How Do You Take Your Ice Cream?
Toward Understanding How Interactivity Can Improve Computer Vision Explanations

VOLUME I

Indu Panigrahi

A MASTER'S THESIS PRESENTED TO THE FACULTY OF
PRINCETON UNIVERSITY IN CANDIDACY FOR THE DEGREE OF
MASTER OF SCIENCE IN ENGINEERING

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF COMPUTER SCIENCE

Advisers: Professor Parastoo Abtahi and Dr. Ruth Fong

May 2025

Abstract

Explanations for computer vision models are important tools for interpreting how the underlying models work. However, they are often presented in static formats, which pose challenges for users, including information overload, a gap between semantic and pixel-level information, and limited opportunities for exploration. We investigate interactivity as a mechanism for tackling these issues in three common explanation types: heatmap-based, concept-based, and prototype-based explanations. We conducted a study (N=24), using a bird identification task, involving participants with diverse technical and domain expertise. We found that while interactivity enhances user control, facilitates rapid convergence to relevant information, and allows users to expand their understanding of the model and explanation, it also introduces new challenges. To address these, we provide design recommendations for interactive computer vision explanations, including carefully selected default views, independent input controls, and constrained output spaces.

Acknowledgements

Firstly, I want to thank my advisers (Parastoo Abtahi and Ruth Fong) and my collaborators (Sunnie S. Y. Kim, Amna Liaqat, Rohan Jinturkar, and Olga Russakovsky) for their frequent guidance and feedback. This was a big project, and I could not have done it without the support of this awesome team. Secondly, my team and I want to thank our participants for taking the time to participate in our study. Additionally, we thank Patrick Newcombe and Fengyi Guo for their help in identifying bird species and images of similar difficulty for the study. Lastly, we thank members of Princeton HCI and the Princeton Visual AI Lab, especially Vikram Ramaswamy, for providing feedback and resources throughout this project.

This material is based upon work supported by the National Science Foundation under grant No. 2145198 (OR) and the Graduate Research Fellowship (SK), as well as the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), the Princeton SEAS Innovation Fund (RF), and Open Philanthropy (RF, OR). Any findings, conclusions, or recommendations expressed in this thesis are those of the author and do not necessarily reflect the views of the sponsors.

This work has been published as part of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems [69], so much of the material is shared between this thesis and the published abstract. However, this thesis provides additional details.

To my parents.

Contents

List of Figures

Chapter 1

Introduction

You can probably tell that Fig. 1.1 is an image of ice cream. What did you pay attention to when making that decision? Was it just the ice cream in the center? Did you look at the bowl and spoon? You could pass the same image into a computer vision model, and it could correctly tell you that it is ice cream. However, just from the model's decision, you do not know what information it used to make that decision.

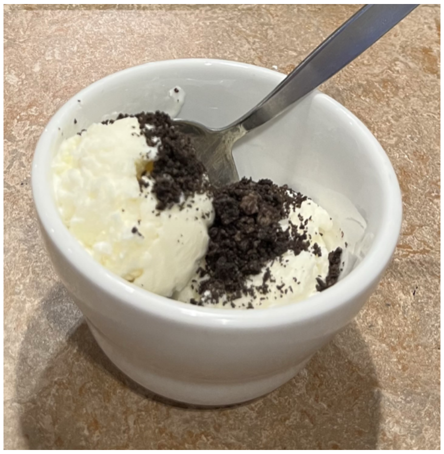


Figure 1.1: **Food for Thought.** Ice cream from the Forbes Dining Hall.

AI explanations help users understand and diagnose complex AI models, such as modern computer vision (CV) models [1, 6]. While the explainable AI (XAI) community has developed many explanation methods, they often present information in a static form which can be challenging for users to interpret [49, 94, 11, 44, 72, 40]. Moreover, a single static explanation cannot meet the diverse needs of users with varying expertise [51], similar to how a single flavor of ice cream cannot satisfy everyone.

Most XAI methods produce static CV explanations [76, 74, 70, 28, 15, 80, 77, 96, 98, 47, 73, 79, 100, 17]. However, recent studies have identified several challenges with static explanations [44, 49, 94, 11, 72, 40] which we categorize into three groups:

1. **Information overload**: Users feel that there is too much information to process at once [49, 94, 11, 44, 72, 40].

2. **Semantic-pixel gap**: Users find it difficult to connect image pixels to objects, attributes, and abstract concepts [72, 44, 11].

3. **Limited exploration**: Users lack the necessary tools to probe the explanation to deepen their understanding [44].

Prior work in Information Visualization has shown that interacting with data helps with interpretation [38, 46, 95, 32], and consequently, several works have called for interactivity in XAI applications [50, 86, 87, 67, 83, 32]. However, interactivity has mostly been explored at the dataset level [95, 12, 32, 27, 8, 92, 23], for visualizing model weights [93, 16, 37, 33, 23], and for non-visual data [52, 83, 9, 10]. In this work, we explore interactivity *at the explanation level* for images. Specifically, we investigate two research questions:

- **RQ1:** How do end-users **leverage interactivity** to help understand information conveyed by computer vision explanations?

- **RQ2:** How do end-users **perceive interactive computer vision explanations**?

To answer these questions, we conducted a within-subjects study using explanations for a bird identification model [91]. We selected bird identification over common object or scene classification use cases [99, 21] to examine the effects of domain knowledge. We focused on three widely used explanation types (heatmap-based, concept-based, and prototype-based explanations) and investigated three interactive mechanisms: **Filtering** to control the amount of information, **Overlays** to connect pixel-level landmarks to semantic labels, and **Counterfactuals** to allow users to edit images and observe a change in the explanation. You can think of the static explanation types as plain vanilla cones and the interactive mechanisms as different types of chocolate-based mix-ins. We recruited 24 participants with varying levels of machine learning (ML) and birding expertise for an $\sim$ 90-minute study. Each participant completed a set of tasks, alongside a questionnaire and interview.

We found that interactivity addresses some limitations of static CV explanations by providing users with tools to bridge the semantic-pixel gap, rapidly pinpoint the information they seek, explore the CV model beyond the immediate task, and gain clarity around the presentation of the static explanation. However, some interactive mechanisms were at times overwhelming for users and introduced new challenges, which informed the design recommendations we present in this paper. Our exploration and study findings contribute to our understanding of interactive mechanisms for XAI and bring us closer to the design of effective interactive CV explanations.

Chapter 2

Related Work

Given the increasing ubiquity of AI[82, 55, 34, 35, 59], there exists a growing need for end-users to understand a model's behavior so they can appropriately trust and use the model. To address this need, many XAI methods have been developed over the past decade across various subfields of AI (e.g., CV and robotics) [53, 18, 74, 60, 71] and across different domains (e.g., healthcare and climate science) [20, 3, 54, 81, 84, 4, 58, 56, 57]. Historically, much work has focused on *developing methods* to explain a model's outputs (i.e., generating an answer for *"why did the model make this prediction?"*). Recently, research has emerged at the intersection of XAI and Human-Computer Interaction (HCI) on *evaluating how useful* explanation methods are for users with varying machine learning (ML) expertise [64, 36, 66, 42], which we build on for CV explanations. In this chapter, we highlight literature most related to our focus on interactive CV explanations.

2.1 Computer Vision Explanations

In this work, we focus on *attribution-based* explanations, which compute a measure of importance by identifying the parts of an input that are critical for a model's decision [29, 61] and have been the most developed in XAI research thus far. We discuss three types of attribution-based explanations:

- *Heatmap-based explanations* typically generate a temperature heatmap over an input image that indicates the importance of each pixel or image region for the model's prediction [76, 74, 70, 28, 15, 80, 77, 96, 98] (Fig. 2.1A).

- *Concept-based explanations* explain a model's output using semantic concepts (e.g. the presence of a "yellow crown" was important for the model's prediction of a "bobolink" in Fig. 2.1B) [47, 73, 79, 100]. These explanations list a set of concepts that contribute to the model's prediction (e.g., "yellow crown", "pointed wings", "gray bill" in Fig. 2.1B) and assign a numerical importance score to each concept (e.g., +2.7 for "yellow crown" in Fig. 2.1B).

- *Prototype-based explanations* learn a set of important image regions (i.e., prototypes) from training images [17, 75, 39, 65, 22, 68]. Given a new image, they then match its regions to the most similar prototypes from the learned set and report those regions alongside similarity scores as contributing to the model's prediction (Fig. 2.1C).

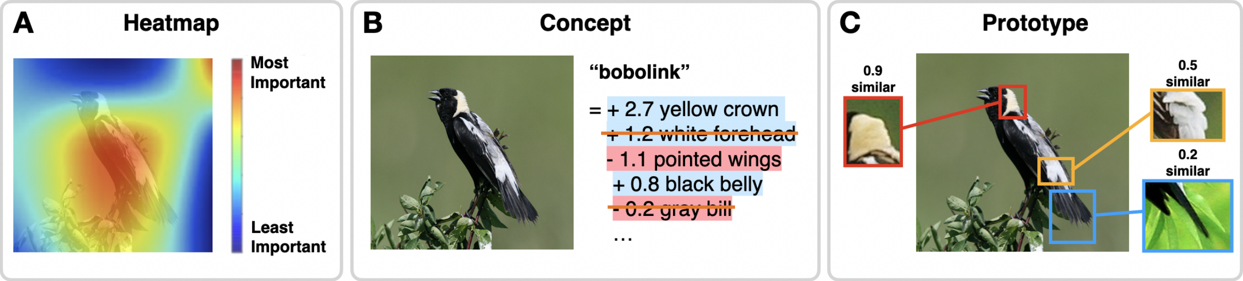


Figure 2.1: **Types of Computer Vision Explanations.** We worked with heatmap-, concept-, and prototype-based explanations, sample mock-ups of which are shown here.

2.2 Problems with Static CV Explanations

Computer vision explanations have primarily been static. However, a number of works in XAI and in Information Visualization have performed user studies that pinpoint various issues with static CV explanations. We group these issues into three overarching problems: 1. information overload, 2. semantic-pixel gap, and 3. limited means for user-driven exploration.

2.2.1 Information Overload

Prior work has found that the amount of information in explanations can be overwhelming. With heatmaps, users often want to view the "extremes" of a heatmap and find intermediate colors distracting [49, 94, 11]. For concepts, users report feeling overwhelmed by the number of items [44, 72], often the number of annotated objects in a dataset (e.g., 1197 in the Broden dataset [7]). Similarly, prototypes can require many regions to explain a model well enough [40].

2.2.2 Semantic-Pixel Gap

Prior work suggests users often find it difficult to connect image regions to the represented object and vice versa. For concepts, users have difficulty identifying the concepts in the images [72]. With heatmaps and prototypes, users struggle to discern what objects are in the highlighted regions [44, 11].

2.2.3 Limited Means for User-Driven Exploration

Although attribution-based explanations show *what* regions are important, they do not explain *why* they are important [61, 90, 24]. Indeed, Kim et al. [44] found that users wanted to know why certain regions were deemed important by exploring the underlying causal relationships.

2.3 Interactivity and Explainable AI

While the problems presented could be partially mitigated by adjusting the amount of information in the static explanation on a case-by-case basis, interactivity has been identified as an effective [86, 46, 87, 85] and customizable [87, 85] way for users to interact with and understand data [12, 95]. Within XAI, there has been some work on incorporating interactivity to explore datasets and models more generally [2, 93]. Much of the existing work in interactivity for XAI has largely implemented interfaces that allow users to explore images at the dataset level by filtering through data points based on model predictions or their calculated similarity [32, 12, 27, 8, 92, 23]. Other work provides tools to examine the inner workings of models (e.g., model weights and learned features) [25, 30, 93, 16, 37, 33, 23].

There is limited research on interactivity at the explanation level for a single image. Some related work includes interactivity for textual or tabular data [52, 83, 9, 10]. For example, Slack et al. [83] develop a chatbot for querying a model trained on tabular data; however, their work focuses on conversational agents, whereas we focus on direct manipulation. As another example, Liu et al. [52] explore interactive explanations that rely on direct manipulation, such as sliders and drop-down menus. However, they focus on comparing how human-AI teams with interactive explanations perform compared to AI-only agents when identifying out-of-distribution points. Within computer vision, Nguyen et al. [67] develop an interface for human-AI teams where users provide feedback on explanations. In contrast, our work seeks to investigate how users leverage and perceive interactive CV explanations and to provide design recommendations.

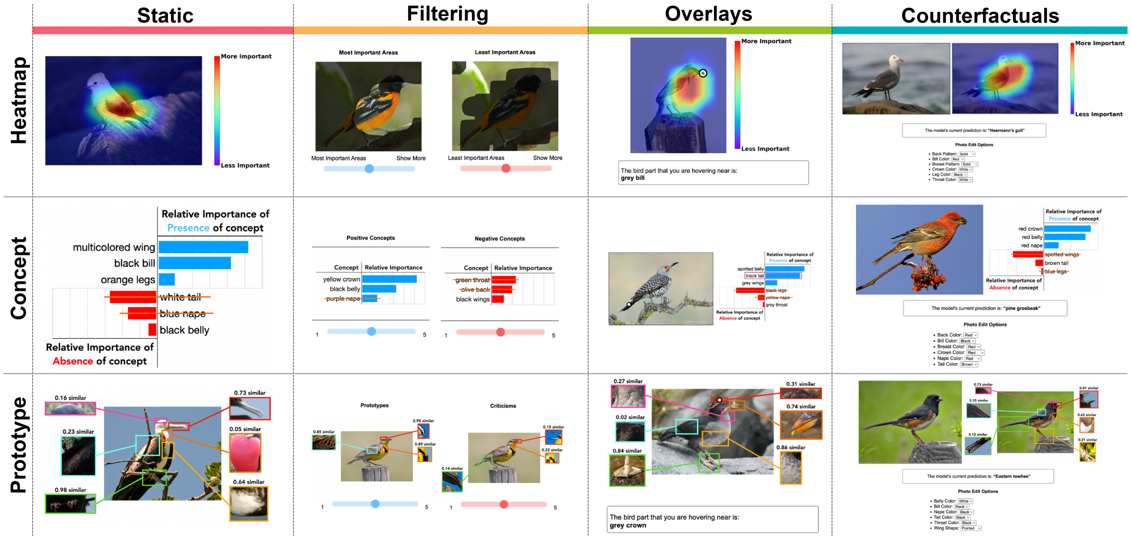
Chapter 3

Exploratory User Study

In this chapter, we describe the details of our IRB-approved study.

3.1 Terminology

We tested three *explanation types*: heatmap-based, concept-based, and prototype-based (i.e., the rows in Fig. 3.1). We created mock-up explanations, allowing us to carefully control the *presentation* of explanations across participants. For each explanation type, we created four *presentation types*: Static, Filtering, Overlays, and Counterfactuals (i.e., the columns in Fig. 3.1). The latter three were interactive mechanisms inspired by the problems outlined in Section 2.1. An *explanation* is a pairing of one explanation type and one presentation type (i.e., one cell in Fig. 3.1).



The bird part that you are hovering near is:
grey bill

The model's current prediction is: **"Heermann's gull"**

Photo Edit Options

- Back Pattern: Solid
- Bill Color: Red
- Breast Pattern: Solid
- Crown Color: White
- Leg Color: Black
- Throat Color: White

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 bird part that you are hovering near is:
grey crown

The model's current prediction is: **"Eastern towhee"**

Photo Edit Options

- Belly Color: White
- Bill Color: Black
- Nape Color: Black
- Tail Color: Black
- Throat Color: Black
- Wing Shape: Pointed

Figure 3.1: **12 Mock-up Explanations.** The 3 explanation types (rows) and 4 presentation types (columns) are shown. All bird images were from the CUB dataset [91]; the dataset creators do not claim copyright over the original images.

3.2 Designing the Mock-up Explanations

For the **Static** baseline (Fig. 3.1, 1st column), we based our mock-ups on prior works [76, 17, 44]. We then adapted them so that they shared some uniform characteristics across presentation types. First, we wanted all explanations to have a visual component, so we used bar graphs instead of numerical scores in the concept version (heatmap and prototype versions already had visual components), similar to TCAV [41]. This design decision was also in line with prior work, which reported that users dislike or feel overwhelmed by the numerical coefficients [44]. Second, we wanted explanations to describe an individual image (i.e., local explanation), not an entire output class (i.e., global explanation) which in our case was a bird species. Concept-based explanations are typically global, as concepts may not appear in every image. Thus, we made local concept-based explanations by crossing out concepts not present via strikethrough [44].

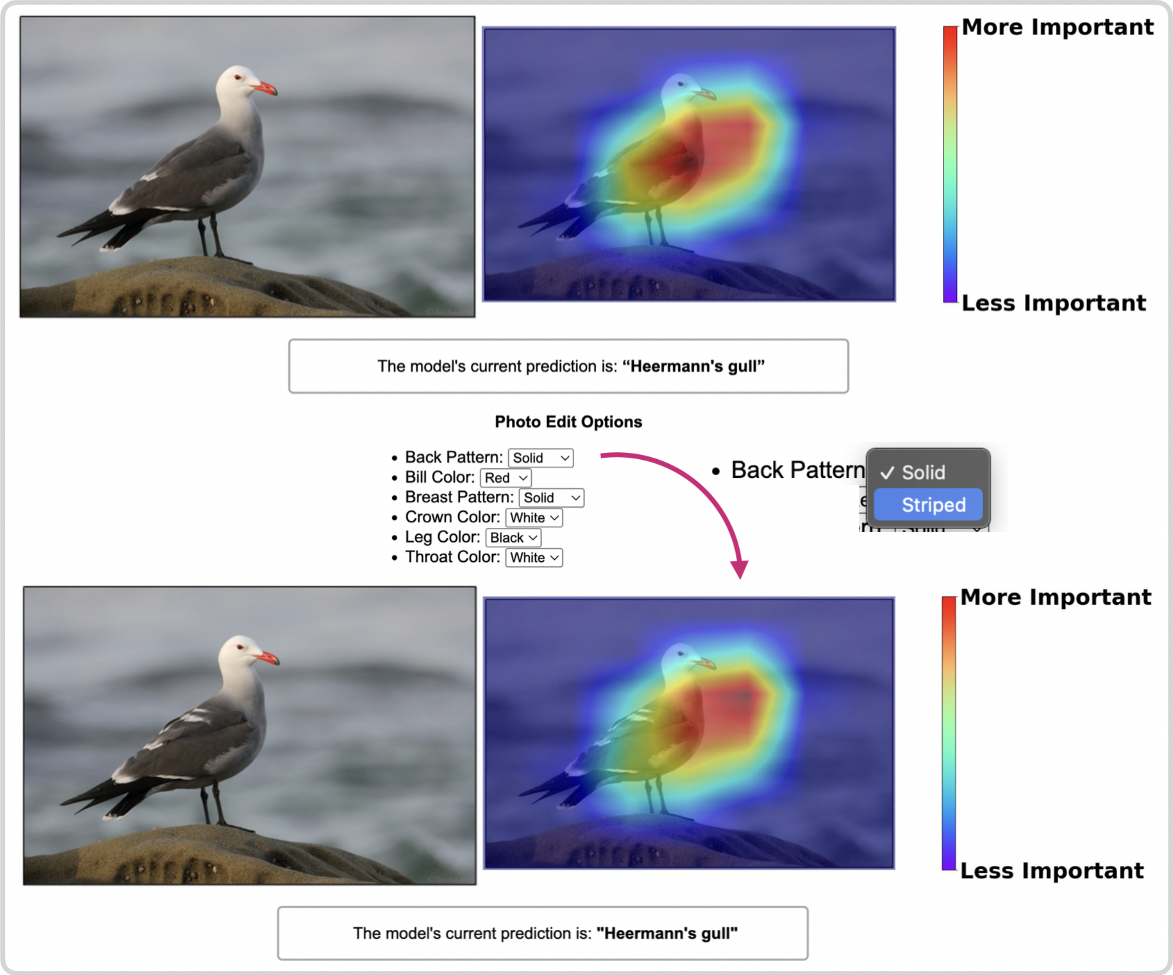
Filtering (Fig. 3.1, 2nd column) is a type of interactivity that allows the user to conditionally add and remove information [26, 38, 95, 78]. We utilized sliders that hide or show components of the explanation, conditioned on their assigned importance [52]. In the concept and prototype versions, the sliders remove or add concepts and prototypes. In the heatmap version, we replaced the color gradient with a binary mask that grows or shrinks as the slider is moved. We chose this design because layering the mask on top of the color gradient makes the underlying image difficult to see [44].

Overlays (Fig. 3.1, 3rd column), described by Shneiderman as “details-on-demand” [78], are extra layers of information [48] that appear only when the user indicates that they want to view them [89]. We incorporate a hover-over that supplies the pixel locations or the semantic labels of bird parts in the explanation, depending on which is missing. When users hover over a heatmap or prototype version, a dot appears on the closest bird part and a dialog box below the explanation shows a textual label for that part. For the concept version, when users hover over a concept in the bar chart that is visible in the image (i.e., not crossed out), a dot representing the concept's location appears on the image.

Counterfactuals (Fig. 3.1, 4th column) explain model predictions by allowing the user to explore how changes to model inputs affect model outputs [62]. Typically, static counterfactual explanations show an edited image with the smallest modification needed to change the model's output prediction from one class to another (e.g., how does an image need to change so that a model predicts "Cardinal" instead of "Blue jay"?) [29, 88]. We utilize Counterfactuals as a form of interactivity by allowing users to edit an image and see how the explanation and prediction subsequently change. Liu et al. [52] use a similar mechanism for tabular data where users can edit the input numbers to a model. An example of our Counterfactuals is shown in Fig. 3.2.

Users are provided with six drop-down menus that we refer to as "edit options". Each menu corresponds to a bird part and provides two options for the attribute of that bird part: one is the original and the other is a randomly selected alternative. For example, if the bird has a solid back pattern and the selected alternative is a striped back pattern, clicking the menu for "Back Pattern" will display the options "Solid" and "Striped". If a user selects the alternative attribute, the image updates to show the edit, and the explanation updates when the edited image changes it. Additionally, we include a dialog box that displays the model's prediction for the current image; this also updates if the edited image changes the model's prediction.

For each explanation, we chose a different image to mitigate learning effects. We used a series of preprocessing steps and consulted with two birding experts to select 12 images of similar difficulty in terms of bird identification and image quality. We performed these checks to ensure some level of consistency across images in terms of lighting, camera angle, or other factors. Additionally, two CV experts (two authors) checked that the explanations were of similar difficulty.



The model's current prediction is: "Heermann's gull"

Photo Edit Options

- Back Pattern: Solid
- Bill Color: Red
- Breast Pattern: Solid
- Crown Color: White
- Leg Color: Black
- Throat Color: White

The model's current prediction is: "Heermann's gull"

Figure 3.2: **Counterfactuals Example.** This figure shows the heatmap version of the Counterfactuals presentation type. Among the six edit options available below the prediction dialog box, the user changes the attribute for a bird part (e.g., changing the Back Pattern to be Striped instead of Solid), and the updated prediction and explanation are shown. In this example, the prediction did not change, and the heatmap shifted slightly to the right.

3.3 Participants

We designed a within-subjects study and recruited 24 participants with varying levels of ML and domain expertise. In the screening form, participants rated their familiarity with ML and birding on a five-point scale, where each point had a one-sentence description of the expected level of familiarity. We categorized participants as follows and recruited 6 participants from each of these categories (Fig. 3.3):

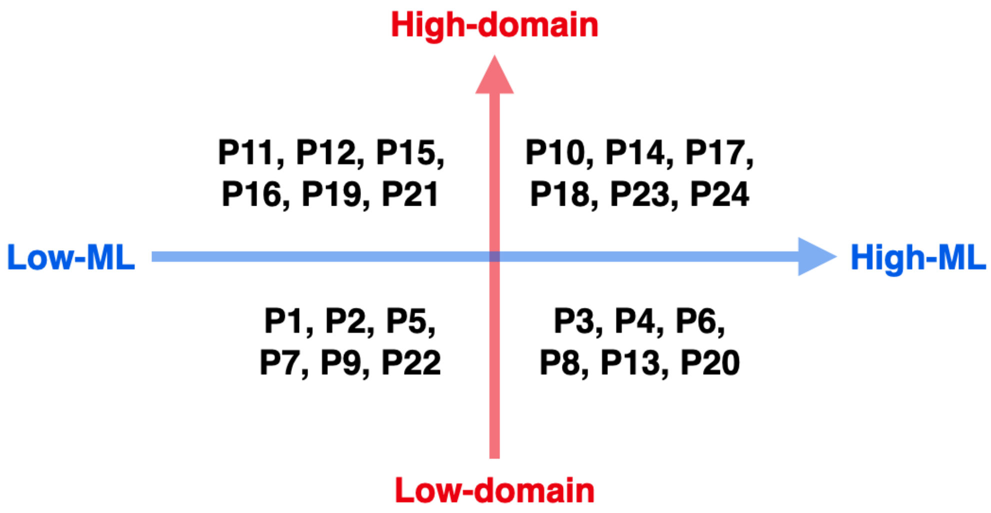


Figure 3.3: **Distribution of participants among ML and domain expertise.** We have four background categories formed by all possible pairings of high and low ML background with high and low domain background. Our 24 participants were distributed evenly across all categories with 6 per category.

- Low-ML: Participants who responded “1: I don’t know anything about ML” to “3: I know the basics of ML and can hold a short conversation about it”.
- High-ML: Participants who responses “4: I have taken a course on ML and/or have experience working with a ML system” to “5: I often use and study ML”.
- Low-Domain: “1: I don’t know anything about birding” to “3: I know the basics of birding and can hold a short conversation about it”.
- High-Domain: “4: I have taken a course on birding and/or have experience in birding” to “5: I often conduct bird-watching and study birding”.

We advertised our study through various platforms, including several institutions’ email lists, birding labs, online conservation and birding forums, and Mastodon. Due to our snowball sampling process, all participants held an academic or research affiliation. Studies were conducted over Zoom video calls and lasted 90–100 minutes. All participants received a $25 gift card.

3.4 Procedure

We designed a within-subjects study during which each participant tried all 12 explanations shown in Fig. 3.1. Our study consisted of an web application for viewing the mock-ups and a survey for providing written answers; we provide snapshots of the interface in the Appendix.

For the study conditions, we formed a balanced Latin square across the three explanation types and randomized the presentation type ordering. Thus, there were 6 possible orderings with 4 participants assigned per ordering. To ensure that participants from different expertise levels experienced the same ordering, each of the 4 participants was randomly selected from a different background category. That is, for each of the 6 possible conditions, there was one participant with a Low-ML and Low-Domain background, one with a Low-ML and High-Domain background, and one each from the remaining two categories. For each explanation, the participant went through an exploration phase and then answered survey questions.

Exploration Phase. First, the participant was asked to read the instructions under the explanation and then view and interact with the explanation. The instructions described the explanation type and how to use the interactive mechanism if present. The participant could choose when to proceed to the questions; however, we enforced a time limit of approximately two minutes.

Tasks. We included tasks to encourage participants to actively engage with the explanation rather than passively view it. Since attribution-based explanations show the most and least important regions of an image, we designed tasks that involved identifying these regions. The survey questions asked participants to identify and rank the three most and three least important parts of a bird based on the explanation. All task questions included a reference to a bird part guide that we created to assist with birding terminology.

Statement Ratings and Preference Rankings. At the end of the study, participants were asked to rate their level of agreement with a set of statements on a five-point Likert scale. The statements were based on the three problems identified in Sec. 2.2.

- "There was too much information provided."
- "It was difficult to connect the explanation to parts of the photo."
- "It was difficult to understand why the model produced this explanation."

Additionally, they ranked the presentation types in terms of their general preference and their preference in the context of learning about a new bird species, allowing for ties:

- "Based on the labels in the table, rank your preference of the following four groups of explanation designs (1 being most preferred and 4 being least preferred) and verbally explain your thought process." The labels were W, X, Y, and Z with each letter representing Static, Filtering, Overlays, and Counterfactuals respectively.
- "Let's say that you needed to learn about a new bird species. Rank how useful the four groups of explanation designs would be for doing so (1 being most useful and 4 being least useful) and verbally explain your thought process." As with the previous question, the labels were W, X, Y, and Z with each letter representing Static, Filtering, Overlays, and Counterfactuals respectively.

3.5 Analysis

We collected qualitative data from the study recordings and quantitative data from the survey questions. For our qualitative data, we created a codebook from the audio and video recordings of the studies and performed a Reflexive Thematic Analysis [14, 13] to extract findings and outline recommendations. Some codes directly came from verbal responses (e.g., "*User prefers Counterfactuals for learning because exploration*"), while other codes described participant interactions (e.g., "*User interacted with each of the edit options individually*"). Two authors created an initial codebook based on four studies, one from each background category (Fig. 3.3) to encompass a range of perspectives. Then, one of the two authors coded the remaining studies. All authors iterated on the codebook to form themes across participant interactions and verbal feedback. Since one author coded all the studies and refined the codebook in agreement with the other authors, we did not calculate inter-rater reliability.

Chapter 4

Results

In this chapter, we present the quantitative results from the surveys followed by the themes from our qualitative analysis of the interviews.

4.1 Quantitative Results

After viewing all 12 explanations, participants rated a set of statements (Sec. 3.4) on a 5-point Likert scale and ranked the four presentation types. The results for the ratings and rankings are shown in Figure 4.1 and Figure 4.2 respectively. Lower numbers indicate better ratings.

For the analysis, we used a two-sided Wilcoxon signed-rank test with Holm correction; we used Holm to avoid assuming independence. Participants rated Counterfactuals as having significantly more information compared to Static, Filtering, and Overlays ($p < .01$). Participants rated Filtering as having significantly more information than Static ($p = .015$). 4 participants noted that while Filtering provided more information than Static, the additional information was not overwhelming. For connecting pixel-level information to semantic-level information, participants rated Overlays as significantly less difficult than Static, Filtering, and Counterfactuals ($p < .01$). There were no significant differences in ratings for the model understanding statement. In general, participants preferred Overlays significantly more than Static ($p < .01$) and Counterfactuals ($p = .02$). For learning about bird species, participants preferred Overlays significantly more than Static presentations ($p < .01$).

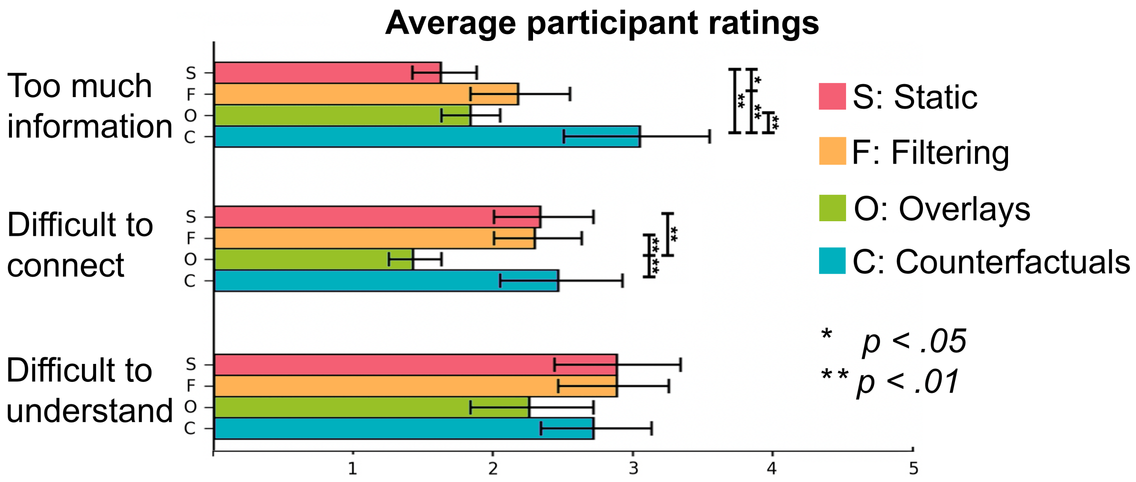


Figure 4.1: **Average participant ratings for survey statements.** Lower is better. Error bars are 95% confidence intervals.

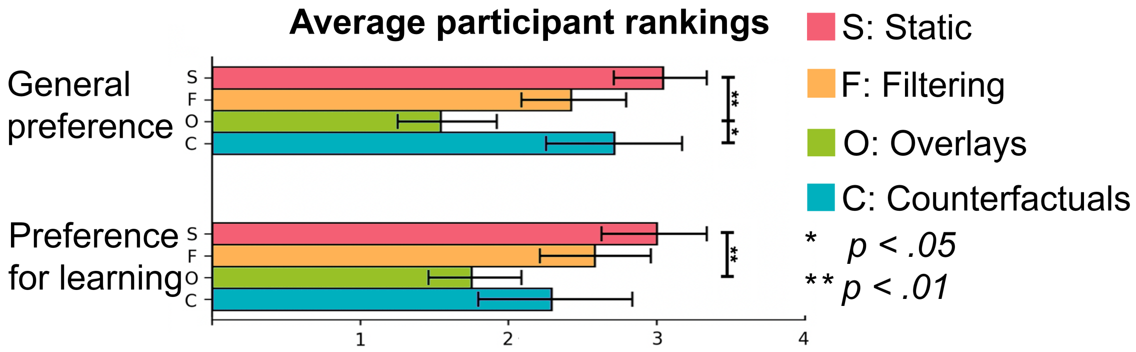


Figure 4.2: **Average participant rankings for general preference and preference for learning bird species.** Ties are allowed. Lower is better. Error bars are 95% confidence intervals.

4.2 Qualitative Findings

In this section, we focus on the themes from our qualitative analysis of participants' verbal feedback and interactions. Participants utilized the interactive mechanisms in most trials (214 out of 216) when these mechanisms were available. We refer to participants from the four background categories using the following notation: HM-HD for "High-ML and High-Domain", LM-LD for "Low-ML and Low-Domain", and similar notation for the combinations HM-LD and LM-HD.

4.2.1 Participants appreciate interactive mechanisms that augment the explanation without changing the underlying explanation.

Participants appreciated Filtering and Overlays as tools for adjusting the amount of detail in the explanation without changing the explanation itself. Specifically, participants mentioned that Filtering made explanations less overwhelming by giving them control over the amount of visible information (4 LM-LD, 2 HM-LD, 2 LM-HD, 4 HM-HD) and this control in turn made explanations easier to digest (2 LM-LD, 1 LM-HD, 1 HM-HD). Participants also used this interactive mechanism to reduce the amount of information (3 LM-LD, 1 HM-LD, 2 LM-HD, 2 HM-HD) or gradually add information (2 LM-LD, 2 HM-LD, 3 LM-HD) during the task. Similarly, participants liked Overlays (5 LM-LD, 4 HM-LD, 3 LM-HD, 6 HM-HD) despite the additional information presented. For example, P2 said: "*[Overlays have] more information than [Static], but I feel like the way they're presented makes it so that you can take it in easier, and it doesn't feel like more information*". On the other hand, participants found Counterfactuals overwhelming (5 LM-LD, 2 HM-LD, 3 LM-HD, 2 HM-HD), as image edits altered the underlying explanations.

4.2.2 Participants find interactive mechanisms that alter the underlying explanation overwhelming.

Participants found that Counterfactuals were overwhelming because there could be simultaneous changes in the explanation when the image was modified. Additionally, they struggled to untangle the causal relationships because the inputs were interdependent (3 LM-LD, 4 HM-LD, 1 LM-HD, 3 HM-HD). P6 commented that "*In my mind, itâs hard to reason about the combination of influences*", and P2 stated for the prototypes version that they "*think there's too much going on with the photo editing and the [prototypes] popping out*". In fact, participants often performed one edit at a time, making sure to revert each edit option to the default choice before trying another (61 out of 72 Counterfactuals trials). Moreover, participants had to explore a large number of possible explanations that resulted from the 64 possible edited images (1 LM-LD, 5 HM-LD, 1 LM-HD, 2 HM-HD). P7 remarked: "*There were so many things to look at—it was confusing*". Overall, participants found Counterfactuals confusing and hard to interpret (5 LM-LD, 2 HM-LD, 3 LM-HD, 2 HM-HD), particularly those with low ML and domain expertise, and needed more time to explore the space of possible inputs and the resulting outputs (1 LM-LD, 2 HM-LD, 3 LM-HD).

4.2.3 Although participants find Counterfactuals overwhelming, they utilize them to resolve confusion around static presentations by inducing systematic changes in model predictions and explanations.

Participants leveraged Counterfactuals to clarify aspects of static explanations they found confusing and to confirm their understanding of the explanations. For example, many participants were confused about static concept-based explanations (4 LM-LD, 5 HM-LD, 3 LM-HD, 4 HM-HD), and some used Counterfactuals to edit the image and observe the resulting changes in the model's prediction and explanation to clarify their understanding, which ultimately led to a more accurate interpretation of the static explanation (3 LM-HD, 3 HM-HD). P17 noted: "*It was helpful to be able to* *change the color of [a bird part] and see how that affected the bar graph...that sort of helped me understand what the [strikethrough] is for*". Additionally, the visual changes helped participants identify the bird parts (3 LM-LD, 5 HM-LD, 5 LM-HD, 2 HM-HD). Hohman et al. observed similar behaviors where "*...participants used Counterfactuals often throughout their exploration, both as a direct task and as a sanity check for feature sensitivity*" in their study with ML practitioners [32].

4.2.4 Participants leverage Counterfactuals to explore a range of explanations to better understand the AI model more broadly and beyond the task at hand.

Many participants found Counterfactuals useful for understanding the AI model's reasoning at a high level (6 LM-LD, 6 HM-LD, 5 LM-HD, 3 HM-HD). Some participants expressed a desire to use Counterfactuals to explore other birds outside of the one image given and to build a macroscopic understanding of the AI model. Specifically, participants appreciated the ability to change the bird species by editing the image (2 HM-LD, 3 LM-HD, 1 HM-HD). This ability to explore allowed participants to develop an understanding of the model more broadly. For example, P15 thought it helped them build rules about how changes in the modelâs input affect its output, and P18 noted that this mechanism would allow them to identify the decision points of the model. These uses were beyond the scope of the tasks which only asked for the most and least important bird parts in the given image. This suggests that participants saw Counterfactuals as a source of information to explore and expand their understanding of the model, even if they were not asked to do so.

4.2.5 Participants felt that Filtering and Overlays allowed them to quickly focus on information of interest.

Some participants appreciated that Filtering automatically sorted the bird parts by importance (e.g., similarity scores sorted by magnitude in prototype-based explanations) because it obviated the need for manual comparison (2 LM-LD, 1 HM-LD, 1 LM-HD, 2 HM-HD). For instance, P6 liked the prototype version of Filtering because "*it produces a ranking of the similarity scores that I don't need to think about myself*". Similarly, participants found Overlays useful for easily and quickly identifying the names of the bird parts, even with the bird part guide provided (1 LM-LD, 2 HM-LD, 2 LM-HD, 3 HM-HD). For example, P21 remarked that Overlays were "*faster because I could hover over it and it could tell me what the parts were*". Thus, participants felt that the automatic sorting and labeling offered by Filtering and Overlays respectively enabled them to efficiently hone in on the information that they were seeking. Participants noted that the lack of these features in Static presentations was inconvenient (1 HM-LD, 1 LM-HD, 3 HM-HD).

Chapter 5

Discussion

In this chapter, we summarize our findings and offer design recommendations. Additionally, we discuss the limitations of our work and identify avenues for future work.

5.1 Summary

Participants found that interactive features augmenting explanations, such as Filtering and Overlays, helped them understand information without being overwhelming. While Counterfactuals introduced significant visual changes by altering the underlying explanations, participants used them to clarify confusing aspects of static explanations, reaffirm their understanding, and explore the AI model more broadly. With Filtering and Overlays, participants felt they could quickly focus on relevant information, and they appreciated the flexibility that these mechanisms offered when interpreting explanations.

5.2 Design Recommendations

Participants expressed that interactive mechanisms were at times confusing or inconvenient. Based on these findings, we recommend the following design considerations for future work on interactive CV explanations.

- **Avoid interdependent input controls.** Participants found Counterfactuals confusing because they could not disentangle the effects of multiple simultaneous edits on the prediction or the explanation. A more appropriate design might involve preventing users from performing more than one edit at a time, thereby clarifying the causal relationship between each input and the corresponding changes.

- **Constrain the input and output space.** With Counterfactuals, many participants felt overwhelmed by the number of possible image edits and the extent of changes to the explanation. We suggest limiting the input and output space to create a manageable range of information, tailored to the application, allowing participants to easily explore all.
- **Design an optimal static default view.** We recommend that users should have the option to choose whether or not to engage with interactive mechanisms. Therefore, the default static presentation should provide enough information for the initial interpretation of the explanation without overwhelming users [72].

5.3 Limitations

Overall, we find that interactivity can benefit users' understanding and experience when interpreting CV explanations. However, developing interactive explanations for real-world applications requires more effort than static explanations. Developers need to consider factors like ease-of-use in generating interactive explanations for novel models and usability testing of interactive elements. Furthermore, evaluating interactive explanations requires a different approach from static explanations. Much work on evaluating static explanations has focused on automatic, qualitative evaluations, as is typical in machine learning research. However, evaluating and improving interactive explanations also necessitates qualitative, human-centered evaluation, especially during the early stages of development when user feedback can inform the design. Due to these barriers, relatively little work has been done on interactive CV explanations, leaving us with limited guidance for making design decisions. Thus, our study is a preliminary step toward evaluating such explanations and has several limitations.

For example, our study sessions were relatively short, with participants only having two minutes to explore each explanation. Some participants expressed a desire for more time, suggesting the need for future studies to explore deeper engagement with interactive mechanisms for CV explanations. Additionally, we tested a limited number of designs. We focused on a subset of three interactive mechanisms, motivated by prior work, and tested one design for each mechanism. However, there are more potential mechanisms and designs (e.g., different ways of allowing users to edit images in Counterfactuals). Lastly, longitudinal studies are needed to understand how users interact across repeated sessions and the role of interactive mechanisms that maintain user history.

We chose a simple bird identification task, which was a reasonable starting point, because it was consistent with prior work [44, 45, 63] and allowed us to recruit participants with low and high domain expertise. However, it is unclear to what extent our findings can be generalized to more complex, high-risk, and high-impact applications such as medical diagnostics. Moreover, our sample size was relatively small, underscoring the need for larger-scale studies. Lastly, our interactive mechanisms were not directly comparable to the static explanations. For example, the Filtering mechanism for heatmap-based explanations lacked the color gradient present in its static counterpart. Carefully controlled experiments are needed for quantitative analysis of performance, trust, and confidence in tasks using various interactive CV explanations.

5.4 Future Work

Future research should explore a broader range of interactive mechanisms for CV explanations beyond the three examined in this work. Additionally, some participants suggested integrating information beyond the image, such as geographical context or general facts, to improve the utility of explanations. Further work is needed to define the boundaries of what information should be included in interactive explanations across different application domains. Collaborating with domain experts and stakeholders through participatory design will be critical for advancing these directions.

Finally, prior work has shown that explanations may lead to misplaced trust in AI systems [19, 42, 43, 97, 5]. Interactivity in explanations could amplify these risks, further misleading users and enabling confirmation bias. For example, users might iteratively interact with Counterfactuals until the explanations appear to align with their preconceptions. More research is needed to carefully examine these possibilities and mitigate potential negative outcomes.

Chapter 6

Conclusion

In this thesis, we investigate how users leverage and perceive interactivity in the context of CV explanations, which have traditionally been static. Drawing from prior XAI research, we identified three key issues with these explanations: information overload, a semantic-pixel gap, and limited means for user-driven exploration. Given the potential of interactivity to remedy these issues, we explored how users engage with interactive explanations by comparing three mechanisms across three explanation types. We conducted a study with 24 participants of varying ML and domain expertise. Our findings showed that interactivity helps address the challenges of static explanations by offering users greater control and quicker access to relevant information. While some mechanisms were overwhelming, they allowed users to expand their understanding of the model and refine their grasp of the static explanation itself. Based on these insights, we offered design recommendations for interactive CV explanations. Of course, there is still much to explore in the space of interactivity. That is, there are many types of mix-ins to try with different flavors of ice cream (Fig. 6.1). Our work takes one step in the right direction by directly collaborating with end-users to identify their needs.

Figure 6.1: **More Food for Thought.** Ice cream mix-ins at the Rockefeller-Mathey Dining Hall.

Bibliography

[1] A. Adadi and M. Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[2] S. Amershi, M. Cakmak, W. B. Knox, T. Kulesza, and T. Lau. IUI workshop on interactive machine learning. In *Proceedings of the Companion Publication of the 2013 International Conference on Intelligent User Interfaces Companion*, IUI '13 Companion, page 121â124, New York, NY, USA, 2013. Association for Computing Machinery.

[3] A. M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences*, 11(11), 2021.

[4] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *CoRR*, abs/2112.11561, 2021.

[5] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. 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*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[6] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58(C):82â115, jun 2020.

[7] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] A. Bäuerle, A. A. Cabrera, F. Hohman, M. Maher, D. Koski, X. Suau, T. Barik, and D. Moritz. Symphony: Composing Interactive Interfaces for Machine Learning. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[9] A. Bhattacharya, J. Ooge, G. Stiglic, and K. Verbert. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 204â219, New York, NY, USA, 2023. Association for Computing Machinery.

[10] A. Bhattacharya, S. Stumpf, L. Gosak, G. Stiglic, and K. Verbert. EXMOS: Explanatory Model Steering through Multifaceted Explanations and Data Configurations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[11] P. Blignaut. Visual span and other parameters for the generation of heatmaps. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10, page 125â128, New York, NY, USA, 2010. Association for Computing Machinery.

[12] A. Boggust, B. Carter, and A. Satyanarayan. Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 746â766, New York, NY, USA, 2022. Association for Computing Machinery.

[13] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[14] V. Braun and V. Clarke. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597, 2019.

[15] W. Brendel and M. Bethge. Approximating CNNs with Bag-of-local-Features Models Works Surprisingly Well on ImageNet. In *International Conference on Learning Representations (ICLR)*, 2019.

[16] S. Chang, P. Dai, L. Hong, C. Sheng, T. Zhang, and E. H. Chi. AppGrouper: Knowledge-based Interactive Clustering Tool for App Search Results. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, page 348â358, New York, NY, USA, 2016. Association for Computing Machinery.

[17] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[18] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen. A survey of the state of explainable AI for natural language processing. In K.-F. Wong, K. Knight, and H. Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China, Dec. 2020. Association for Computational Linguistics.

[19] V. Danry, P. Pataranutaporn, Y. Mao, and P. Maes. Donât Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[20] A. J. DeGrave, Z. R. Cai, J. D. Janizek, R. Daneshjou, and S.-I. Lee. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *Nat. Biomed. Eng.*, Dec. 2023.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[22] J. Donnelly, A. J. Barnett, and C. Chen. Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[23] J. J. Dudley and P. O. Kristensson. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.*, 8(2), jun 2018.

[24] N. El Bekri, J. Kling, and M. F. Huber. A study on trust in black box models and post-hoc explanations. In F. Martínez Álvarez, A. Troncoso Lora, J. A. Sáez Muñoz, H. Quintián, and E. Corchado, editors, *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019)*, pages 35–46, Cham, 2020. Springer International Publishing.

[25] K. J. K. Feng and D. W. Mcdonald. Addressing UX Practitionersâ Challenges in Designing ML Applications: an Interactive Machine Learning Approach. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 337â352, New York, NY, USA, 2023. Association for Computing Machinery.

[26] A. Figueiras. Towards the Understanding of Interaction in Information Visualization. In *2015 19th International Conference on Information Visualisation*, pages 140–147, 2015.

[27] J. Fogarty, D. Tan, A. Kapoor, and S. Winder. CueFlik: interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing*

Systems, CHI ’08, page 29â38, New York, NY, USA, 2008. Association for Computing Machinery.

[28] R. Fong and A. Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *International Conference on Computer Vision (ICCV)*, 2017.

[29] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual Visual Explanations. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR, 09–15 Jun 2019.

[30] L. Guo, E. M. Daly, O. Alkan, M. Mattetti, O. Cornec, and B. Knijnenburg. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI ’22, page 537â548, New York, NY, USA, 2022. Association for Computing Machinery.

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[32] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1â13, New York, NY, USA, 2019. Association for Computing Machinery.

[33] F. Hohman, H. Park, C. Robinson, and D. H. Chau. Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2020.

[34] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med*, 29(9):2307–2316, Sep 2023.

[35] J. Janai, F. GÃ¼ney, A. Behl, and A. Geiger. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Foundations and TrendsÂ® in Computer Graphics and Vision*, 12(1â3):1–308, 2020.

[36] W. Jin, X. Li, and G. Hamarneh. Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11945–11953, Jun. 2022.

[37] A. Kapoor, B. Lee, D. Tan, and E. Horvitz. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1343â1352, New York, NY, USA, 2010. Association for Computing Machinery.

[38] D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[39] E. M. Kenny, M. Tucker, and J. Shah. Towards Interpretable Deep Reinforcement Learning with Human-Friendly Prototypes. In *The Eleventh International Conference on Learning Representations*, 2023.

[40] B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! Criticism for Interpretability. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[41] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 10–15 Jul 2018.

[42] S. S. Y. Kim, N. Meister, V. V. Ramaswamy, R. Fong, and O. Russakovsky. HIVE: Evaluating the Human Interpretability of Visual Explanations. In *European Conference on Computer Vision (ECCV)*, 2022.

[43] S. S. Y. Kim, J. W. Vaughan, Q. V. Liao, T. Lombrozo, and O. Russakovsky. Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2025.

[44] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández. "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*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[45] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernandez. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[46] Y.-S. Kim, K. Reinecke, and J. Hullman. Explaining the Gap: Visualizing One's Predictions Improves Recall and Comprehension of Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 1375â1386, New York, NY, USA, 2017. Association for Computing Machinery.

[47] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept Bottleneck Models. In *International Conference on Machine Learning (ICML)*, 2020.

[48] N. Kong and M. Agrawala. Graphical Overlays: Using Layered Elements to Aid Chart Reading. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2631–2638, 2012.

[49] S. Kriglstein, G. Wallner, and M. Pohl. A user study of different gameplay visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 361â370, New York, NY, USA, 2014. Association for Computing Machinery.

[50] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, page 126â137, New York, NY, USA, 2015. Association for Computing Machinery.

[51] Q. V. Liao, D. Gruen, and S. Miller. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1â15, New York, NY, USA, 2020. Association for Computing Machinery.

[52] H. Liu, V. Lai, and C. Tan. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[53] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768â4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[54] J. LÃPtsch, D. Kringel, and A. Ultsch. Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients. *BioMedInformatics*, 2(1):1–17, 2022.

[55] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment Anything in Medical Images. *Nature Communications*, 15:1–9, 2024.

[56] R. Machlev, L. Heistrene, M. Perl, K. Levy, J. Belikov, S. Mannor, and Y. Levron. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, 9:100169, 2022.

[57] A. Mamalakis, I. Ebert-Uphoff, and E. A. Barnes. *Explainable Artificial Intelligence in Meteorology and Climate Science: Model Fine-Tuning, Calibrating Trust and Learning New Science*, pages 315–339. Springer International Publishing, Cham, 2022.

[58] Q. Meteier, M. Capallera, L. Angelini, E. Mugellini, O. A. Khaled, S. Carrino, E. De Salis, S. Galland, and S. Boll. Workshop on Explainable AI in Automated Driving: A User-Centered Interaction Approach. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, AutomotiveUI '19, page 32â37, New York, NY, USA, 2019. Association for Computing Machinery.

[59] S. Michielssen, A. Maloof, J. Haumacher, A. Dreger, K. Bonicki, and K. Hallgren. Using large language models to generate baseball spray charts in the absence of numerical data. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, (0):1–8, 07/2024 2024.

[60] S. Milani, N. Topin, M. Veloso, and F. Fang. Explainable Reinforcement Learning: A Survey and Comparative Review. *ACM Comput. Surv.*, 56(7), apr 2024.

[61] T. Miller. Contrastive explanation: a structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021.

[62] T. Miller. 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*, FAccT '23, page 333â342, New York, NY, USA, 2023. Association for Computing Machinery.

[63] K. Morrison, P. Spitzer, V. Turri, M. Feng, N. Kühl, and A. Perer. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1), Apr. 2024.

[64] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.*, 55(13s), jul 2023.

[65] M. Nauta, R. van Bree, and C. Seifert. Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[66] M. Nazar, M. M. Alam, E. Yafi, and M. M. Suâud. A Systematic Review of Humanâ-Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques. *IEEE Access*, 9:153316–153348, 2021.

[67] G. Nguyen, M. R. Taesiri, S. S. Y. Kim, and A. Nguyen. Allowing Humans to Interactively Guide Machines Where to Look Does Not Always Improve Human-AI Team's Classification Accuracy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8169–8175, June 2024.

[68] G. Nguyen, M. R. Taesiri, and A. Nguyen. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[69] I. Panigrahi, S. S. Y. Kim*, A. Liaqat*, R. Jinturkar, O. Russakovsky, R. Fong, and P. Abtahi. Interactivity x Explainability: Toward Understanding How Interactivity Can Improve Computer Vision Explanations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA)*, 2025.

[70] V. Petsiuk, A. Das, and K. Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC)*, 2018.

[71] K. Qian, M. Danilevsky, Y. Katsis, B. Kawas, E. Oduor, L. Popa, and Y. Li. XNLP: A Living Survey for XAI Research in Natural Language Processing. In *Companion Proceedings*

of the 26th International Conference on Intelligent User Interfaces, IUI '21 Companion, page 78â80, New York, NY, USA, 2021. Association for Computing Machinery.

[72] V. V. Ramaswamy, S. S. Y. Kim, R. Fong, and O. Russakovsky. Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10932–10941, June 2023.

[73] V. V. Ramaswamy, S. S. Y. Kim, N. Meister, R. Fong, and O. Russakovsky. ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features, 2022.

[74] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135â1144, New York, NY, USA, 2016. Association for Computing Machinery.

[75] D. Rymarczyk, L. Struski, M. Górszczak, K. Lewandowska, J. Tabor, and B. Zieliński. Interpretable Image Classification with Differentiable Prototypes Assignment. In *Computer Vision â ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23â27, 2022, Proceedings, Part XII*, page 351â368, Berlin, Heidelberg, 2022. Springer-Verlag.

[76] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[77] V. Shitole, F. Li, M. Kahng, P. Tadepalli, and A. Fern. One Explanation is Not Enough: Structured Attention Graphs for Image Classification. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[78] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[79] P. Shukla, S. Bharati, and M. Turk. CAVLI - Using Image Associations To Produce Local Concept-Based Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3750–3755, June 2023.

[80] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR) Workshops*, 2014.

[81] A. Singh, S. Sengupta, and V. Lakshminarayanan. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*, 6, 2020.

[82] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera Y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, Aug. 2023.

[83] D. Slack, S. Krishna, H. Lakkaraju, and S. Singh. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, 5(8):873–883, 2023.

[84] H. Smith. Clinical AI: opacity, accountability, responsibility and liability. *AI & SOCIETY*, 36(2):535–545, 2021.

[85] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):672–681, 2019.

[86] S. S. Sundar, J. Oh, S. Bellur, H. Jia, and H.-S. Kim. Interactivity as self-expression: a field experiment with customization and blogging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 395â404, New York, NY, USA, 2012. Association for Computing Machinery.

[87] S. S. Sundar, Q. Xu, and S. Bellur. Designing interactivity in media interfaces: a communications perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 2247â2256, New York, NY, USA, 2010. Association for Computing Machinery.

[88] S. Vandenhende, D. Mahajan, F. Radenovic, and D. Ghadiyaram. Making Heads or Tails: Towards Semantically Consistent Visual Counterfactuals. In *European Conference on Computer Vision (ECCV)*, 2022.

[89] L. Vermette, S. Dembla, A. Y. Wang, J. McGrenere, and P. K. Chilana. Social CheatSheet: An Interactive Community-Curated Information Overlay for Web Applications. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), dec 2017.

[90] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2):841–887, 2018.

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

[92] Z. J. Wang, F. Hohman, and D. H. Chau. WizMap: Scalable Interactive Visualization for Exploring Large Machine Learning Embeddings. In *Demo, Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[93] Z. J. Wang, R. Turko, O. Shaikh, H. Park, N. Das, F. Hohman, M. Kahng, and D. H. Polo Chau. CNN Explainer: Learning Convolutional Neural Networks with Interactive

Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1396–1406, 2021.

[94] J. Xia and D. C. Wilson. Instructor Perspectives on Comparative Heatmap Visualizations of Student Engagement with Lecture Video. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18, page 251â256, New York, NY, USA, 2018. Association for Computing Machinery.

[95] J. S. Yi, Y. a. Kang, J. Stasko, and J. Jacko. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007.

[96] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014.

[97] Y. Zhang, Q. V. Liao, and R. K. Bellamy. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

[98] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

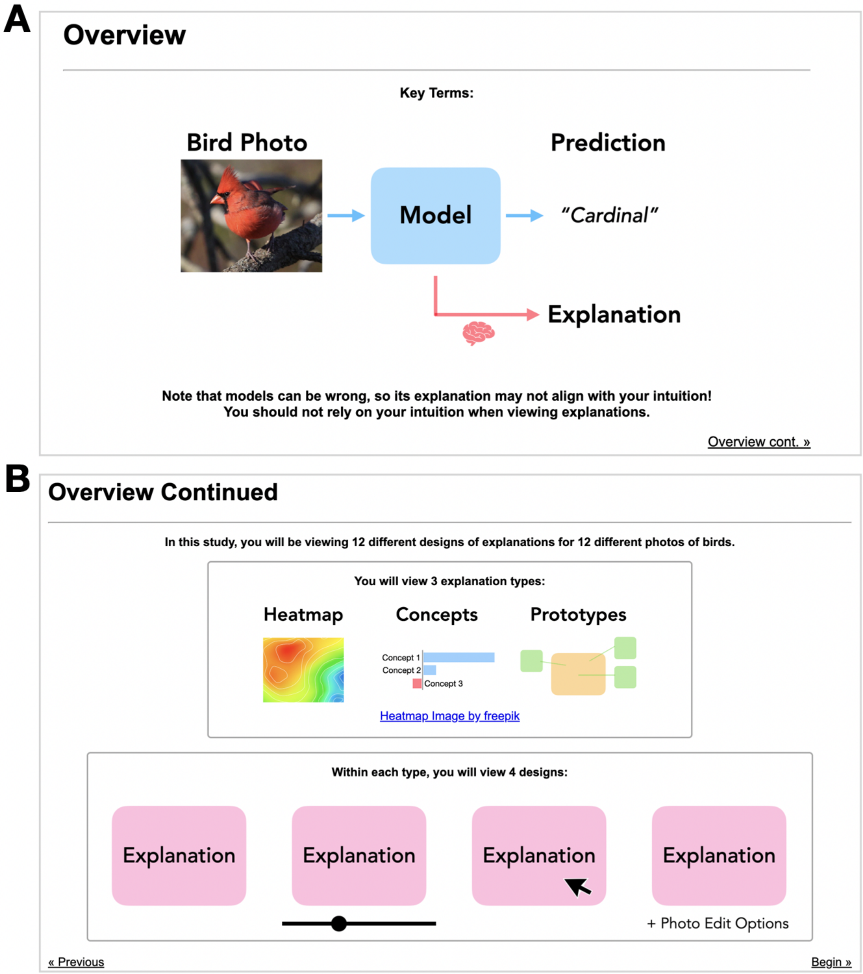
[99] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[100] B. Zhou, Y. Sun, D. Bau, and A. Torralba. Interpretable basis decomposition for visual explanation. In *European Conference on Computer Vision (ECCV)*, 2018.

Appendix

A.1 Study Interface

We built a Flask application, deployed using Render, to implement our mock-ups and guide participants through the study. In line with our verbal script, participants began by viewing the overview pages shown in Fig. A.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. A.2). Next, for each presentation type, participants viewed and, when applicable, interacted with the explanation (Fig. A.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 [91]. 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. A.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. A.5). Similar to the explanation type summary pages, participants provided ratings and rankings on the presentation types, this time generalizing along each column.



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.

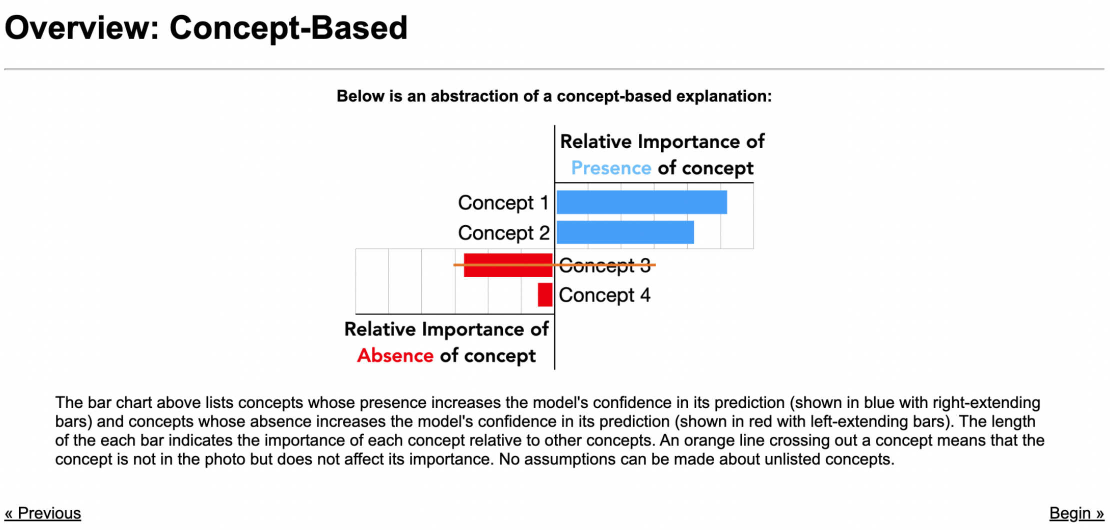
In this study, you will be viewing 12 different designs of explanations for 12 different photos of birds.

Figure A.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. Heatmap image attributed in the Panel B was designed by Freepik (https://www.freepik.com).

A.2 Constructing Mock-ups

All bird images were from the Caltech-UCSD Birds-200-2011 dataset [91]. 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).



Overview: Concept-Based

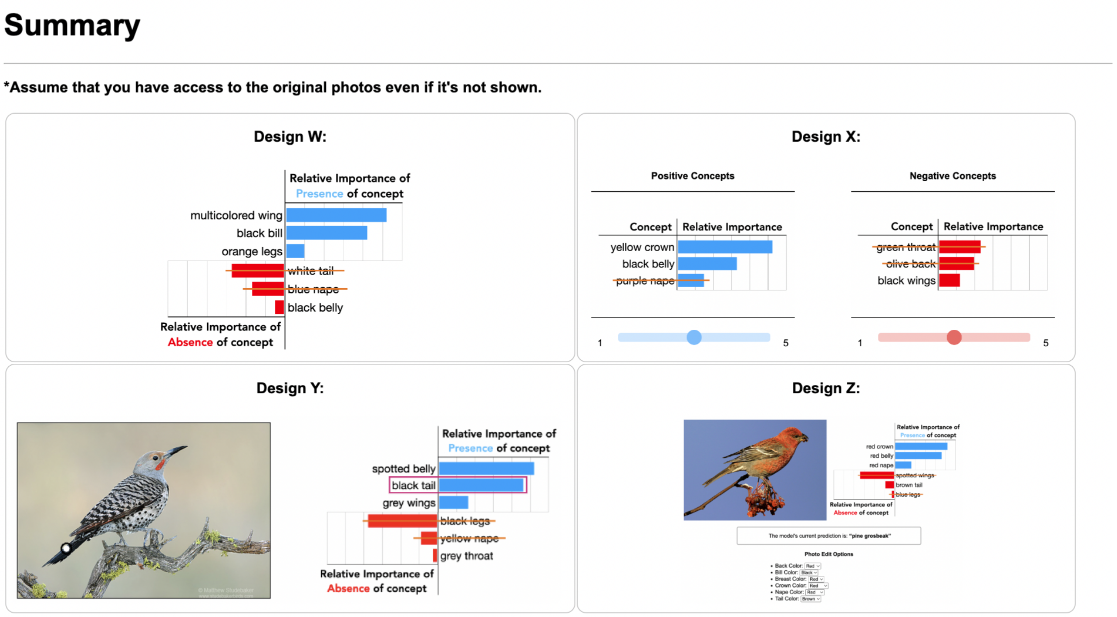
Below is an abstraction of a concept-based explanation:

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.

« Previous

Begin »

Figure A.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

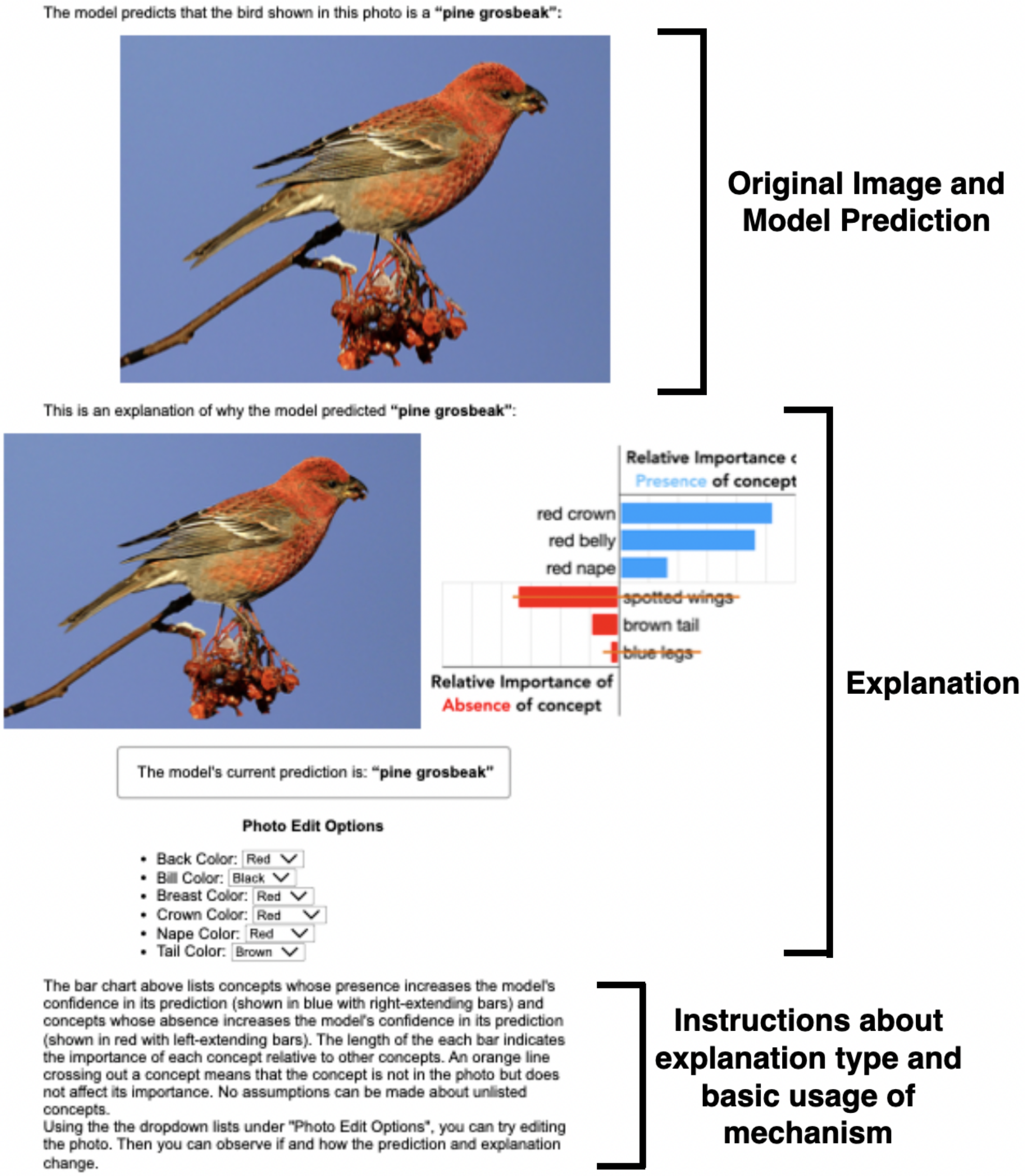
***Assume that you have access to the original photos even if it's not shown.**

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 A.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.



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"**:

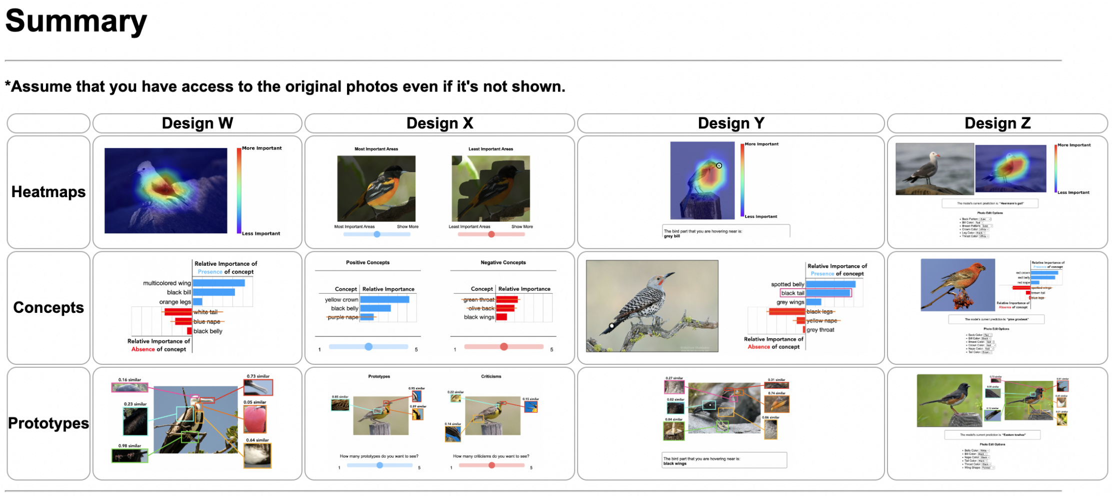
The model's current prediction is: **"pine grosbeak"**

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 A.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. A.2) and includes instructions for basic usage of the mechanism, if present.



Summary

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

Figure A.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.

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 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 [31] that had been trained on the CUB dataset [91].

A.3 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.