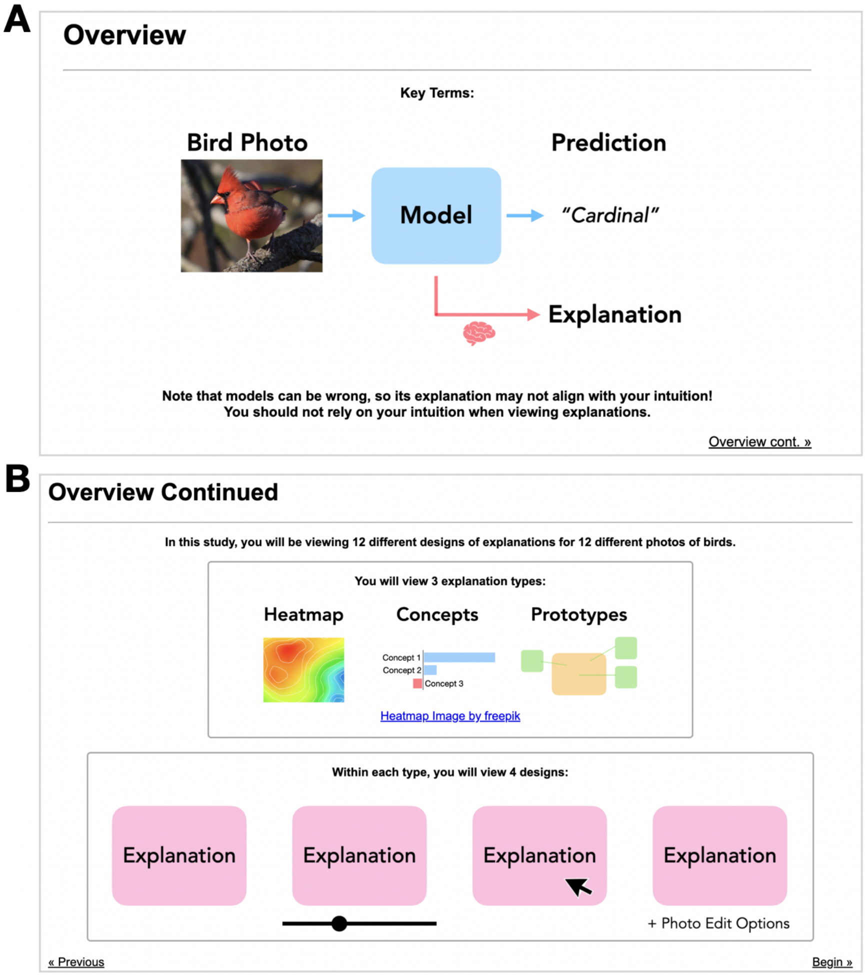
SUPPLEMENTARY MATERIALS

C STUDY INTERFACE

We built a Flask application, deployed using Render, to implement our mock-ups and help guide participants through the study. In line with our verbal script, participants began by viewing the overview pages shown in Fig. 1 to obtain a general idea of what an explanation is and the structure of the study. Then, for the first explanation type that they were assigned, they progressed to a page that briefly explained how to interpret the explanation type accompanied by a generic figure of the explanation type (Fig. 2). Next, for each presentation type, participants viewed and, when applicable, interacted with the explanation (Fig. 4). They then answered the task questions as described in the main paper using a Google Form. Participants were also provided with a bird part guide to help with identifying bird parts; the image and annotations for the bird part guide were from the Caltech-UCSD Birds-200-2011 dataset[A2]. After viewing the four presentation types for the current explanation type, participants viewed a page that displayed snapshots of the four explanations that they just interacted with (Fig. 3) and provided ratings and rankings for these explanations. They then repeated this process for the two remaining explanation types. Finally, at the end of the study when participants finished working through all twelve explanations, they viewed a table of snapshots of each explanation, organized by explanation type along the rows and by presentation type along the columns (Fig. 5). Similar to the explanation type summary pages, participants provided ratings and rankings on the presentation types, this time generalizing along each column.



Key Terms:

Note that models can be wrong, so its explanation may not align with your intuition!
You should not rely on your intuition when viewing explanations.

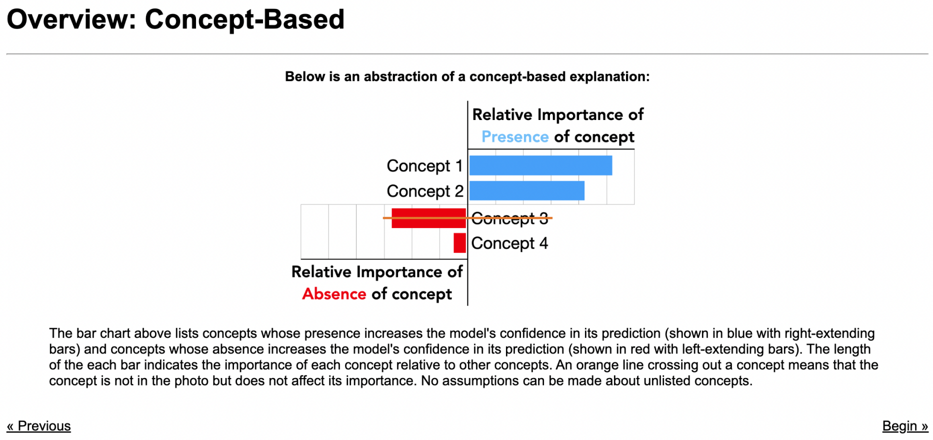
In this study, you will be viewing 12 different designs of explanations for 12 different photos of birds.

You will view 3 explanation types:

Within each type, you will view 4 designs:

Figure 1: Overview Pages. (A) Page that provides a simple overview of what an explanation is. (B) Page that provides an outline of what explanation and presentation (referred to as "design" during the study) types the participant was about to encounter; we made abstract representations of the explanation and presentation types[1].

[1] Heatmap image attributed in the Panel B was designed by Freepik.



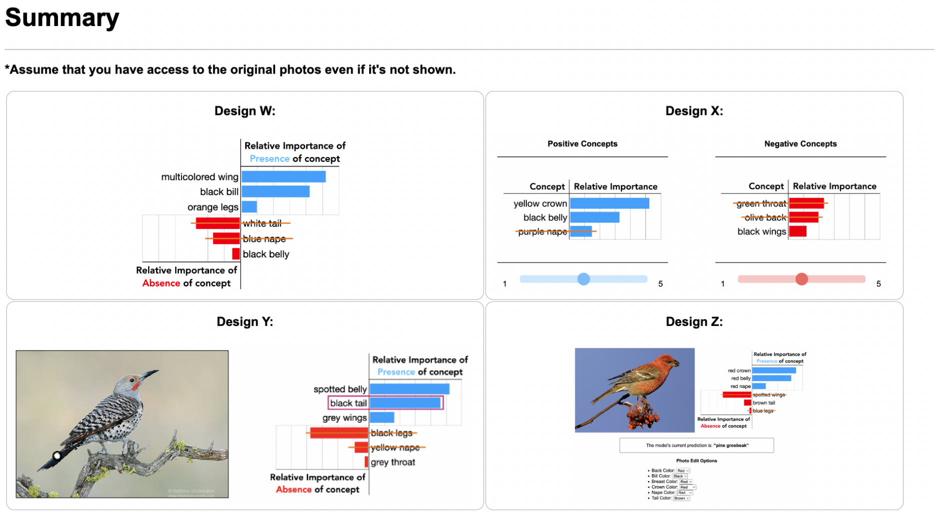
Overview: Concept-Based

Below is an abstraction of a concept-based explanation:

Relative Importance of Presence of concept
Concept 1
Concept 2
~~Concept 3~~
Concept 4
Relative Importance of Absence of concept

The bar chart above lists concepts whose presence increases the model's confidence in its prediction (shown in blue with right-extending bars) and concepts whose absence increases the model's confidence in its prediction (shown in red with left-extending bars). The length of the each bar indicates the importance of each concept relative to other concepts. An orange line crossing out a concept means that the concept is not in the photo but does not affect its importance. No assumptions can be made about unlisted concepts.

Figure 2: Explanation Type Overview Page Example. This is an example of an overview page for an explanation type, in this case concept-based explanations, which was shown before the participant encountered the four presentation types. An abstract figure of the explanation type is shown with a short description of how to interpret its components underneath.



Summary

*Assume that you have access to the original photos even if it's not shown.

The model's current prediction is: **"pine grosbeak"**

Photo Edit Options

- Back Color: Red
- Bill Color: Black
- Breast Color: Red
- Crown Color: Red
- Nape Color: Red
- Tail Color: Brown

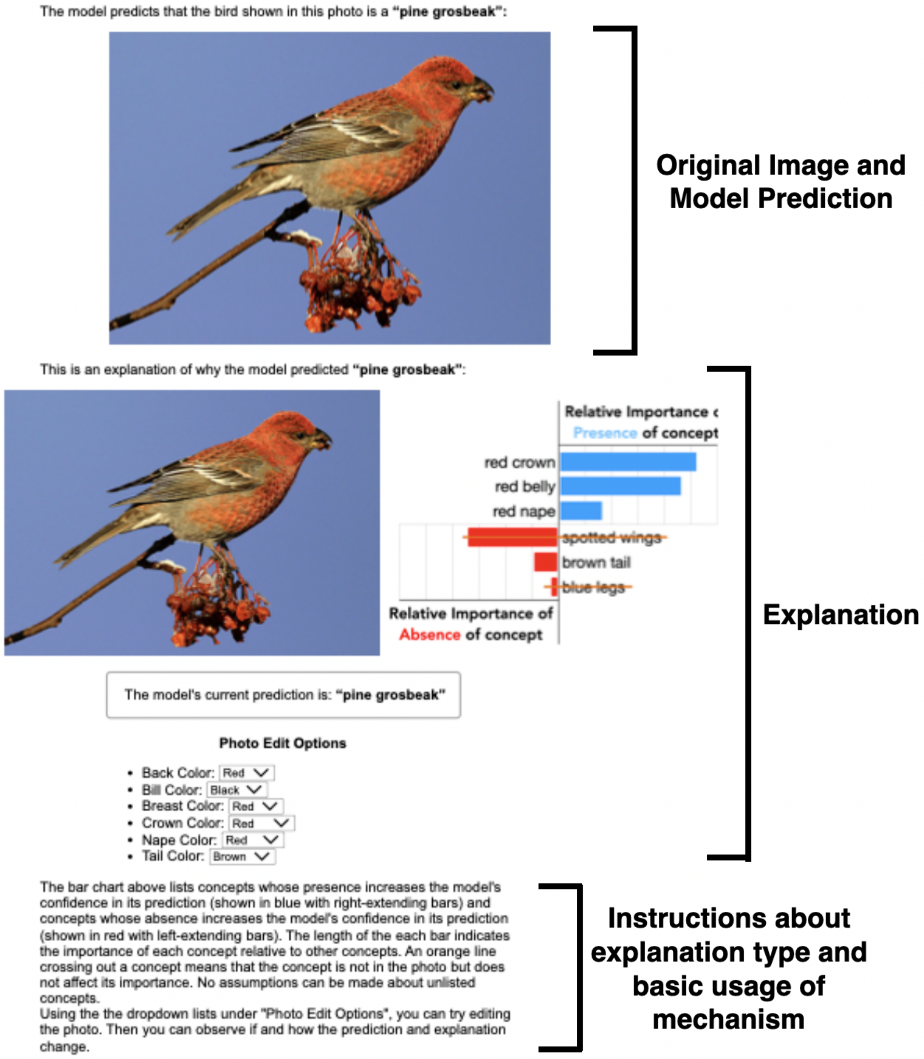
Figure 3: Explanation Type Summary Page Example. This is an example of a summary page for an explanation type, in this case for concept-based explanations, that was shown after the participant finished viewing the four presentation types for the explanation type. The four presentation types were given the generic labels W-Z (for Static, Filtering, Overlays, and Counterfactuals respectively) for the participant to rate and rank in the survey.

D CONSTRUCTING MOCK-UPS

All bird images were from the Caltech-UCSD Birds-200-2011 dataset[A2]. We began with the set of images that the model had not been trained on (i.e., the test set) and correctly classifies. We then perform the following preprocessing steps:

(1) Removed categories that directly had concept names (e.g., Yellow-billed Cuckoo) or hinted at the concepts (e.g., Sooty Albatross).
(2) Filtered by model confidence, keeping images that had scores > 0.99.
(3) Manually screened the images to remove images that did not clearly display one real bird in the main center focus (e.g., images that were drawings rather than real bird photos, images that cutoff part of the bird, images that contained more than one bird, etc).

After these preprocessing steps, we consulted with two birding experts to iterate through a randomly selected subset to construct a set of 12 images of similar difficulty. The decisions during these consultations depended on how difficult the bird species is to identify when in the field and also the quality of the images. As described in the main text, we created mock-up explanations for the 12 images that resulted from



The model predicts that the bird shown in this photo is a **"pine grosbeak"**:

This is an explanation of why the model predicted **"pine grosbeak"**:

The model's current prediction is: **"pine grosbeak"**

Photo Edit Options

- Back Color: Red
- Bill Color: Black
- Breast Color: Red
- Crown Color: Red
- Nape Color: Red
- Tail Color: Brown

The bar chart above lists concepts whose presence increases the model's confidence in its prediction (shown in blue with right-extending bars) and concepts whose absence increases the model's confidence in its prediction (shown in red with left-extending bars). The length of the each bar indicates the importance of each concept relative to other concepts. An orange line crossing out a concept means that the concept is not in the photo but does not affect its importance. No assumptions can be made about unlisted concepts.
Using the the dropdown lists under "Photo Edit Options", you can try editing the photo. Then you can observe if and how the prediction and explanation change.

Figure 4: Explanation Page Example. This is an example of a page for an explanation. From top to bottom: The original image along with the model's prediction on that image is shown. Then, the bottom half depicts the explanation and a paragraph that repeats the description of the explanation type (Fig. 2) and includes instructions for basic usage of the mechanism, if present.

this process, and the designs of our mock-ups were grounded in prior work. We generating heatmap mock-ups by using Grad-CAM on a ResNet-18 model [A1] that had been trained on the CUB dataset[A2].

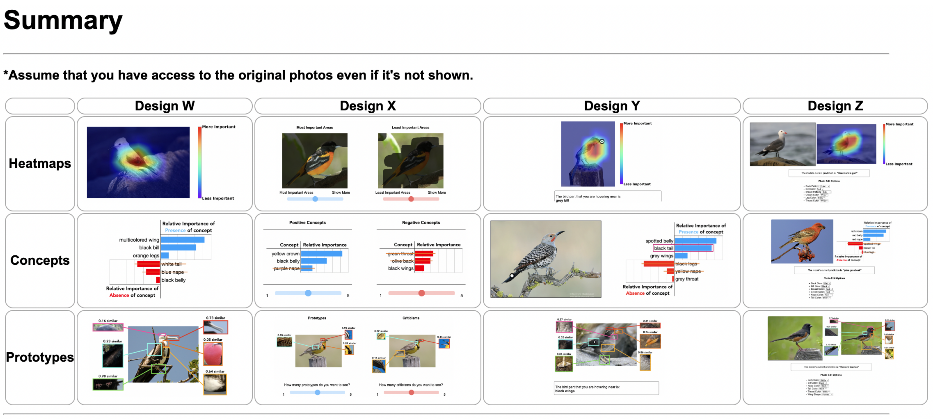
E RECRUITING DETAILS

We advertised our study through various platforms, including several institutions' student group lists, birding labs, birding clubs, the AI for Conservation Slack, the Birding International Discord, the Climate Change AI community forum, the WildLabs.net community forum, X, and Mastodon.

REFERENCES

[A1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[A2] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset.* Technical Report CNS-TR-2011-001. California Institute of Technology.



Summary

*Assume that you have access to the original photos even if it's not shown.

Figure 5: Explanations Summary Page. This is the summary page that was shown at the end of the study (i.e., after the participant finished viewing all twelve explanations). The rows represent the explanation types (i.e., Heatmaps, Concepts, and Prototypes), and the columns represent the four general groups of presentation type (i.e., Static, Filtering, Overlays, and Counterfactuals). The four presentation types were given the generic labels W-Z respectively for the participant to rate and rank in the survey.