XAI Show and Tell: Understanding the Impact of Visual and Natural Language Explanations on Non-Experts' Decisions

Katelyn Morrison
kcmorris@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

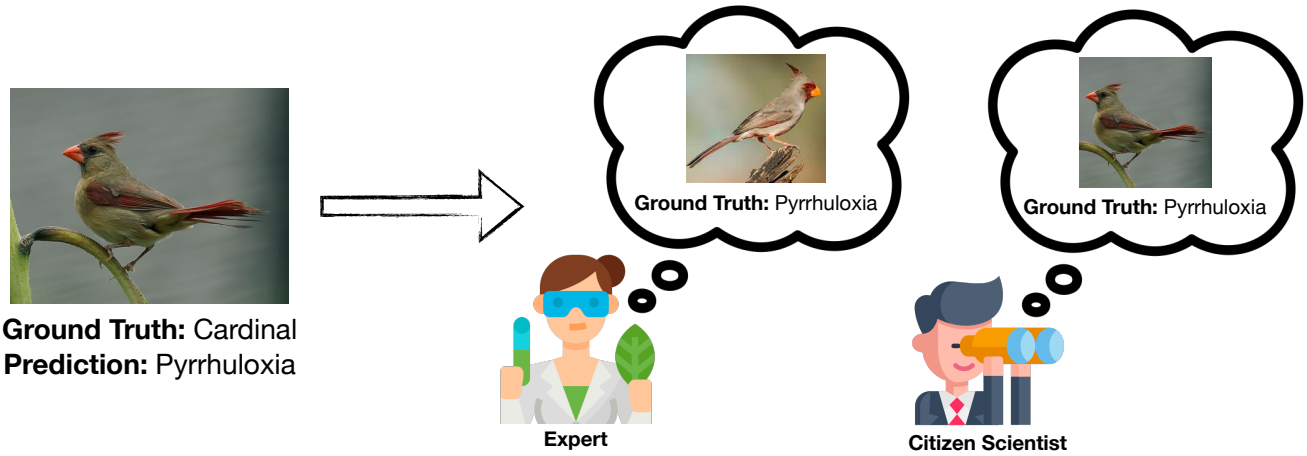


Figure 1: Motivating the need to understand how different explanation techniques impact non-experts' decisions to benefit human-AI collaboration tools used across the wildlife conservation domain. Citizen scientists, or non-experts, may be lacking information that experts are aware of, which is why it's necessary and important to explore which explanations improve their decision-making when collaborating with AI.

ABSTRACT

Novel explainability techniques are rapidly being developed to improve human-AI collaborations across various domains. As a result, several studies have evaluated how these techniques in various domains impact end-users' trust in AI and task accuracy when collaborating with AI. Other studies have empirically evaluated how interpretable one technique is over another in various domains. However, some techniques and domains receive more attention than others. Few works have evaluated and compared how natural language explanations and example-based explanations (visual) impact non-experts' trust in AI and task accuracy. We take the task of bird species classification, where a user collaborates with AI to classify pictures of birds without explanations, with natural language explanations, and with visual explanations. We provide preliminary findings on how these explanations impact non-experts' decision-making and provide directions for future works.

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

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Computing methodologies** → **Intelligent agents**; • **Human-centered computing** → *Empirical studies in HCI*.

KEYWORDS

human-AI collaboration, trust, natural language explanations, visual explanations

ACM Reference Format:
Katelyn Morrison. 2022. XAI Show and Tell: Understanding the Impact of Visual and Natural Language Explanations on Non-Experts' Decisions. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Visual decision-making is prevalent across the life sciences such as identifying abnormalities in medical imagery [11]; monitoring biodiversity by analyzing and counting species such as sea turtles or moths from camera traps [2], and identifying damaged buildings after natural disasters and armed conflicts from satellite imagery [14]. Specifically, in the wildlife domains, there are numerous citizen scientists and activists who want to partake in the efforts just as much as researchers and domain-experts [2, 9]. This domain is unique in the sense that non-experts (or *novices*) and experts are working on the same task with the same AI models. However, experts may have

more context outside of the classification and confidence the AI provides them that a non-expert may not. For example, an expert ornithologist may know that female cardinals are mostly a light brown color instead of bright red like the male cardinals. If an AI predicted a female cardinal to be a cardinal but did not explain why and the novice user was only familiar with what male cardinals look like, the novice may incorrectly classify the bird and inappropriately disagree with the AI. Experts tend to also be more aware of features related to the bird's habitat, migration patterns, and movements whereas citizen scientists may not. In this sense, not only do we need to tailor explanations toward users with different levels of expertise for the same human-AI collaboration task, but we need to understand what information is most appropriate to show them and how to show it in order to achieve appropriate trust. Achieving appropriate trust in human-AI collaboration generally has been widely studied, however understanding how the level of expertise plays a role is under-explored.

Recently, human-computer interaction researchers have pushed for more human-centered explainability techniques. However, some machine learning researchers claim their explainability techniques are more human-centered simply because they are in the form of natural language which they claim is natural to humans and easier for humans to interpret [5, 7]. Natural language explanation can involve a lot more than just rationalizing a prediction, though. For example, the assertiveness of a message (or in our case a natural language explanation) can potentially impact that message is interpreted and perceived [1]. Furthermore, the information within the natural language explanation can be full of jargon from the perception of a non-expert which could make the explanation misleading or confusing. Another explanation technique that has been under-explored is normative explanations which is a type of example-based explanations [3]. These explanations may be less informative than the information presented in a natural language explanation, however, they may be more natural to non-experts providing the "right" information for the non-expert to appropriately agree or disagree with the AI.

We explore the **impact of natural language explanations and visual explanations** on non-experts' decisions as well as **how the assertiveness of the natural language explanation impacts non-experts' decisions**. We use mixed methods and a self-assessment methodology through a human-AI visual decision-making task of bird species identification to address our research questions.

2 RELATED WORK

I will briefly touch on the few core areas that are important to this work: human-AI collaboration in wildlife conservation; the creation and evaluation of visual and natural language explanations; and the impact of assertiveness on team performance. I will only touch on one to three core papers that I believe are the most valuable to discuss.

2.1 Human-AI Collaboration in Wildlife Conservation

Based on my current knowledge, I have seen very few works addressing the challenges of human-AI collaboration within wildlife conservation, a high-stakes visual decision-making task. From informal conversations, I have had with co-founders at WildMe.org and from reading their publications, I will briefly describe what human-AI collaboration looks like for wildlife conservation efforts and briefly describe some of the challenges. As a preface, WildMe is a non-profit organization that developed an open-source tool for researchers, biologists, and citizen scientists to help them analyze photos and videos of wildlife [8]. The image analysis pipeline at WildMe performs object detection (showing bounding boxes), species classification, and unique individual identification. This pipeline allows researchers and citizen scientists from around the world to work together in order to predict population sizes, understand migration trends and relationships, and implement policies to protect wildlife. However, in some cases, the object detection model does not pick up on species that are partially occluded by trucks or trees and bushes which frustrates the end-users. In other cases, researchers or citizen scientists may not fully trust that an individual was or was not already logged by another researcher or a citizen scientist.

2.2 Visual and Natural Language Explanations

Previous literature across different fields of machine learning has proposed generating natural language explanations. For example, Hendricks et al propose an inherently interpretable model that generates visual explanations for fine-grained image classification where the visual explanations are natural language explanations that are class- and image-relevant [5]. The authors evaluate their method on the CUB-200-2011 dataset [12]. In reinforcement learning applications, Nguyen et al propose generating explanations from RL agent's actions to improve interpretability [7]. They evaluate their method on two different games: pong and MiniGgrid.

Aside from natural language explanations, another approach is to use example-based explanations. Cai et all is one paper that proposes and evaluates example-based explanations with non-experts for a drawing guessing game [3]. The authors found the normative explanations to help users better understand how the AI made decisions. Another paper by Yang et al investigates example-based explanations in a slightly different format from Cai [13]. They similarly found the example-based explanations improved the users' appropriate trust in the classifier.

2.3 Assertiveness & Team Performance

Very few works have considered looking at how the assertiveness of natural language explanations impacts decision-makers' trust in AI primarily because the technique is still very novel and not widely implemented or stable. However, we can look at studies from psychology to understand how various characteristics and personalities of team players impact the team's performance overall. One particular study found that the most critical member of the team improved team performance and satisfaction if they were assertive [10]. Similarly, studies [4, 6] both found that assertiveness is necessary for effective collaborations.

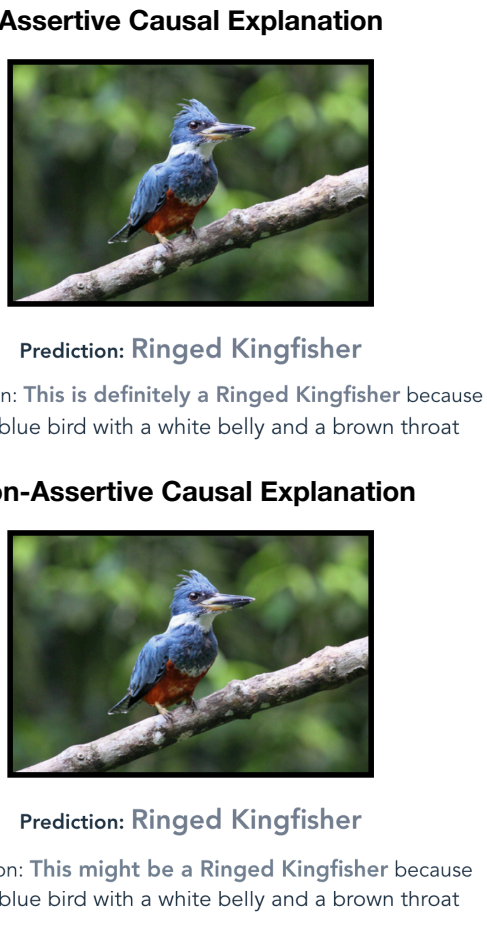


Figure 2: Assertive and non-assertive explanations for this Ringed Kingfisher. This is an example of when the generated natural language explanation was misleading because it did not align with the visual features shown in the image.

3 EXPERIMENTAL DESIGN

To address our research questions, we designed a pilot study and post-task survey based on bird classification. We use the CUB-200-2011 dataset [12], a dataset of 200 different bird species that can be seen throughout North America. This dataset has been used for a wide variety of machine-learning tasks, including fine-grained image classification and image captioning tasks. The dataset contains 11788 images, with 5794 of those images belonging to the test set. We describe the design of this pilot study using the CUB-200-2011 dataset as seen in Figure 3 and survey below.

Since we are showing birds to participants with potentially varying levels of expertise in bird watching, we ask participants to identify the bird species with no assistance from AI. Participants were given a list of 200 bird species and rated their confidence in their guesses. We start off with this so we can determine if the participant already knew the bird species before they started collaborating with the AI. Next, participants were shown the AI's classification for the bird, and participants were asked if they agreed or disagreed with the AI. We also asked participants to rate their confidence in their decision to agree or disagree. Next, each participant was shown two different types of explanations along with the AI's classification in random order: a natural language explanation and a normative explanation. The conditions of natural language explanation, example-based explanation, or no explanation were designed to be within-subject variables for this study so that we could ask participants at the end to compare the three conditions. We also designed the study to also have a between-subjects variable which is the assertiveness of the natural language explanation. Participants were assigned to one of two natural language explanations for the study: **assertive** (top-half of Figure 2) or **non-assertive** (bottom-half of Figure 2).

Each participant was shown the same four bird images. To minimize the ordering effect, the bird images were shown to participants in random order. Two of the four birds were incorrectly predicted by the AI. One of the images predicted incorrectly by the Awe also has a natural language explanation that is explicitly misleading, and the other predicted incorrectly has a natural language explanation that is not explicitly misleading. we present participants two cases where the Awe correctly classifies the bird and two where the Awe incorrectly classifies the bird to control for random guessing from participants. This means that if participants get an appropriate trust score above 50%, then it is less likely that they randomly guessed to agree or disagree with the Awe throughout the task.

Post-Interview Survey. All participants were required to complete a short post-interview survey after they finished the main task. The survey consisted of seven questions. A complete list of the questions is provided in the appendix. The purpose of the survey is to collect self-reporting measures for the participant's level of expertise in bird watching and occupation.

Metrics. We define the **appropriate trust** metric to be a percentage based on a sum of the number of times the participant correctly agreed with the Awe when the classification was correct and the number of times the participant correctly disagreed with the Awe when the Awe was wrong out of the total number of birds that were classified. We use **confidence in decision** as a proxy for subjective trust in the Awe (*which granted is not the best metric or proxy for subjective trust and is further discussed in the Limitations section*). Participants were presented with a 4-point Likert scale (extremely not confident to extremely confident).

Recruitment. We recruited 16 participants in total. However, four participants did not complete the post-task survey. Participants were not paid to participate in this pilot study which was made clear in the consent form before the study began. Participants were recruited from Facebook, personal connections, and various Slack channels. The participant recruitment was limited to a few outlets because we did not want to exhaust the supply of participants that we knew would provide valuable data for a pilot study where they would not get paid.

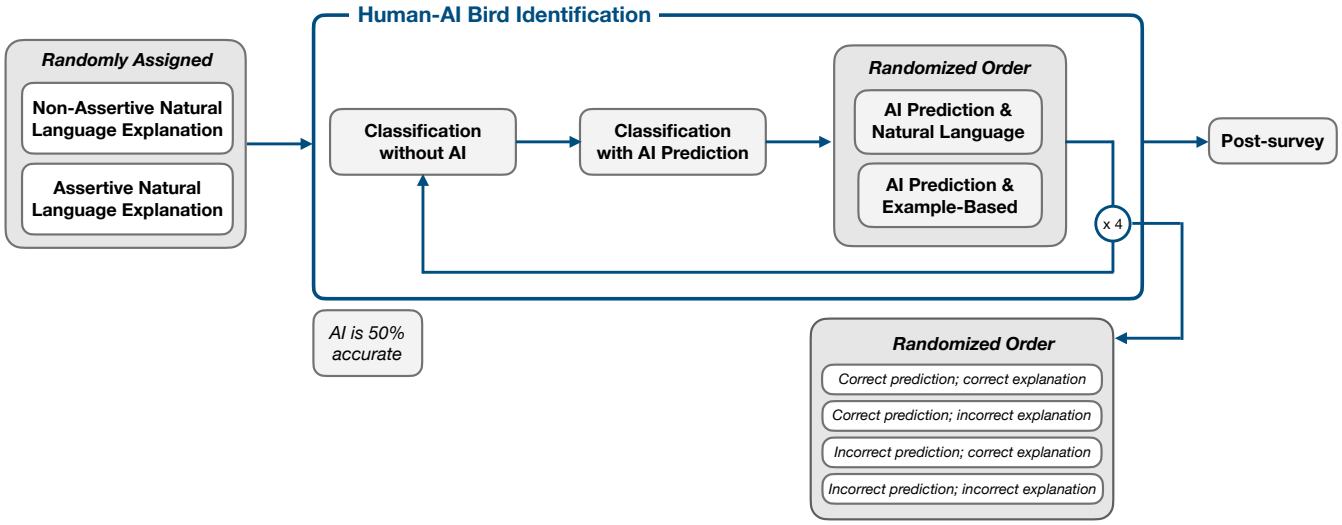


Figure 3: Diagram showing the flow of the study. Every participant was randomly assigned to either a non-assertive or an assertive natural language explanation. For four different birds, participants classify the bird without the AI, then see the AI's classification and determine whether they agree or disagree with the AI, then see a natural language explanation and example-based explanation in a random order along with the AI's classification. Lastly, they take a survey at the end to detail their expertise in bird watching.

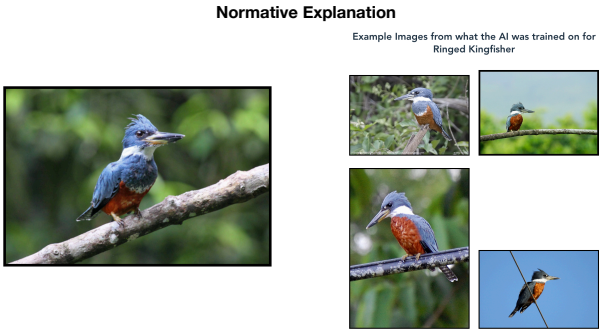


Figure 4: The four most similar images to the image being classified are selected from the class that was predicted by the AI.

3.1 Generating Natural Language Explanations

To generate natural language explanations for each bird image, we use the open-sourced PyTorch implementation[1] of the *Generating Visual Explanations* paper [5]. The model that the authors propose in [5] is inherently interpretable that combines a fine-grained classification model with an LSTM network. In short, the class predicted by the fine-grained classification model and ground truth descriptions of each image are sent through the sentence classifier to generate a natural language explanation that is class- and image-relevant. The authors provide a dictionary of their model's class predictions for each test set image in the CUB-200-2011 dataset. Their classification model is 83.5% accurate on the test set. we used the open-sourced PyTorch implementation to evaluate the model on the test set to generate the natural language explanations. For example, for the `Ringed Kingfisher`, the generated natural language explanation was, "this is a blue bird with a white belly and a brown throat". For this study, we append the non-assertive causal and assertive causal portions such as, "This is definitely a Ringed Kingfisher because ...".

We searched through the test set class predictions and natural language explanations to find images we could use in the study. we wanted to evaluate scenarios when the Awe classification was correct and incorrect. we also wanted to evaluate when the natural language explanation was misleading. we identified only four images (one for each scenario) to use in the pilot study. A full set of the images and resulting natural language explanations can be found in the Appendix.

3.2 Generating Normative Explanations

To create the normative explanations slightly modified from the definition from [3], we used the DeepImageSearch Python library[2] to find the most similar images to a given image. We limited it to show the top four most similar images from the class that the AI predicted. The images were only selected from the training set images for that class. All example-based images that were generated for each bird used are shown in the Appendix.

[1] https://github.com/salaniz/pytorch-gve-lrcn

[2] https://github.com/TechyNilesh/DeepImageSearch

4 RESULTS

4.1 Preliminary Analyses

For my preliminary analyses, we only report graphical exploratory data analyses and perform some statistical tests to determine the significance of observed trends.

Visual vs. Natural Language Explanations. One of the main research questions we asked was how visual and natural language explanations impact non-experts' decisions. In Figure 5, we show each scenario and the number of participants that agreed with the AI and disagreed with the AI for each explanation type (none, natural language, and example-based). When the classification and explanation were correct (top left facet in Figure 5), we see that 100% of participants correctly agreed with the AI for the natural language and example-based explanations. However, we see that 25% (4 out of 16 participants) incorrectly disagreed with the AI's classification when shown no explanation.

The extreme opposite scenario (incorrect prediction and explanation) shown in the bottom right of Figure 5 shows that only 2 participants incorrectly agreed with the AI when the classification was incorrect. We also see that 14 out of 16 participants correctly disagreed with the AI's classification when not shown any explanation and all participants correctly disagreed with the AI when shown the example-based explanation.

While Figure 5 explicitly shows how many participants disagree and agree with the AI for each scenario, Figure 6 shows the average appropriate trust score across participants for each explanation type when the AI is correct and incorrect. A participant's appropriate

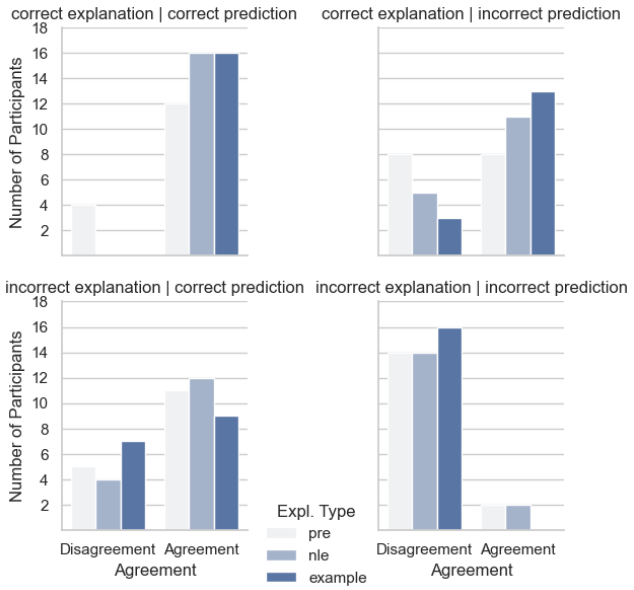


Figure 5: The x-axis shows the agreement variable (disagreement or agreement), and the y-axis shows the number of participants. Each facet is labeled and shows the scenario. The legend shows the type of explanation where 'pre' is no explanation

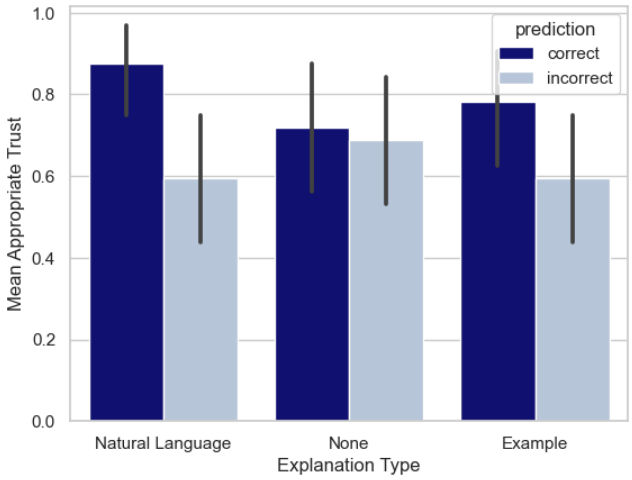


Figure 6: Mean appropriate trust for each explanation type when the AI's classification is correct and incorrect.

trust score is the sum of the number of times they correctly agree or disagree with the AI divided by the total number of images. For example, if a participant agrees with the AI both times it is correct and agrees with the AI both times it is incorrect, then this participant's appropriate trust score is 0.5 or 50%. To determine if the trends we see in Figure 6 are statistically significant, we conduct a two-way repeated measures ANOVA test where our dependent variable is the appropriate trust and our within-subject variables are the explanation type, and the prediction is correct or incorrect. We don't see a significant interaction between the explanation type and the prediction. We also do not see a statistically significant main effect for the explanation type. However, we see a weak significant main effect for the prediction variable on the appropriate trust (pp-value = 0.056). We can conclude that the appropriate trust is higher when the AI is correct than when it is incorrect.

Assertive vs. Non-assertive. When breaking down the natural language explanations into assertive and non-assertive explanations, we can again look at the participant's appropriate trust scores. As shown in Figure 7, we see a higher mean appropriate trust for the assertive natural language explanations when the AI is correct and lower when it is incorrect. The mean appropriate trust for non-assertive, incorrect is slightly higher than assertive, incorrect. However, the mean appropriate trust for non-assertive, correct is slightly higher than non-assertive correct.

Unfortunately, due to a slight counting error in the study system, we have two extra participants for the assertive group than the non-assertive group. This error will not allow us to properly conduct the appropriate statistical tests (repeated measures ANOVA). We can however, perform a different statistical test to see if the difference between the appropriate trust for assertive and non-assertive explanations when the AI is correct is significant. Conducting an independent t-test, we get a p-value of 0.03, so we can conclude that assertive explanations when the AI is correct result in a higher appropriate trust compared to non-assertive explanations when the AI is correct.

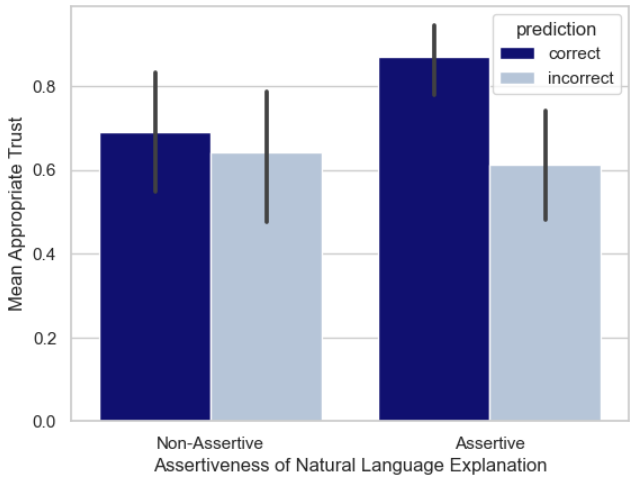


Figure 7: Mean appropriate trust for the non-assertive and assertive natural language explanations when the AI's classification is correct and incorrect.

Explanation preference. In the post-task survey, I asked every participant which explanation they preferred to have if they collaborated with the AI on bird species classification. 12 of the 16 participants answered this question. It is unclear why four of the participants did not answer this question.

7 of the 12 participants said they preferred the example-based explanations; 4 participants mentioned that they would prefer to see both of the explanations at the same time; and one participant said they would prefer the natural language explanations.

5 PRELIMINARY DISCUSSION

5.1 No significant difference between visual and natural language explanations

While we do see differences across the appropriate trust in Figure 6, we did not find a significant difference. This could mean that for this task with non-experts, the modality of explanation did not make any significant impact on decision-making. More data should be collected to have a more powerful analysis and conclusion.

Subjectively prefer visual explanations While the type of explanation does not significantly impact decision-making quantitatively, we see qualitatively through the survey results that most non-experts preferred the visual explanation over the natural language explanation. Further studies should conduct more rigorous tests before implementing visual explanations over natural language explanations in a tool.

5.2 Non-assertive explanations negatively impact appropriate trust when AI is correct

We saw that participants who were assigned to non-assertive natural language explanations had a lower appropriate trust score than the participants who were assigned to assertive natural language explanations with statistical significance. This finding aligns with findings from Pearsall and Ellis [10] where they found that critical team members that were assertive improved team performance and satisfaction. This is an interesting finding for the HCI and CSCW communities because it is important to understand when to include which type of explanations. For example, one design guideline to derive from this finding might be to show assertive explanations when the AI's confidence is above a certain threshold when it is correct instead of non-assertive explanations.

6 LIMITATIONS & NEXT STEPS

This study was treated as a pilot study to help me iterate on the questions that I want to ask, the final study design that I should use to answer questions like this, and in what ways I can ensure the analyses are rigorous enough. Throughout the course of designing the study and collecting pilot study data, I have kept track of everything I want to change with the design of this study which I will elaborate on below:

6.1 Study Design

One explanation type that I did not offer in this study, but considered was showing a combination of the natural language explanation and the example-based explanation or the natural language explanation and a saliency map. If I can implement this into the study design, I would add this as another group to the between-subjects explanation type variable.

Currently, the only between-subjects part of this study is whether the participant is assigned to a non-assertive natural language explanation or an assertive natural language explanation. The type of explanation (visual or natural language) or no explanation was a within-subject variable. While this allowed me to ask each participant their preference on explanation technique, I think this impacted the quantitative results I got for the visual and natural language explanations (despite using counterbalancing). I will change the study design to make the explanation type (visual, natural language, combination of visual and natural language, or none as the control) between subjects so I can perform more rigorous statistical analyses on the data.

Participants are initially asked to guess the bird species without assistance from the AI, along with their confidence in their answer, as one way to gauge their level of expertise and see if they already knew the bird species before working with the AI. When shown the AI classification and explanations, the participants are not allowed to select another bird species if they disagree with the AI's classification. In the future design of this study, I will allow participants to select a new species if they disagree with the AI and change their minds about their original answer.

I only used four different bird images (one for each scenario that I identified). I think the use of one bird image for each scenario significantly impacted my findings because some birds could have been harder in general for people to classify. Moving forward, I will ensure that there are at least four different birds for each scenario so I can try to produce results that are not reliant on how hard it is to classify a given bird.

6.2 Evaluation Metrics

I evaluated how confident users were in their decision to agree or disagree with the AI. While this metric was interesting and did

produce almost statistically significant findings, I would like to change this question to gather a subjective rating on trust in the AI's classification, not confidence in your own decision. Collecting this subjective rating of trust would be based on a 5-point Likert scale and would reveal quantitatively how much trust the user thinks they have in the model.

6.3 Visual Explanations

Instead of using the Python library DeepImageSearch, it would be more appropriate to perform the similarity search based on the extracted features for each image since that is based on how the model views the images. This will allow for the possibility of showing misleading example-based explanations as well.

6.4 Participant Pool

For this class project, I only recruited people from Facebook, personal contacts, and the HCII Ph.D. slack channel. Moving forward, I will be recruiting from the AI for Conservation slack channel, which has over 1,000 users. I will also make a post on the Wildlabs.net community forum, Climate Change AI community forum, Twitter, Prolific, and my connections at WildMe which is an NGO to help ecologists collaborate with AI to detect and identify species from camera traps.

7 CONCLUSION

Citizen scientists and experts have been collaborating with AI for wildlife conservation efforts, but these tools are not being designed to take into account the different perceptions and knowledge that the user has. For example, researchers are more aware of features that should be considered when reviewing an AI's prediction, while citizen scientists are not experts and may not be aware of this causing a negative interaction with the AI. We study two forms of explanations (natural language explanations and example-based explanations) within human-AI collaboration for bird species classification and assertive and non-assertive natural language explanations. While we found that example-based versus natural language explanations did not significantly impact the non-experts' appropriate trust, we saw that non-experts have a higher appropriate trust when shown assertive natural language explanations when the AI is correct compared to non-assertive ones. These findings beg several research questions and challenge designers to ideate new human-centered explanations and interfaces for human-AI collaborations.

ACKNOWLEDGMENTS

I would like to thank Hao-Fei Cheng, Haiyi Zhu, and Adam Perer for helping me iterate through study designs and research questions. I would also like to thank all the participants who completed the pilot study phase. Icons for various figures from Flaticon.

REFERENCES

[1] Tae Hyun Baek, Sukki Yoon, and Seeun Kim. 2015. When environmental messages should be assertive: Examining the moderating role of effort investment. *International Journal of Advertising* 34, 1 (2015), 135–157.

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

[3] 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.

[4] Terri A Fredrick. 2008. Facilitating better teamwork: Analyzing the challenges and strategies of classroom-based collaboration. *Business Communication Quarterly* 71, 4 (2008), 439–455.

[5] 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.

[6] Megan M Lambertz-Berndt and Michael G Blight. 2016. "You Don't Have to Like Me, But You Have to Respect Me" The Impacts of Assertiveness, Cooperativeness, and Group Satisfaction in Collaborative Assignments. *Business and Professional Communication Quarterly* 79, 2 (2016), 180–199.

[7] X Phong Nguyen, Tho H Tran, Nguyen B Pham, Dung N Do, and Takehisa Yairi. 2022. Human Language Explanation for a Decision Making Agent via Automated Rationale Generation. *IEEE Access* 10 (2022), 110727–110741.

[8] Jason Parham, Charles Stewart, Jonathan Crall, Daniel Rubenstein, Jason Holmberg, and Tanya Berger-Wolf. 2018. An animal detection pipeline for identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1075–1083.

[9] 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.

[10] Matthew J Pearsall and Aleksander PJ Ellis. 2006. The effects of critical team member assertiveness on team performance and satisfaction. *Journal of Management* 32, 4 (2006), 575–594.

[11] 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.

[12] 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.

[13] 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.

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

A APPENDIX

A.1 Visual and Natural Language Explanations

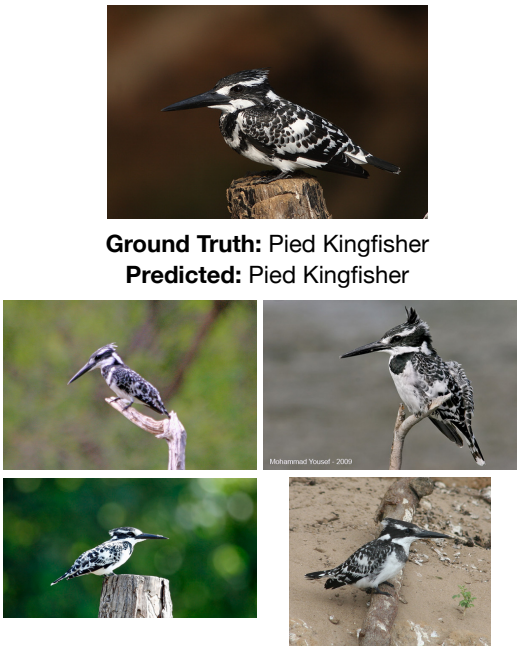
We show the visual and natural language explanations that we used for this study along side each image that was used along with the ground truth and predicted classes.

A.2 Survey Questions

The first question of the survey is, "How experienced are you as a bird watcher?". A 5-point Likert scale was provided with this question showing the following options: novice, advanced beginner, competent, proficient, expert.

The second question of the survey is, "How often do you identify bird species (as a hobby or for occupation)?". A 5-point Likert scale was provided with the following options: several times a week, a couple times a week, a couple times a month, a couple times a year, never.

The third survey question is open-ended and asked people to write down their occupation. The fourth question asked people to describe what channel of communication they heard about this study from. This question was more for me personally to see what populations I was getting responses from, whether that be Facebook friends, people I reached out personally, or various slack channels.



Ground Truth: Pied Kingfisher
Predicted: Pied Kingfisher

Natural Language Explanation: this is a black and white bird with a long black bill

Figure 9: This shows one of the four bird images that were used in the study.

Ground Truth: Ringed Kingfisher
Predicted: Ringed Kingfisher

Natural Language Explanation: this is a blue bird with a white belly and a brown throat

Ground Truth: Hooded Warbler
Predicted: Wilson Kingfisher

Natural Language Explanation: this is a yellow bird with a grey wing and a small pointy beak

Figure 10: This shows one of the four bird images that were used in the study.

Ground Truth: Red winged blackbird
Predicted: Rose breasted Grosbeak

Natural Language Explanation: this is a black bird with red breast and a red spot on its wing

Figure 11: This shows one of the four bird images that were used in the study.

The next survey question asked for participants to state their preference regarding the explanations used in the study: "Which explanation technique (natural language sentences or pictures of examples) would you prefer to see when collaborating with AI to identify birds and why? Response must be a minimum of 50 characters.".

The last two questions ask people to make note of any technical errors or difficulties they encountered during the study and if they had any other comments to share about the study.