculate the deception caused by imperfect XAI (refer to Equation (A5) in Appendix A.4). We show how explanations can deceive decision-makers for different levels of expertise (experts vs. non-experts) and the correctness of explanations (correct vs. incorrect).

The figure shows that experts have a higher RSR than non-experts for incorrect and correct explanations across both modalities, validating that experts rely more on their initial decisions when given AI advice. 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 < 0.001). As a result, experts are more often falsely relying on the incorrect 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. For example-based explanations, the DoR_{RAIR} is positive, meaning that experts rightly follow correct AI advice more often when provided with correct explanations than with incorrect explanations. Similarly to the RSR cases, for the RAIR cases, the experts are being 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 a similar DoR_{RSR} in both modalities, indicating no 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 ones. For the latter, the three examples can show inconsistent bird species that are different from the image's ground truth. Thus, the DoR_{RAIR} for non-experts is at 0.26. Interestingly, for natural language explanations, the incorrect explanations are not as misleading (DoR_{RAIR} = 0.03). 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 the correctness of example-based explanations especially has an impact on humans' decision-making behavior. Following the analyses of different factors influencing decision-makers' appropriate reliance, we explore how these factors impact the human–AI team performance in the following subsections.

5.3 Effects on CTP

To explore not only how the correctness of explanations in different modalities and human factors influence decision-making behavior when humans collaborate with an AI but also to understand how these factors impact performance, we reveal the impact of these factors on human–AI team performance. With this, we answer *RQ*3. Such an analysis is especially crucial to effectively deploy AI within real-world applications and gain insights into factors that can lead to CTP [37]. Schemmer et al. [88] outline that RAIR and RSR are two factors impacting CTP. Analyzing CTP allows us to connect it to decision-makers' reliance behavior and reveal the role that explanations' correctness and modality, as well as human factors, play. For this analysis, we perform the same regression analysis as in Section 5.1 but this time with human–AI team performance as the dependent variable. The results are shown in Table A3 in Appendix A.5. In Figure 8, we show the effect sizes on the research model.

[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 40% of the *incorrect advice, incorrect explanation* cases in our study.

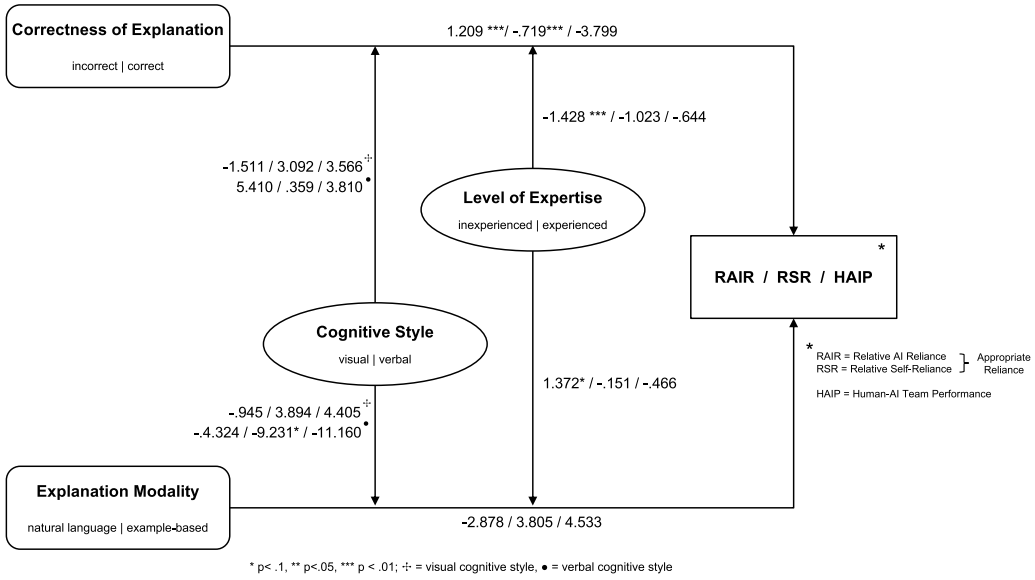


Fig. 8. The results of the regression analyses for our research model.

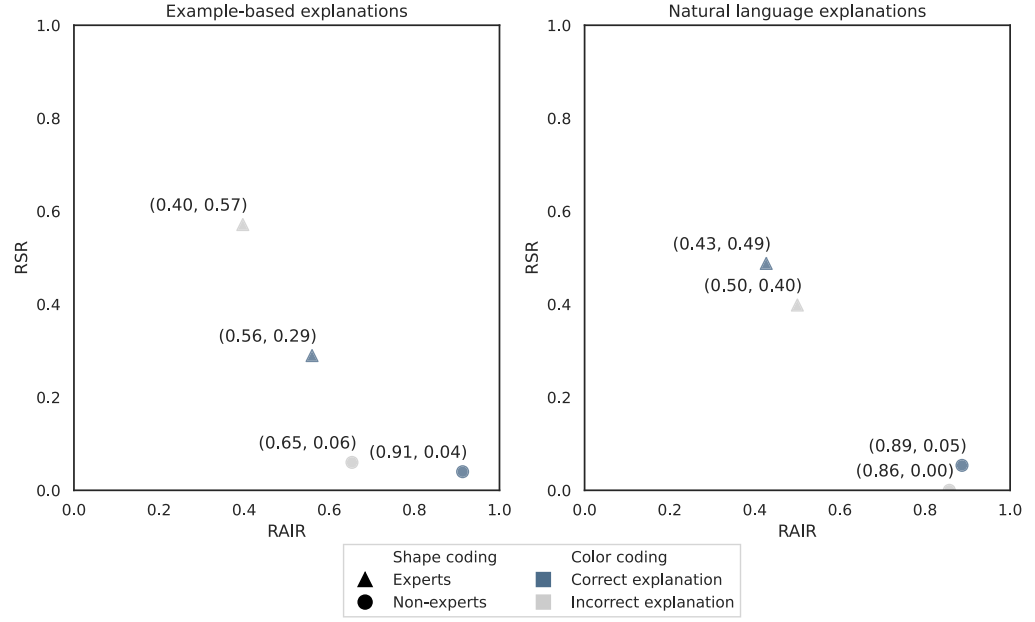


Fig. 9. Average observed RAIR (correct AI advice) and RSR (incorrect AI advice) for example-based explanations (on the left side) and natural language explanations (on the right side). We show the average RAIR and RSR for both levels of expertise as well as correct and incorrect explanations .

The Effect of the Explanation's Modality on Human–AI Team Performance Is Higher for Non-Verbal Participants. The regression analysis reveals that there is an interaction effect between the explanations' modality and the verbal cognitive style (coeff = −11.160, p-value = .044). The

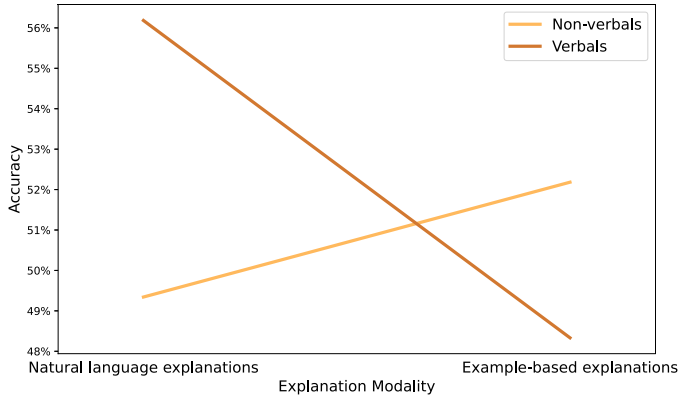


Fig. 10. The moderating effect of verbal cognitive style on the relationship of explanation modality on human–AI team performance.

negative effect indicates that the explanations' modality affects human–AI team performance more for non-verbal participants. With a coefficient magnitude of 11.160, this effect is high. We display the moderation effect in Figure 10. The figure shows that verbal participants have a higher RSR for natural language explanations, while non-verbal participants prefer example-based explanations. Since no further interaction effects exist, we drop the non-significant interaction terms in the regression analysis as in Section 5.1. The new regression model reveals a direct effect of expertise on human–AI team performance (coeff = 2.923, p-value = .002). This data shows that more experienced participants achieve a higher team performance when collaborating with the AI. Thus, our results support hypothesis 18 but they do not provide evidence for hypotheses 3, 6, 9, 12, and 15.

The insights of the moderation analyses on human–AI team performance are twofold: First, we see that the explanation modality impacts human–AI team performance—and the verbal cognitive style moderates this effect. This finding seems intuitive, as participants are presented with textual and visual explanations, so their cognitive style impacts the relationship of the type of explanation on human–AI team performance. Additionally, as the different explanation modalities contain different information, they also affect decision-making differently. Second, the expertise of participants influences the human–AI team performance. Thus, we conduct subgroup analyses in the next subchapter to better understand these effects and answer *RQ*4.

5.4 Subgroup Analyses for Factors Impacting Human–AI Team Performance

5.4.1 Level of Expertise on Human–AI Team Performance. As the analyses in Section 5.3 reveal, the level of expertise impacts human–AI team performance. Thus, in comparing the human–AI team performance for different levels of expertise, we gain further insights for a deeper understanding of how to achieve CTP. Figure 11 presents the performance of AI and humans for each treatment.

Analysis for Non-Experts versus Experts. In Figure 11, 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 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

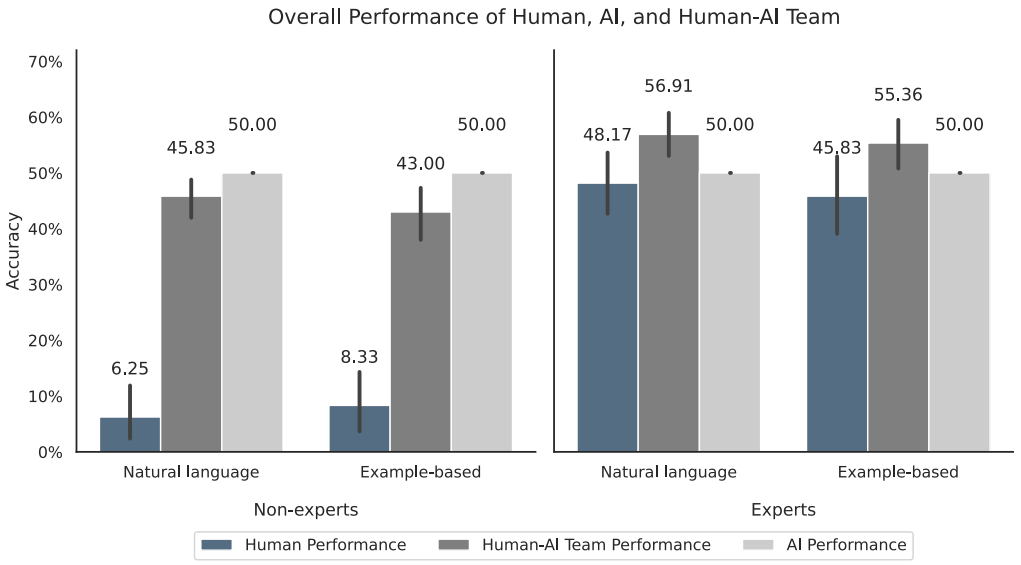


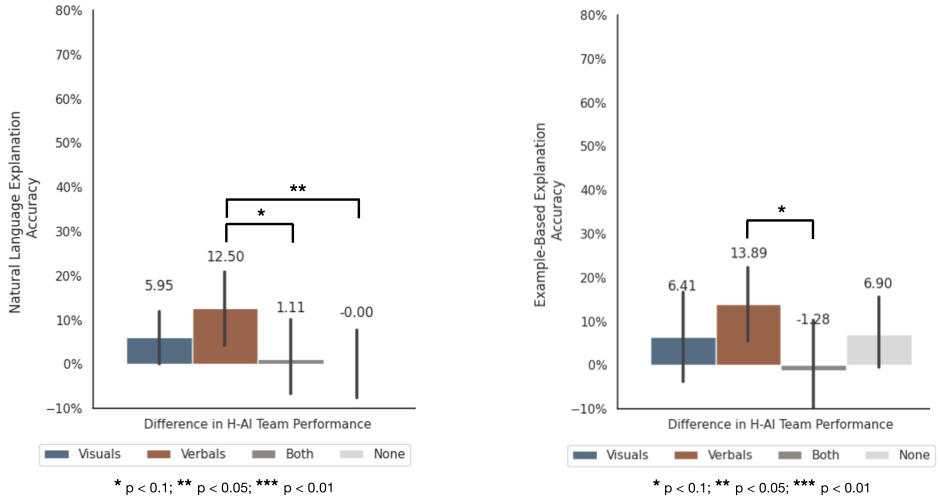
Fig. 11. The average overall performance of the human, AI, and human–AI teams for identifying 12 birds. The bar chart on the left shows the performance of the non-experts, while the bar chart on the right shows the performance of the experts .

example-based explanations improve their performance by 34.67% (task accuracy of 43.00%) when paired with the AI.

Analysis for Incorrect versus Correct Explanations. We can separate this figure into correct and incorrect explanations. When we only consider cases with correct explanations (Figure A2 in Appendix A.6, 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 A3 in Appendix A.6), we still observe CTP for the experts. However, the accuracy of non-experts' task fulfillment suffers more when incorrect explanations are shown. Non-experts' task accuracy for natural language explanations is 42.86% and 36.67% for example-based explanations.

5.4.2 Cognitive Styles on Human–AI Team Performance. We further conduct subgroup analyses based on these factors to understand the relationship between cognitive styles and imperfect explanations of human–AI team performance. When shown correct and incorrect natural language explanations, we report the average human–AI team performance for each cognitive style. We specifically analyze whether the human–AI team performs better than the human or AI alone for each cognitive style.

When explanations are correct, we observe that participants with a verbal cognitive style have the highest human–AI team performance (63.54%), achieving CTP (human performance = 43.75%). When we separate out the participants with a verbal cognitive style into experts and non-experts, we see that experts with a verbal cognitive style have the highest human–AI team performance (60.70%), while the non-experts with a verbal cognitive style have the same human–AI team performance as the AI alone (50%). While the participants with a verbal cognitive style still achieve CTP for incorrect explanations (51.04%; human performance = 35.42%), they do not have the highest human–AI team performance.



(a) The difference between the human-AI team performance when shown natural language explanations.

(b) The difference between the human-AI team performance when shown example-based explanations.

Fig. 12. Differences in human–AI team performance when shown different explanation modalities.

For incorrect explanations, participants who identify with both visual and verbal cognitive styles have the best human–AI team performance (54.44%), achieving CTP (human performance = 38.89%). This trend is also seen for experts and non-experts who identify with both visual and verbal cognitive styles. This is possibly the case because the task itself requires processing images, and the explanation requires processing text. Those with a visual cognitive style have the lowest human–AI team performance when explanations are correct (50.00%) and incorrect (44.05%). Interestingly, the participants with visual cognitive style do not achieve CTP when given natural language explanations. Given that the explanations are in a natural language format instead of a visual format, those with a visual cognitive style will struggle to integrate these explanations into their decision-making.

5.4.3 Impact of Imperfect XAI by Different Cognitive Styles and Human–AI Team Performance. We measure the impact of imperfect XAI on the human–AI team performance by looking at the difference in performance when shown correct and incorrect explanations for each cognitive style. A positive difference means that human–AI team performance is worse when shown incorrect explanations than correct explanations. A negative difference value means that human–AI team performance is better when incorrect explanations are shown than correct explanations. The difference for each cognitive style is shown in Figure 12(a).

We observe that those with a verbal cognitive style are most impacted when collaborating with an AI presenting incorrect natural language explanations (see Figure 12(a)). When shown incorrect versus correct explanations, the difference in human–AI team performance for those with a verbal cognitive style is 12.50%. Compared to participants who identify with both styles, we can see a trend that the difference in human–AI team performance is lower. Compared to participants who do not identify with either style, the difference in human–AI team performance is significantly higher. This difference in human–AI team performance could be attributed to how people with a verbal cognitive style rely on textual explanations, causing greater impact when the explanation is wrong. Participants with a visual cognitive style also decrease in human–AI team performance by

5.95% when shown incorrect natural language explanations. The human–AI team performance of participants who identify with both visual and verbal styles is only slightly worse when incorrect natural language explanations are shown. Because these participants identify as processing both types of information well, it is possible that they rely solely on textual information than people with a verbal cognitive style and rely less on the explanations for this visual classification task.

We observe that the participants who do not identify with either cognitive style have the highest team performance (56.32%) when the explanations are correct, achieving CTP (human performance = 39.08%). However, when shown incorrect explanations, the human–AI team performance for this subgroup of participants drops below 50%. It's possible these participants are following the AI advice without carefully processing the explanations.

Similar to the natural language explanations, we again see that the human–AI team performance for participants with a verbal cognitive style is impacted the most by the incorrect example-based explanations. We observe that those who identify with both cognitive styles are impacted the least. The marginally significant comparison to verbal participants shows a trend that their human–AI team performance is higher than that of those who identify with both cognitive styles. This could possibly be due to the explanations consisting of both visual and natural language formats. These findings also reveal that those with a visual cognitive style can make sense of incorrect explanations more than those with a verbal cognitive style.

6 Discussion

In our work, we investigate how imperfect XAI impacts humans' decision-making when collaborating with AI. More precisely, we assess how imperfect explanations affect humans' reliance behavior and examine the effects on the human–AI team performance. To answer our RQs, 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 [50]. With recent research looking into how humans and AI can achieve CTP [7], Schemmer et al. [88] conceptualize the role of appropriate reliance in human–AI collaboration. We extend the framework of Schemmer et al. [88] by adding another dimension: The correctness of 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 the current literature and the implications for HCI.

The Impact of Imperfect XAI Depends on Participants' Knowledge. In our study, we observe a significant moderation of humans' level of expertise on the effect of explanations' correctness on RAIR. However, we do not see this moderation for RSR. When humans are provided with wrong AI advice, their level of expertise does not moderate the impact of imperfect explanations on humans' RSR. We do see 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' RAIR, inappropriately relying on the AI. Thus, this study sets a starting point for investigating the effect of imperfect XAI on different explanation modalities.

Our Findings Show That Imperfect Explanations Impact Human–AI Team Performance. Hemmer et al. [37] argue that interpretability is a key component of human–AI complementarity. Previous user studies fail to show that incorporating XAI into AI systems can lead to CTP [31]. 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. We observe that experts reach CTP when imperfect explanations are provided. This holds true for natural language and example-based explanations. While non-experts do not reach CTP, 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 CTP when the correctness, or fidelity, of the explanation changes (see Figures A2 and A3 in Appendix A.6). Previous research discusses the impact of explanations' fidelity on humans' reliance on AI and hypothesizes that fidelity positively impacts humans' reliance behavior on AI [37]. Our results confirm this hypothesis. Furthermore, Papenmeier et al. [70] 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 A3 in Appendix A.6). Similar to our findings, Nourani et al. [67] 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. Our 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, our results can inform managers on how to assign tasks to humans with varying 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 learn that organizations using the tool modify their workflow to incorporate "checks-and-balances." For example, intro-level biologists 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.

Visual, Example-Based Explanations Can Be More Deceptive Than Natural Language Explanations. To account for the impact of imperfect XAI on humans' appropriate reliance, we establish a novel metric *DoR*, 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. Regarding 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), we find that the correctness of example-based explanations impacts RSR, while this is not the case for natural-language explanations. This suggests that participants may be more sensitive to inconsistencies or misleading cues in visual explanations than in textual ones. One possible interpretation is that visual comparisons, such as showing three visually similar bird species, make contradictions between the AI's prediction and the explanation more salient. For instance, if a participant notices that the examples do not resemble the query image—or that the examples vary among themselves—they may be more inclined to doubt the AI and revert to their own judgment. This type of visual mismatch might trigger a stronger corrective reaction than ambiguous or subtly incorrect textual explanations, which could be less obviously flawed or more easily misinterpreted. Additionally, visual explanations might activate pattern-recognition processes that experts especially rely on, which—when violated—prompt self-reliance. This aligns with prior findings that visual information, while powerful, can overwhelm users and pose challenges [97].

[8]WildMe.org.

Additionally, experts are deceived by correct explanations. This is an interesting observation that the consistency of the visual examples may explain. 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 5). 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 into no longer relying on AI 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 predicts. 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 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 incorrect explanations deceive them, 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.

Our Findings Can Guide Researchers and Practitioners on How to Assess and Design for Imperfect XAI in Human–AI Collaborations. Regardless of the modality of the explanation, 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., [47, 60]), 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 [95]. While our study is situated in a constrained bird identification task, the core insights—such as how imperfect explanations impact trust and reliance differently based on expertise—may be informative for domains where visual classification is central and expertise varies (e.g., medical imaging, biodiversity monitoring, quality control). In such contexts, AI systems can support knowledge transfer and help organizations retain and distribute expert knowledge [43, 94, 109], although further domain-specific validation is needed. Rather than providing direct prescriptions, our findings should be seen as a step toward understanding the nuanced effects of imperfect XAI that knowledge managers may eventually consider when evaluating human–AI workflows.

Building on Our Study's Findings, Several Implications Emerge for the Types of Intelligent Technologies Used When Interacting with AI Systems and the Design Considerations They Require. Our findings suggest that imperfect XAI systems should consider human cognitive styles—such as verbal and visual cognitive styles—and expertise levels, as doing so may lead to improvements in human–AI collaboration. For instance, AI systems supporting decision-making in high-stakes scenarios, such as medical imaging or biodiversity monitoring, could integrate adaptive explanation modalities to better align with the user's cognitive preferences. Furthermore, as our results show, the presence of imperfect explanations can mislead users and reduce the effectiveness of human–AI teams. This underscores the need for intelligent technologies to incorporate mechanisms that detect and mitigate the effects of explanation errors, such as confidence metrics or interactive feedback loops. By embedding these design principles, intelligent technologies can better support human decision-making, improve team performance, and reduce the risks associated with inappropriate reliance or deception in human–AI collaboration.

Our Findings Show That Human Factors (i.e., Cognitive Styles) Influence the Decision-Making Performance in the Presence of Imperfect XAI. In our study, participants conduct a visual classification task. The data in our study shows a difference in decision-making for humans with different cognitive styles. In general, humans who identify with a verbal cognitive style are impaired the most by

imperfect explanations: When shown incorrect explanations, their performance drops the most. Previous research in HCI that studies humans' cognitive styles reveals similar findings. Riefle et al. [81] show that humans who differ in information processing understand explanations differently. Similarly, Felmingham et al. [30] and Ramon et al. [75] point out the implications of cognitive styles on humans' perception of explanations in human–AI decision-making. Building on this, our findings highlight the need to broaden the human factors considered in XAI research. For instance, cognitive biases such as confirmation bias or automation bias can influence how users interpret and act on AI explanations [89]. Future work should explore how explanation design can mitigate such biases, e.g., through cognitive forcing functions [15]. Additionally, contextual and task-level factors—including task complexity, uncertainty, and decision stakes—play a significant role in shaping human–AI interaction [86, 93]. For example, explanations that support effective collaboration in low-stakes scenarios may not generalize to high-stakes environments like healthcare or conservation. Our findings also relate to AI system-level factors, such as model accuracy, which strongly influence user trust and reliance behavior [71]. The growing body of work on evaluative AI [62] emphasizes that AI systems must be designed to support not recommendations but to provide evidence for human decisions. Integrating these perspectives—on cognitive, contextual, task, and system dimensions—can help guide the development of more robust, trustworthy XAI systems that account for the complex ecosystem in which human–AI decisions unfold.

Although our findings are specific to visual decision-making and bird species classification, they highlight implications worth considering and research directions worth exploring in other domains.

7 Limitations and 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. Regarding the ecological validity of our experiment, it is important to acknowledge how experts actually go about identifying bird species for real-world tasks. For example, 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 where 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 influenced 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 be either incorrect or correct. However, as we mentioned in our work, some explanations that we classify as incorrect can contain correct evidence, making it difficult to have only 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 continuously and categorized as low fidelity and high fidelity [70]. Additionally, the correctness of explanations can be measured along several dimensions. For instance, Cabitza et al. [16] define incorrect explanations along the two dimensions of coherence and relevance. Cabitza et al. [16] define relevance as relevant information for the instance that is being explained and coherence as the AI correctly outlining the reasoning for the predicted class. Luo et al. [59] use a similar approach. Next to those dimensions, the factual correctness of how the AI is explaining an instance also plays an important role. Miró-Nicolau et al. [63] define this dimension as faithfulness of explanations. We encourage future work to explore explanation fidelity using multiple categories instead of two to understand the differences between low- and medium-fidelity explanations regarding task performance and appropriate reliance. This will provide insight into the impact of noisy explanations on

decision-making, such as when an explanation reveals some information aligned with the ground truth class and some information aligned with the predicted class.

Nuances of Error Tolerance and Impact. Our study treats AI and explanation correctness as binary conditions—correct or incorrect—to facilitate a controlled analysis. However, this design choice overlooks the nuanced spectrum of error tolerance in real-world decision-making contexts. Not all incorrect AI advice or explanations exert the same influence on decision-makers. For instance, some incorrect advice may be close to the correct class (e.g., misclassifying a closely related bird species), and some explanations, although technically incorrect, may still be useful or persuasive depending on their perceived plausibility or the human's prior knowledge. This distinction matters, as prior work suggests that humans may tolerate or even prefer certain types of "helpful" errors over others that are technically accurate but uninformative [7]. By grouping error types into binary categories, our study may obscure these nuances. Future research should examine graded measures of error severity and perceived usefulness to better understand which kinds of imperfections in AI advice and XAI truly degrade human–AI collaboration and which may be more benign or even beneficial under certain conditions.

Different Explanations Convey Different Information. The two explanation modalities throughout our analyses 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., [22, 26, 46, 47, 97]), 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 [83]. For example, the natural language explanations from Hendricks et al. [38] are feature-based, providing descriptions of features present in the image [38]. 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 [92]), there are logical errors (e.g., inconsistencies) within example-based explanations. This opens avenues for future research to investigate how different characteristics of explanation modalities impact AI-assisted decision-making.

Measurements of Cognitive Style. While this study draws on cognitive style theory to interpret individual differences, we acknowledge that the concept of cognitive styles remains debated within the psychological literature. Critics have questioned the empirical robustness of cognitive style constructs, pointing to inconsistent definitions, limited predictive validity, and challenges in measurement reliability [51]. As such, while cognitive style provides a useful conceptual lens, our interpretations should be viewed with caution and ideally complemented by more operationalized individual difference measures in future work.

Generalization to Other Image Classification Tasks. Given that the study task was repetitive in that participants had to classify several bird species, answering the same questions, it is possible that participants could have become fatigued over time. This fatigue could have impaired their willingness to exert effort in interpreting the AI explanations, regardless of their cognitive style. Future work could look into mechanisms to measure and counter such fatigue effects.

8 Conclusion

This article sets out a research model to investigate the effect of imperfect XAI on human–AI decision-making. Thus far, HCI 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 CTP 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 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 understand and fully exploit imperfect XAI in decision-making. We invite researchers to take part in this debate and hope to inspire scientists to participate in this endeavor actively.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Acknowledgements

Generative AI tools were utilized throughout this work. Specifically, ChatGPT, Claude and GitHub Copilot were employed to generate code for visualizations. Additionally, ChatGPT, DeepL Write, and Grammarly were used to enhance the writing quality of tutorials and explanations provided to participants during the experiments, as well as to improve the language across all sections of this article.

References

[1] iNaturalist. 2023. Retrieved July 2, 2023 from https://www.inaturalist.org/
[2] Merlin Bird App. 2023. Retrieved July 2, 2023 from https://merlin.allaboutbirds.org/
[3] Wildbooks from WildMe.org. 2023. Retrieved July 2, 2023 from https://www.wildme.org/
[4] Cornell Lab of Ornithology All About Birds Guide. 2023. Retrieved July 2, 2023 from https://www.allaboutbirds.org/guide/
[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.
[6] Stephan Alaniz. 2018. pytorch-gve-lrcn: PyTorch Implementation of Visual Generation and Execution for Long-Term Predictions. Retrieved from 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.
[8] Catarina Barata and Carlos Santiago. 2021. Improving the explainability of skin cancer diagnosis using CBIR. In *Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI '21)*, Part III. Springer, 550–559.
[9] Woodrow Barfield. 1986. Expert-novice differences for software: Implications for problem-solving and knowledge acquisition. *Behaviour & Information Technology* 5, 1 (1986), 15–29.
[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.

[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* 32, 1 (2021), 110–138.
[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:1710.08880. Retrieved from 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:2306.05115. Retrieved from https://arxiv.org/abs/2306.05115
[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.
[15] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
[16] Federico Cabitza, Caterina Fregosi, Andrea Campagner, and Chiara Natali. 2024. Explanations considered harmful: The impact of misleading explanations on accuracy in hybrid human-AI decision making. In *Proceedings of the World Conference on Explainable Artificial Intelligence*. Springer, 255–269.
[17] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving human-AI collaboration with descriptions of AI behavior. arXiv:2301.06937. Retrieved from https://arxiv.org/abs/2301.06937
[18] 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.
[19] 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.
[20] 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.
[21] Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to Tango: Towards theory of AI's mind. arXiv:1704.00717. Retrieved from https://arxiv.org/abs/1704.00717
[22] 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:2301.07255. Retrieved from https://arxiv.org/abs/2301.07255
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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 *Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–7.
[28] Jure Erjavec, Nadia Zaheer Khan, and Peter Trkman. 2016. The impact of personality traits and domain knowledge on decision making—A behavioral experiment. In *Proceeding of the European Conference on Information Systems (ECIS)*.
[29] Alexander Erlei, Abhinav Sharma, and Ujwal Gadiraju. 2024. Understanding choice independence and error types in human-AI collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.
[30] Claire M. Felmingham, Nikki R. Adler, Zongyuan Ge, Rachael L. Morton, Monika Janda, and Victoria J. Mar. 2021. The importance of incorporating human factors in the design and implementation of artificial intelligence for skin cancer diagnosis in the real world. *American Journal of Clinical Dermatology* 22, 2 (2021), 233–242.
[31] Raymond Fok and Daniel S. Weld. 2023. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. arXiv:2305.07722. Retrieved from https://arxiv.org/abs/2305.07722
[32] 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:2212.09342. Retrieved from https://arxiv.org/abs/2212.09342

[33] Nigel Ford and Sherry Y. Chen. 2000. Individual differences, hypermedia navigation, and learning: An empirical study. *Journal of Educational Multimedia and Hypermedia* 9, 4 (2000), 281–311.
[34] 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.
[35] 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.
[36] Andrew F. Hayes. 2017. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. Guilford Publications.
[37] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI complementarity in hybrid intelligence systems: A structured literature review. In *Proceedings of the 25th Pacific Asia Conference on Information Systems (PACIS)*, 78.
[38] 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.
[39] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Explaining review-based recommendations: Effects of profile transparency, presentation style and user characteristics. *i-Com* 19, 3 (2021), 181–200.
[40] Christina Humer, Andreas Hinterreiter, Benedikt Leichtmann, Martina Mara, and Marc Streit. 2023. Reassuring, misleading, debunking: Comparing effects of XAI methods on human decisions. *ACM Transactions on Interactive Intelligent Systems* 14, 3 (2024), 1–36.
[41] 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. *OSF Preprints* 2.
[42] 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:2306.02980. Retrieved from https://arxiv.org/abs/2306.02980
[43] 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.
[44] Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. 2022. Explaining chest X-ray pathologies in natural language. In *Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI '22)*, Part V. Springer, 701–713.
[45] 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:2309.14049. Retrieved from https://arxiv.org/abs/2309.14049
[46] 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.
[47] Sunnie S. Y. 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.
[48] Taenyun Kim and Hayeon Song. 2020. The effect of message framing and timing on the acceptance of artificial intelligence's suggestion. In *Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.
[49] John R. Kirby, Phillip J. Moore, and Neville J. Schofield. 1988. Verbal and visual learning styles. *Contemporary Educational Psychology* 13, 2 (1988), 169–184.
[50] 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.
[51] Maria Kozhevnikov. 2007. Cognitive styles in the context of modern psychology: Toward an integrated framework of cognitive style. *Psychological Bulletin* 133, 3 (2007), 464.
[52] Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. 2023. Are large language models post hoc explainers? arXiv:2310.05797. Retrieved from https://arxiv.org/abs/2310.05797
[53] 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.
[54] Benedikt Leichtmann, Andreas Hinterreiter, Christina Humer, Marc Streit, and Martina Mara. 2024. Explainable artificial intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. *International Journal of Human–Computer Interaction* 40, 17 (2024), 4787–4804.

[55] 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.

[56] Q. Vera Liao and Kush R. Varshney. 2021. Human-centered explainable AI (XAI): From algorithms to user experiences. arXiv:2110.10790. Retrieved from https://arxiv.org/abs/2110.10790

[57] Martin Lindvall, Claes Lundström, and Jonas Löwgren. 2021. Rapid assisted visual search: Supporting digital pathologists with imperfect AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 504–513.

[58] 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:2212.09968. Retrieved from https://arxiv.org/abs/2212.09968

[59] Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2022. Evaluating explanation correctness in legal decision making. In *Proceedings of the Canadian AI*.

[60] 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.

[61] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[62] Tim Miller. 2023. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–342.

[63] Miquel Miró-Nicolau, Antoni Jaume-I Capó, and Gabriel Moyà-Alcover. 2024. Assessing fidelity in XAI post-hoc techniques: A comparative study with ground truth explanations datasets. *Artificial Intelligence* 335 (2024), 104179.

[64] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. 2023. Evaluating the impact of human explanation strategies on human-AI visual decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW (Apr. 2023), 1–37.

[65] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 34, 26422–26436.

[66] Giang Nguyen, Mohammad Reza Taesiri, and Anh Nguyen. 2022. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

[67] 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.

[68] Jeroen Ooge and Katrien Verbert. 2021. Trust in prediction models: A mixed-methods pilot study on the impact of domain expertise. In *Proceedings of the 2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*. IEEE, 8–13.

[69] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[70] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. arXiv:1907.12652. Retrieved from https://arxiv.org/abs/1907.12652

[71] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction*) 29, 4 (2022), 1–33.

[72] 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.

[73] 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.

[74] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. 2021. Finding and fixing spurious patterns with explanations. arXiv:2106.02112. Retrieved from https://arxiv.org/abs/2106.02112

[75] Yanou Ramon, Tom Vermeire, David Martens, Theodoros Evgeniou, and Olivier Toubia. 2021. *How Should Artificial Intelligence Explain Itself? Understanding Preferences for Explanations Generated by XAI Algorithms*. Columbia Business School Research Paper.

[76] Alan Richardson. 1977. Verbalizer-visualizer: A cognitive style dimension. *Journal of Mental Imagery* 1, 1 (1977), 109–125.

[77] Richard Riding and Indra Cheema. 1991. Cognitive styles—An overview and integration. *Educational Psychology* 11, 3–4 (1991), 193–215.

[78] Richard J. Riding, Alan Glass, and Graeme Douglas. 1993. Individual differences in thinking: Cognitive and neurophysiological perspectives. *Educational Psychology* 13, 3–4 (1993), 267–279.
[79] Mark O. Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies* 1, 1 (2019), 33–36.
[80] Lara Riefle, Alexa Brand, Johannes Mietz, Laurin Rombach, Christian Szekat, and Carina Benz. 2022. What fits Tim might not fit Tom: Exploring the impact of user characteristics on users' experience with conversational interaction modalities. In *Wirtschaftsinformatik 2022 Proceedings*, 13.
[81] 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 43rd International Conference on Information Systems (ICIS)*.
[82] 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.
[83] 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 *Proceedings of the IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 444–448.
[84] Mahya Sadeghi, Parmit K. Chilana, and M. Stella Atkins. 2018. How users perceive content-based image retrieval for identifying skin images. In *Proceedings of the 1st International Workshops on Understanding and Interpreting Machine Learning in Medical Image Computing Applications (MLCN '18), DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018*. Springer, 141–148.
[85] Mersedeh Sadeghi, Daniel Pöttgen, Patrick Ebel, and Andreas Vogelsang. 2024. Explaining the unexplainable: The impact of misleading explanations on trust in unreliable predictions for hardly assessable tasks. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, 36–46.
[86] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A missing piece in the puzzle: Considering the role of task complexity in human-AI decision making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, 215–227.
[87] 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.
[88] 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.
[89] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. On the influence of explainable AI on automation bias. In *Proceedings of the European Conference on Information Systems (ECIS)*.
[90] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On explanations, fairness, and appropriate reliance in human-AI decision-making. arXiv:2209.11812. Retrieved from https://arxiv.org/abs/2209.11812
[91] Ben Shneiderman. 2021. Human-centered AI. *Issues in Science and Technology* 37, 2 (2021), 56–61.
[92] Francesco Sovrano, Kevin Ashley, and Alberto Bacchelli. 2023. Toward eliminating hallucinations: GPT-based explanatory AI for intelligent textbooks and documentation. In *CEUR Workshop Proceedings*, No. 3444. CEUR-WS, 54–65.
[93] Philipp Spitzer, Joshua Holstein, Patrick Hemmer, Michael Vössing, Niklas Kühl, Dominik Martin, and Gerhard Satzger. 2025. Human delegation behavior in human-AI collaboration: The effect of contextual information. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–28.
[94] Philipp Spitzer, Niklas Kühl, and Marc Goutier. 2022. Training novices: The role of human-AI collaboration and knowledge transfer. In *Proceedings of the Workshop on Human-Machine Collaboration and Teaming (HM-CaT '22)*.
[95] Philipp Spitzer, Niklas Kühl, Daniel Heinz, and Gerhard Satzger. 2023. ML-based teaching systems: A conceptual framework. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–25.
[96] Ramya Srinivasan and Ajay Chander. 2021. Explanation perspectives from the cognitive sciences—A survey. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*, 4812–4818.
[97] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: The effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 109–119.
[98] 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.
[99] Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. 2023. How displaying AI confidence affects reliance and hybrid human-AI performance. In *Proceedings of the HHAI 2023: Augmenting Human Intellect*. IOS Press, 234–242.

[100] 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.
[101] 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.
[102] 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:2304.02202. Retrieved from https://arxiv.org/abs/2304.02202
[103] 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:2302.07248. Retrieved from https://arxiv.org/abs/2302.07248
[104] 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.
[105] 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.
[106] 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.
[107] 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.
[108] Rebecca M. Warner. 2012. *Applied Statistics: From Bivariate through Multivariate Techniques*. Sage Publications.
[109] 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.
[110] 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: The Preprint Server for Health Sciences*. DOI: 10.1101/2023.04.18.23288752
[111] Wei Xu. 2019. Toward human-centered AI: A perspective from human-computer interaction. *Interactions* 26, 4 (2019), 42–46.
[112] 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.
[113] 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.
[114] 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 *Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1221–1232.
[115] 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.
[116] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G. Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *Proceedings of the 2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8.

Appendix A

A.1 Appropriate Reliance with Imperfect XAI

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

$$RSR\ (Relative\ Self-Reliance) = \frac{\sum_{i=0}^{N}\left(CSR_{IC,i} + CSR_{II,i}\right)}{\sum_{i=0}^{N} IA_i}, \tag{A1}$$

$$RAIR\ (Relative\ AI\ Reliance) = \frac{\sum_{i=0}^{N}\left(CAIR_{CC,i} + CAIR_{CI,i}\right)}{\sum_{i=0}^{N} CA_i}. \tag{A2}$$

Table A1. 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.

A correct AI explanation corresponds 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.

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.

In Figure A1, we show the performance on the bird test based on the experts and non-experts grouping we do.

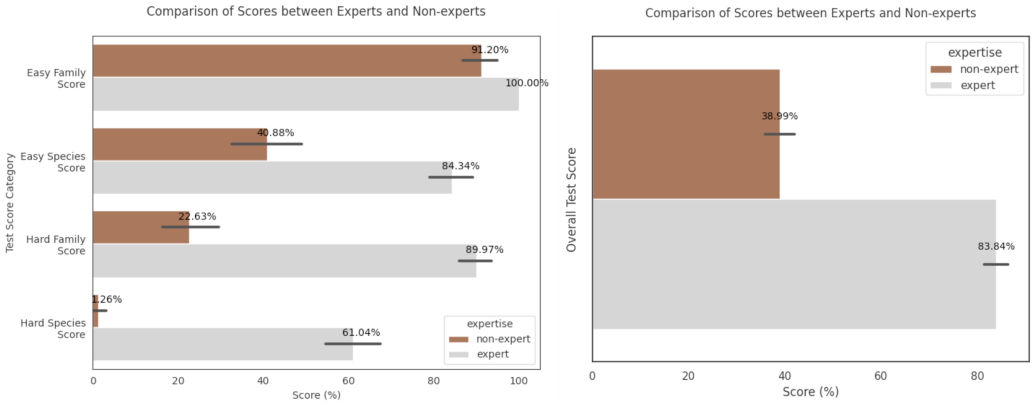


Fig. A1. The left half of this figure shows how participants (experts and non-experts) perform on average for the easy birds and three hard birds. We calculate their family accuracy as well as species accuracy. The right half of this figure shows the average overall score by combining the four scores from the left .

A.3 Explanation Modalities

Natural Language Explanations. The natural language explanations are generated by the model proposed by Hendricks et al. [38]. We follow the PyTorch implementation [6] of the model by Hendricks et al. to obtain the natural language explanations since the original model from Hendricks et al. is 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 [18]. Another study generates the example-based explanation by calculating the L_2 distance of the embedded features [65]. We generate the example-based explanations by following methods used in previous works [8, 100]. As done by Tschandl et al. [100] 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. [38]. 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 show participants the three most similar examples [32].

We do not provide additional details on the assertiveness component of the explanations as we did not find any statistical significance between experts and non-experts. Please refer to for additional details and findings in [64].

A.4 Quantitative and Qualitative Metrics

Quantitative Metrics. We quantitatively calculate appropriate reliance across the four dimensions defined by Schemmer et al. [88]: 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. Contrarily, over-reliance represents the scenario in which the human makes an initial correct decision but overrides her own decision with the incorrect AI advice. Following the appropriate reliance metrics defined by Schemmer et al. [88], 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. To measure this impact, we look at the DoR caused by imperfect XAI. For RAIR, we can apply the following:

$$DoR_{RAIR} = RAIR_C - RAIR_I. \tag{A3}$$

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

$$DoR_{RSR} = RSR_C - RSR_I. \tag{A4}$$

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:

$$DoR(RAIR, RSR) = \sqrt{(RAIR_C - RAIR_I)^2 + (RSR_C - RSR_I)^2}. \tag{A5}$$

According to the conceptualization of Appropriateness of Reliance by Schemmer et al. [88], this results in the following:

$$DoR_{AoR} = AoR_C(RAIR, RSR) - AoR_I(RAIR, RSR). \tag{A6}$$

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. For instance, if ornithologists use an AI application to help them classify birds species for which they are supported through advice and explanations, DoR measures the impact of the correctness of explanations on their reliance behavior. By taking into account RAIR and RSR, it is possible to quantify how much an incorrect explanation leads to over-relying or under-relying on the AI compared to correct explanations. If all the AI application's support for incorrect explanations lead to a lower correct self-reliance (all the instances in which the human is correct, the AI is incorrect, and the human relies on themselves) compared to correct explanations, then DoR would be positive and quantify how high this negative effect of incorrect explanations is.

The assessment of participants' cognitive styles relies on rigorously validated items, as initially proposed by Kirby et al. [49], originating from the work of Richardson [76]. Overall, participants have to rate ten items for each cognitive style (verbal and visual) on a five-point Likert scale. We separate participants into visual and non-visual groups (and similarly for the verbal style) by differentiating participants based on the sum of their ratings and comparing them to the median.

Lastly, as defined by previous work (e.g., [7, 34, 37]), 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 the 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% .

A.5 Moderation Analyses

Table A2. Moderation Analyses of the Correctness of Natural Language and Example-Based Explanations on RAIR and RSR with the Level of Expertise and Cognitive Styles

	RAIR		RSR	
	Coeff	SE	Coeff	SE
(Intercept)	2.288	2.799	−4.183	2.807
Explanation modality	−2.878	3.429	3.805	3.351
Correctness	.380	2.018	−3.138	2.588
Expertise	**−2.709*****	.511	**3.654*****	.943
Visual	**7.367***	4.397	**−9.994****	4.420
Verbal	−6.633	4.123	**8.642***	4.507
Cognitive load	**−2.423***	1.389	.022	1.306
Correctness × visual	−1.511	3.092	5.410	3.846
Correctness × verbal	2.949	3.107	.359	3.851
Correctness × expertise	**−1.428*****	.240	−1.023	.949
Explanation modality × visual	−.945	5.235	3.894	5.125
Explanation modality × verbal	4.324	5.162	**−9.231***	5.297
Explanation modality × expertise	**1.372***	.668	−.151	1.024

*p < .1; **p < .05; ***p < .01. Significant results are highlighted in bold.

Table A3. Moderation Analyses of the Correctness of Natural Language and Example-Based Explanations on RAIR and RSR with the Level of Expertise and Cognitive Styles

	Human–AI team performance	
	Coeff	SE
(Intercept)	−2.234	3.141
Explanation modality	4.533	3.497
Correctness	−3.799	3.459
Expertise	**2.923****	.952
Visual	**−8.328***	4.970
Verbal	6.639	5.039
Cognitive load	−.117	.270
Correctness × visual	3.566	5.281
Correctness × verbal	3.810	5.259
Correctness × expertise	−.664	.989
Explanation modality × visual	4.405	5.368
Explanation modality × verbal	**−11.160****	5.537
Explanation modality × expertise	−.466	.971

*p < .1; **p < .05; ***p < .01. Significant results are highlighted in bold.

A.6 Human–AI Team Performance

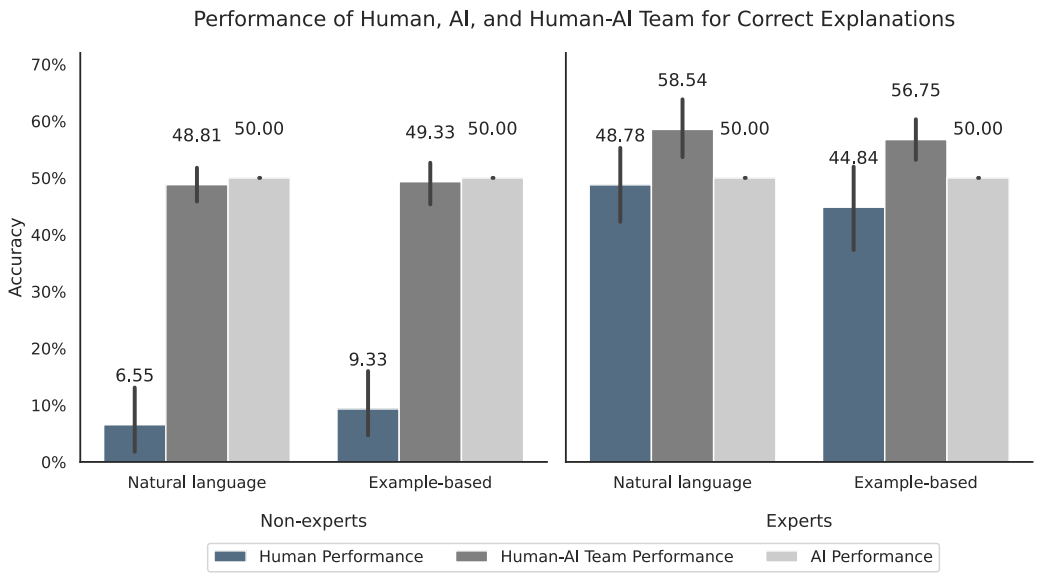


Fig. A2. Performance of the human, AI, and human–AI team specifically for correct explanations. This represents 6 birds from the 12 that participants saw .

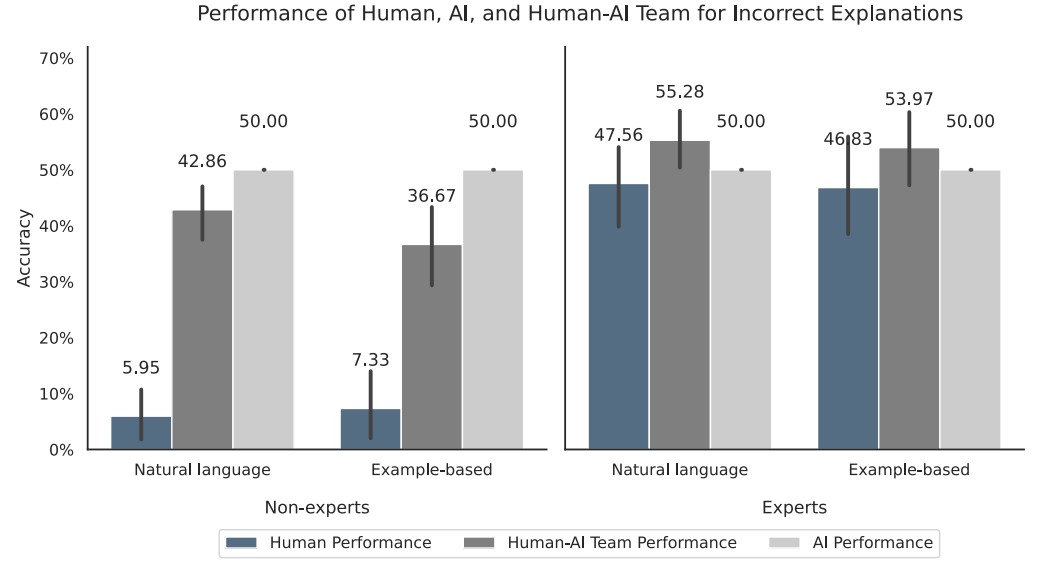


Fig. A3. Performance of the human, AI, and human–AI team specifically for incorrect explanations. This represents 6 birds from the 12 that participants saw .

Received 8 January 2025; revised 15 May 2025; accepted 13 July 2025