

# Supplementary Material for "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction

SUNNIE S. Y. KIM, Princeton University, USA

ELIZABETH ANNE WATKINS, Intel Labs, USA

OLGA RUSSAKOVSKY, Princeton University, USA

RUTH FONG, Princeton University, USA

ANDRÉS MONROY-HERNÁNDEZ, Princeton University, USA

## A INTERVIEW PROTOCOL

We conducted our interviews based on the following questions.

### Part 1: Context

*Domain and AI Background.*

- (1) How would you describe your knowledge of birds?
- (2) How would you describe your knowledge of machine learning and artificial intelligence?

*Use of app.*

- (1) Which features of Merlin do you use among Bird ID, Photo ID, Sound ID, Explore Birds? Why do you not use features XYZ?
- (2) For what tasks do you use the app?
- (3) How successful are you at accomplishing those tasks?
- (4) In what scenarios or circumstances do you decide to use the app?

*Stakes in use.*

- (1) What do you gain when Merlin is successful? What do you lose when Merlin is unsuccessful?
- (2) How important is it to you that Merlin gets each and every prediction correct?

### Part 2: XAI needs and more

As you may know, Merlin uses machine learning-based AI models to identify birds in photos and audio recordings. We will now ask questions about your experiences and thoughts on Merlin's AI models.

*Knowledge and perception of AI.*

- (1) What do you know about Merlin's AI?
- (2) How accurate do you think Merlin's bird identification is?
- (3) How well did you expect Merlin to work? How well did it actually work?
- (4) How do you know if Merlin is correct or incorrect?
- (5) Do you know when Merlin works well and not? For example, have you noticed that it works better on certain types of inputs or certain bird species?

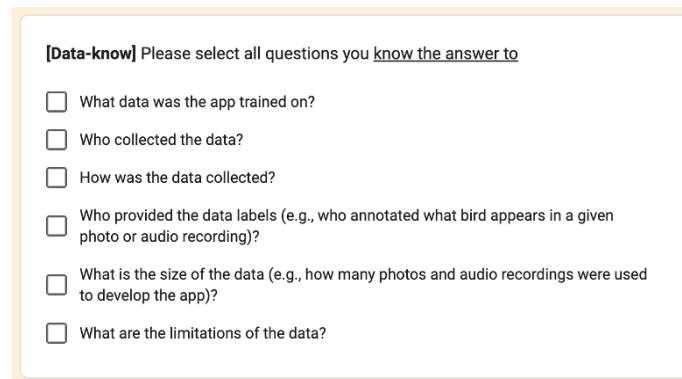
*Trust.* We will now present two scenarios to you and ask whether you would use Merlin in them.

- (1) Scenario 1: Suppose you find a sick bird and take it to the vet. The vet is not sure what bird it is. Would you recommend Merlin to identify the bird species so that the vet can determine the course of treatment?
- (2) Scenario 2: Suppose you are entering a game show where you can win or lose money based on how well you can identify birds from photos or audio recordings. You can only use one resource among Merlin, books (e.g., field guides), the Internet (e.g., search engine, online birder community), and so on. Which resource would you use? Does your answer change depending on certain factors?

*XAI needs (open-ended questions).*

- (1) Could you reflect on a time when Merlin didn't work as you expected and/or the last time you used Merlin?
- (2) During this time what questions did you have? Did you want any explanations of why or how Merlin made its identification?
- (3) In general, what more do you wish to know about Merlin's AI?

*XAI needs (survey).* We will now direct you to a survey. Please take a few minutes to fill it out. When you're done, let us know and we will ask a few follow-up questions about your responses. [Share survey link and wait for completion.]



**[Data-know]** Please select all questions you know the answer to

- ☐ What data was the app trained on?
- ☐ Who collected the data?
- ☐ How was the data collected?
- ☐ Who provided the data labels (e.g., who annotated what bird appears in a given photo or audio recording)?
- ☐ What is the size of the data (e.g., how many photos and audio recordings were used to develop the app)?
- ☐ What are the limitations of the data?

Fig. 1. Snapshot of the survey. For each of the 10 question categories (e.g., *Data*), we ask participants to select all questions in the category they “know the answer to” in one block (as shown in the figure) and “curious to know (more)” in another block. See Sec. 4.2 in the main paper for details about the survey and Tab. 1 for the full survey questions and participants’ responses.

Thank you for filling out the survey.

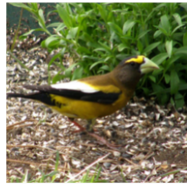
- (1) How do you know the answers to questions XYZ?
- (2) What made you select some questions and not others? Could you explain your thought process?
- (3) How much would you go out of your way to get the answers to these questions?

### Part 3: XAI perceptions

Finally, we will show different options for explaining AI’s outputs and ask for your thoughts. [Start showing slides.]

### Participant's explanation

Before we show options for explaining Merlin's identifications, we want to first ask how you explain your bird identification. Here are three example photos. How would you explain your identification to others?



Evening Grosbeak



Marsh Wren



Airplane

### Perceptions of existing XAI approaches

We uploaded these photos to Merlin Photo ID to see what it returns. For the first photo, Merlin correctly identified Evening Grosbeak. For the second photo, Merlin misidentified Marsh Wren as House Wren. For the third photo, Merlin misidentified Airplane as Ruby-throated Hummingbird.

We will now show you four different approaches for explaining these identifications. These are not actual explanations of how Merlin makes identifications, but mock-ups we created to get your opinions on which explanations would be good to potentially implement in Merlin. [Introduce XAI approaches, one at a time and in random order.<sup>1</sup>]

*After introducing an approach, ask:*

- (1) Do you think you have a better understanding of how Merlin's AI model makes its identification?
- (2) Would you like to see this explanation in Merlin?
- (3) What do you like and dislike about this explanation?
- (4) What can be improved about this explanation?
- (5) Is this explanation satisfying?
- (6) How would you use this explanation?

*After going through all approaches, ask:*

- (1) Which explanation was your favorite and least favorite? Why?
- (2) Having seen different explanation options, would you change how you explain your own bird identification?

*Closing.* Is there anything that you want the research team to know that we haven't been able to cover yet?

## B MATERIALS FOR STUDYING XAI PERCEPTIONS

In this section we describe how we created the materials used in Part 3 of the interview.

### Selecting examples

Merlin's results can be divided into three categories: (1) correct identification, (2) misidentification that people, even experienced birders, would make, (3) misidentification that people wouldn't make. We decided to show an example for each so that participants think about explanations in context of both successful and unsuccessful identifications.

<sup>1</sup>See Fig. 2 in the main paper for details on how we introduced each approach to participants.

For (1) we looked for birds with salient features. We decided on Evening Grosbeak because it has distinctive color and beak size. For (2) we looked for birds that are difficult for human birders to identify. We decided on Marsh Wren because it is one of “little brown birds” that are known to be notoriously difficult to distinguish. For (3) we looked for an object that is not a bird but can potentially fool Merlin. We decided on Airplane because it has a similar shape to birds, although people wouldn’t mistake it for birds.

For each, we selected a few candidate photos from the CUB dataset [6] and the Internet. We then inputted them to Merlin Photo ID, and based on the identification results, chose (1) an Evening Grosbeak photo that Merlin correctly identifies, (2) a Marsh Wren photo that Merlin misidentifies as House Wren, and (3) an Airplane photo that Merlin misidentifies as Ruby-throated Hummingbird. The misidentification in (3) may be due to classification models typically being unable to reject an example that doesn’t belong to a pre-defined set of classes. That is, Merlin Photo ID may not have a “not bird” option and always try to output a bird species even if the input photo does not contain a bird.

### Mocking up XAI explanations

Since we didn’t have access to Merlin’s AI models, it was not possible to produce actual explanations of how Merlin identifies birds. Hence, we created mock-ups of representative XAI approaches in the following way.

*Heatmap.* We created our heatmap-based explanations by training a bird image classification model and generating GradCAM [5] heatmaps for the example photos. Concretely, we trained a standard ResNet50-based model on the CUB dataset [6] that achieves 81.0% accuracy in 200 birds classification. See Fig. 2 for a comparison of explanations from an existing method (GradCAM [5]) vs. our mock-up.

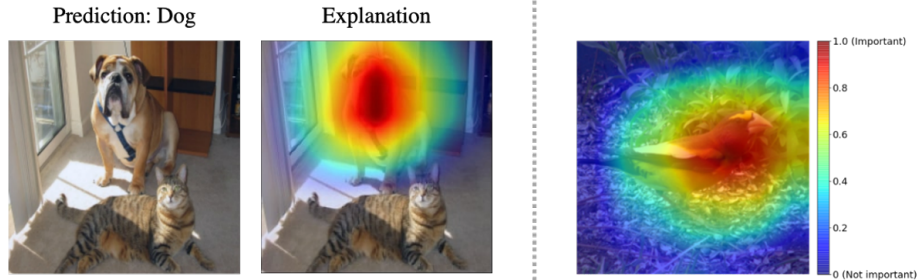


Fig. 2. Heatmap-based explanations. (Left) Example explanation from Selvaraju and colleagues’ work [5]. (Right) Our mock-up.

*Example.* For example-based explanations, we looked for photos from the identified bird species that looked similar to the input photo in the CUB dataset [6] and the Internet. We then showed three photos in each explanation mock-up. See Fig. 3 for a comparison of explanations from an existing method (Representer Point Selection [7]) vs. our mock-up.

*Concept.* For concept-based explanations, we used attributes in the CUB dataset [6] as concepts, following prior work [2]. For each example photo, we manually selected concepts and coefficients, and calculated a class score based on whether the selected concepts were present or absent in the input photo. See Fig. 4 for a comparison of explanations from an existing method (example from [4]) vs. our mock-up. To improve readability, in our mock-ups, we limited the number of concepts, presented concepts vertically with one concept in each row, highlighted positive and negative concept coefficients in different color, and crossed out concepts that were absent.

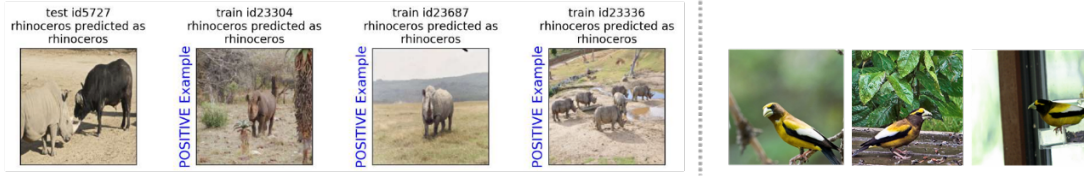


Fig. 3. Example-based explanations. (Left) Example explanation from Yeh, Kim and colleagues' work [7]. (Right) Our mock-up.

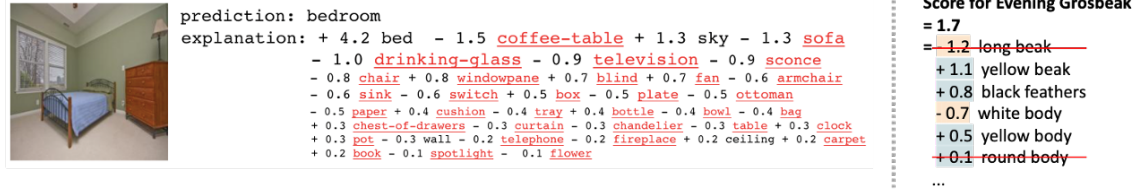


Fig. 4. Concept-based explanations. (Left) Example explanation from Ramaswamy and colleagues' work [4]. (Right) Our mock-up.

*Prototype.* For prototype-based explanations, we identified representative bird parts (e.g., wing, beak, body) and manually selected prototypes, matching photo regions, and similarity scores. See Fig. 5 for a comparison of explanations from an existing method (ProtoPNet [1]) vs. our mock-up. In our mock-ups, we reduced the explanation complexity by showing the input photo once with all prototype-photo region matches overlaid on top, and removing all technical details except similarity scores for each match.

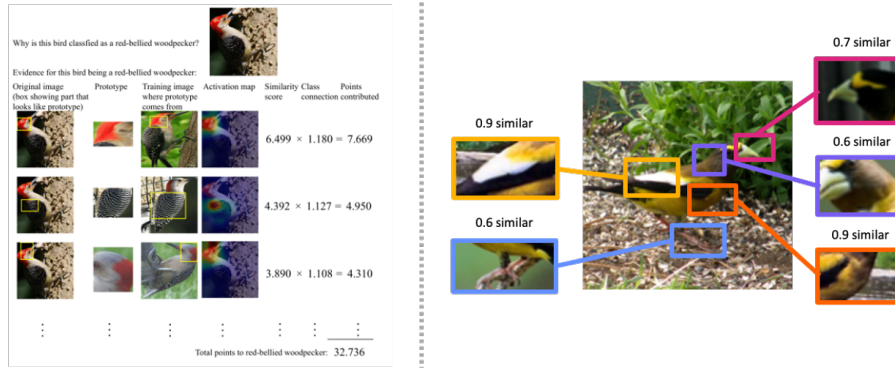


Fig. 5. Prototype-based explanations. (Left) Example explanation from Chen, Li, and colleagues' work [1]. The full explanation consists of up to 10 rows. (Right) Our mock-up.

## C SURVEY QUESTIONS AND RESULTS

Finally, we provide the full survey questions and results in Tab. 1. See Sec. 4.2 in the main paper for details about the survey. Overall, participants were curious about the listed questions. For each, most participants selected “know the answer to” or “curious to know (more).” Concretely, out of 19 participants who filled out the survey,<sup>2</sup> only 0–4

<sup>2</sup>We had 20 participants in total. One participant did not fill out the survey due to time constraints.

participants selected neither (don't know and not curious) for any question. These participants tended to have low-AI background and lower explainability needs as described in Sec. 5.1 in the main paper.

The five questions with the highest number of “know the answer to” selections were:

- Q1. *What kind of output does the app give?*
- Q2. *For Sound ID, what is the spectrogram showing?*
- Q3. *What does the output mean?*
- Q4. *Why is this instance predicted X?*
- Q5. *In what situations is the app likely to be correct?*

Q1, Q2, Q3 indicate that participants have a good understanding of the AI's output. Q4 suggests that participants find the AI's output understandable. Q5 suggests that participants are aware when the AI is likely to be correct. These selections were consistent with participants' experiences with the AI application. Overall, participants were knowledgeable about the app and aware of when it works well and not.

The five questions with the highest number of “curious to know (more)” selections were:

- Q1. *Who collected the data?*
- Q2. *What is the size of the data?*
- Q3. *How was the data collected?*
- Q4. *For Photo ID, when does the app return one vs. multiple birds?*
- Q5. *What features does the app consider to make predictions?*

Q1, Q2, Q3 illustrate that participants are curious about the data that powers the AI. Q4 indicates that participants are not sure why or when different types of outputs are given. Q5 suggests that participants want to know the specific features the AI uses to make its identification, as described in Sec. 5 of the main paper. We note that Q1, Q2, Q3, Q4 convey transparency needs which can be met relatively easily by providing more documentation. Q5 conveys an explainability need, one that many XAI methods are developed to meet.

Table 1. Summary of 19 participants’ responses to the XAI needs survey we developed from Liao and colleagues’ XAI Question Bank [3]. For each question, we report the number of participants who selected “**Know** the answer to,” who selected “**Curious** to know (more),” and who selected neither (don’t know and **Not curious**). The numbers in each row don’t always add up to 19 because some participants selected both “**Know** the answer to” and “**Curious** to know (more)” for a question. See Sec. 4.2 in the main paper for details about the survey. See Sec. 5.1 in the main paper and Sec. C for discussions of the results.

Question about AI	Know	Curious	Not curious
<b>Data</b>			
What data was the app trained on?	8	13	1
Who collected the data?	5	17	0
How was the data collected?	1	16	2
Who provided the data labels?	4	11	4
What is the size of the data?	2	17	0
What are the limitations of the data?	5	13	1
<b>Output</b>			
What kind of output does the app give?	15	3	1
What does the output mean?	12	5	2
For Photo ID, when does the app return one vs. multiple birds?	4	15	1
For Sound ID, what is the spectrogram showing?	13	5	1
<b>Performance</b>			
How accurate is the app’s prediction?	5	14	1
How often does the app make mistakes?	5	13	1
In what situations is the app likely to be correct?	10	9	1
In what situations is the app likely to be incorrect?	9	11	0
<b>How</b>			
How does the app make predictions?	6	13	0
What features does the app consider to make predictions?	4	15	0
<b>Transparency</b>			
How do others use the app?	5	11	3
Do domain experts use the app?	6	12	3
Did domain experts help develop the app?	9	8	4
<b>Why</b>			
Why is this instance predicted X?	12	7	1
Why are instances A and B given the same prediction?	9	7	3
<b>Why not</b>			
Why is this instance NOT predicted Y?	9	8	3
Why is this instance predicted X instead of Y?	8	7	4
Why are instances A and B given different predictions?	8	10	2
<b>What if</b>			
What would the app predict if this instance is changed in some way?	5	11	3
What would the app predict for a different instance?	3	14	2
<b>How to be that</b>			
How should this instance change to get a different prediction?	5	12	2
What kind of instance gets a different prediction?	6	11	2
<b>How to still be this</b>			
What is the scope of change permitted to still get the same prediction?	2	14	3
What kind of instance gets this prediction?	4	12	3

## REFERENCES

- [1] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Neural Information Processing Systems (NeurIPS)*.
- [2] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *International Conference on Machine Learning (ICML)*.
- [3] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. *Questioning the AI: Informing Design Practices for Explainable AI User Experiences*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3313831.3376590>
- [4] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. 2022. Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability. <https://doi.org/10.48550/ARXIV.2207.09615>
- [5] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *International Conference on Computer Vision (ICCV)*.
- [6] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. *The Caltech-UCSD Birds-200-2011 dataset*. Technical Report CNS-TR-2011-001. California Institute of Technology.
- [7] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. 2018. Representer Point Selection for Explaining Deep Neural Networks. In *Neural Information Processing Systems (NeurIPS)*.