arXiv:2112.03184v1 [cs.CV] 6 Dec 2021

HIVE: Evaluating the Human Interpretability of Visual Explanations

Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky
Princeton University
{sunniesuhyoung, nmeister, vr23, ruthfong, olgarus}@princeton.edu

Abstract

As machine learning is increasingly applied to high-impact, high-risk domains, there have been a number of new methods aimed at making AI models more human interpretable. Despite the recent growth of interpretability work, there is a lack of systematic evaluation of proposed techniques. In this work, we propose a novel human evaluation framework HIVE (Human Interpretability of Visual Explanations) for diverse interpretability methods in computer vision; to the best of our knowledge, this is the first work of its kind. We argue that human studies should be the gold standard in properly evaluating how interpretable a method is to human users. While human studies are often avoided due to challenges associated with cost, study design, and cross-method comparison, we describe how our framework mitigates these issues and conduct IRB-approved studies of four methods that represent the diversity of interpretability works: GradCAM, BagNet, ProtoPNet, and ProtoTree. Our results suggest that explanations (regardless of if they are actually correct) engender human trust, yet are not distinct enough for users to distinguish between correct and incorrect predictions. Lastly, we also open-source our framework to enable future studies and to encourage more human-centered approaches to interpretability. HIVE can be found at https://princetonvisualai.github.io/HIVE.

1. Introduction

Complex artificial intelligence (AI) systems are increasingly deployed in important and high-risk domains, such as medical diagnosis, biometric recognition, and autonomous driving. In such settings, it is crucial to understand the behavior and relative strengths and weaknesses of these models. The *interpretability* research field tackles these questions and is comprised of a diversity of works, including those that design inherently interpretable models [9, 12, 30, 41], those that provide explanations of the behavior and inner workings of models [6, 8, 16, 18, 42, 50, 52, 58, 63], and those that seek to understand what is easy vs. difficult for these models and thereby make their behavior more interpretable [3, 54, 60].

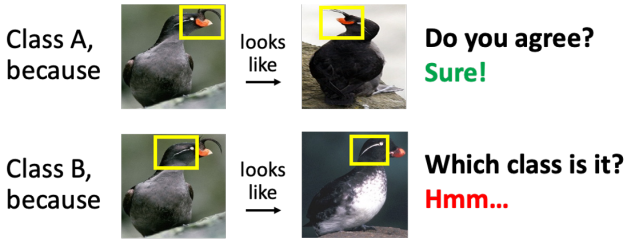

Figure 1. Our *agreement* (top) and *distinction* (bottom) tasks for evaluating the human interpretability of visual explanations.

Evaluating intepretability methods. Despite much methods development, there is a relative lack of standardized evaluation metrics for proposed techniques. The main challenge is the lack of ground truth for many problems. For example, consider GradCAM [50], a popular post-hoc explanation method that produces a heatmap, which highlights image regions relevant to a model's prediction. However, we lack ground-truth knowledge about which regions are *actually* responsible for a model's prediction; thus, different methods propose different proxy tasks for verifying "important" image regions (e.g., measuring the impact of deleting regions or the overlap between ground-truth objects and highlighted regions) [17, 26, 42, 43, 55, 59].

In part due to these challenges, interpretability is often argued through a few exemplar explanations that highlight how a method is more interpretable than a baseline model. However, recent works suggest that some methods are not as interpretable as originally imagined [1, 25, 40]; these works caution against an over-reliance on intuition-based justifications and raise awareness for the need of proper evaluation and falsifiable hypotheses in interpretability research [37]. The lack of ground truth for interpretability also highlights a need for human studies. In some cases, human judgment can serve as a gold standard and provide valuable insights into what forms of explanations are useful for human decision making.

Challenges of human studies. Conducting human evaluation for interpretability methods has its own set of chal-

lenges. First, designing the study and its user interface (UI) is non-trivial. Interpretability methods themselves can be complex; thus, for each method, researchers must design custom tasks that maintain the essence of the target application [14]. While this has been done for methods involving text and tabular data [23, 44], UI design is particularly challenging for vision-based explanations given the complexity of natural images.

Further, it is difficult to create standardized evaluation that allows for cross-method comparisons. Prior works typically conduct human studies to either demonstrate that their proposed method is reasonable and convincing to humans [6, 7, 64] or to compare their method to another similar interpretability method [11, 27]. However, to the best of our knowledge, there has been no work that conducts a cross-method comparison of different types of interpretability methods for computer vision.

Finally, human studies are typically costly and cumbersome to conduct at scale.

Our contributions. In our work, we tackle these challenges head-on and propose a novel human evaluation framework: HIVE (Human Interpretability of Visual Explanations). First, we design a study and its associated UIs to fairly and thoroughly evaluate multiple computer vision interpretability methods. As highlighted in [31, 39], model interpretability can have multiple meanings, from understanding how a whole model (or its component parts) works to being able to simulate a model's decision. We are primarily concerned with the use of models to aid human decision making; thus, we measure interpretability through participants' 1) level of *agreement* with provided explanations and 2) ability to *distinguish* between correct and incorrect predictions based on the provided explanations (see Fig. 1).

Second, to demonstrate the extensibility and applicability of HIVE, we conduct IRB-approved human studies and evaluate four existing computer vision interpretability methods that represent different streams of interpretability work (e.g., post-hoc explanations, interpretable-by-design models, heatmaps, and prototype-based explanations): GradCAM [50], BagNet [9], ProtoPNet [12], ProtoTree [41]. To the best of our knowledge, there have been no human studies for the studied interpretable-by-design models and no studies with the proposed tasks, likely due to the high complexity of the produced explanations. Through our experiments, we demonstrate a number of key findings:

- Participants struggle to distinguish between correct and incorrect explanations for all four methods. This suggests that interpretability works need to improve and evaluate their ability to identify and explain model errors.
- Participants tend to believe explanations are correct (regardless of if they are actually correct) revealing an issue of *confirmation bias* in interpretability work. Prior works have made similar observations for non-visual interpretability methods [39, 44]; however, we substantiate them for visual explanations and demonstrate a need for falsifiable explanations in computer vision.
- We quantify prior work's [25, 41] anecdotal observation that similarity of prototypes in prototype-based models are not consistent with human similarity judgements.
- Participants prefer a model with an explanation over a baseline model. Before switching their preference, they require a baseline model to have higher accuracy (and by a greater margin for higher-risk settings).

Lastly, to help mitigate costs and enable reproducible and extensible research, we open-source HIVE for future evaluation of novel methods.[1] Snapshots of our user interface can be found in the Appendix which starts on page 12.

2. Related work

Interpretability landscape. Interpretability research can be described in several ways: first, whether a method is post-hoc or interpretable-by-design; second, whether it is global or local; and third, the form of an explanation (see [4, 10, 13, 15, 20, 22, 46, 48] for surveys). *Post-hoc explanations* focus on explaining predictions made by already-trained models, whereas *interpretable-by-design* models are intentionally designed to possess a more explicitly interpretable decision-making process [9, 12, 30, 41]. Furthermore, explanations can either be *local explanations* of a single input-output example or *global explanations* of a network (or its component parts). Local, post-hoc methods include heatmap [16, 42, 50–52, 58, 63], counterfactual explanation [21], approximation [45], and sample importance [29, 56] methods. In contrast, global, post-hoc methods aim to understand global properties of CNNs, often by treating them as an object of scientific study [6, 8, 18, 28]. Because we focus on evaluating interpretability in AI-assisted decision making scenarios, we do not evaluate global, post-hoc methods. Furthermore, interpretable-by-design (ibd) models can also provide local and/or global explanations. Lastly, explanations can take a variety of forms: two more popular ones we study are *heatmaps* highlighting important image regions and *prototypes* from the training set (i.e., image patches) that form interpretable decisions. In our work, we investigate four popular methods that span these types of interpretability work: GradCAM [50] (post-hoc, heatmap), BagNet [9] (ibd, heatmap), ProtoPNet [12] (ibd, prototypes), and ProtoTree [41] (ibd, prototypes).

Evaluating heatmaps. Heatmap methods are arguably the well-developed class of interpretability work. However, there is a lack of consensus on how to evaluate these methods. Several metrics have been proposed [5, 17, 26, 42, 43, 55, 59]. In particular, the authors of [1, 2] and BAM [55]

[1] https://princetonvisualai.github.io/HIVE

highlight how several methods fail basic "sanity checks" such as being specific to model parameters and robust to spurious correlations, thereby signaling that more comprehensive metrics are much needed.

Evaluating interpretable-by-design models. In contrast, there has been relatively little work on assessing interpretable-by-design models. Quantitative evaluations of these methods typically focus on demonstrating their competitive performance with a baseline CNN, while the quality of their interpretability is often demonstrated through qualitative examples. Recently, a few works revisited several methods' interpretability claims. Hoffmann et al. [25] highlight that prototype similarity of ProtoPNet [12] does not correspond to semantic similarity and that this disconnect can be exploited. Margeloiu et al. [40] analyze concept bottleneck models [30] and demonstrate that learned concepts fail to correspond to real-world, semantic concepts.

Evaluating interpretability with human studies. Previous human studies for computer vision interpretability methods have been limited in scope: They typically ask participants which explanation method they find more reasonable (e.g., GradCAM [50] vs. IBD [64]) or which model they find more trustworthy based on explanations (e.g., AlexNet vs. VGG16 based on GradCAM). Recent work by Shitole et al. [51] compares heatmap-based explanations through counterfactual questions, asking users to reason about how a model will classify two different occluded versions of an image. To our knowledge, human studies have not been conducted for interpretable-by-design models. More similar to our work are human studies [33, 34, 36, 44, 62] conducted for models trained on tabular datasets. Lage et al. [33] evaluate interpretability of decision sets by asking participants to simulate a model's behavior when varying its complexity and measuring task time and accuracy as proxies for interpretability. Poursabzi-Sangdeh et al. [44] also conduct human studies to test the interpretability of linear regression models by measuring how much participants trust a model and if they can simulate it. However, these studies do not scale to the complexity of modern computer vision models.

In our work, we introduce HIVE, a human evaluation framework that can be used to evaluate a variety of interpretability methods (e.g., post-hoc, interpretable-by-design, heatmaps, prototype explanations) for computer vision, in a similar spirit to work by Zhou et al. [65] on evaluating generative models.

3. Evaluation framework

In this section, we describe HIVE by laying out two desiderata of explanations used to assist human decision making (Sec. 3.1). We then design tasks to evaluate these two desiderata (Secs. 3.2 and 3.3), adapting them for the 4 different methods we study.

3.1. Tasks for objective evaluation

We are primarily concerned with AI-assisted decision making scenarios where humans use AI models and interpretability techniques to make decisions. In such scenarios, the model presents a user with a suggested decision or prediction, along with an explanation for why this decision is selected. The goal of the explanation is to provide insight into the model, to both instill trust but also to help the user identify if the model is making a error. Prior work suggests that subjective self-reported trust may not be a reliable indicator of trust [32, 49]. Hence, we follow the recent line of works that measure users' trust in a model with behavioral indicators [35, 44, 57, 61]. We design tasks to allow for more objective evaluation and testing of falsifiable hypotheses. Concretely, we identify two desirable qualities of explanations in assisting human decision making, and design corresponding tasks to evaluate these properties.

Distinction. First, explanations should allow users to distinguish between correct and incorrect predictions. If explanations are always convincing, even for incorrect predictions, they yield little practical value in assisting human decision making. In our first task, a participant is shown an image along with several explanations corresponding to different possible outputs (e.g., different class labels in object recognition) and the goal is to distinguish the correct prediction from incorrect ones based on provided explanations. One challenge is that the task needs to be complex (for example, bird species classification [53]), forcing the user to rely on provided explanations rather than prior knowledge. To quantify the distinction task, we measure the accuracy of selecting the correct prediction out of a set of possible predictions produced by a visual recognition model.

Agreement. Second, the explanation should be understandable to users. For this task, given an input image, output prediction, and explanation, the participant is asked whether they agree or disagree with the explanation for the input and output pair. This is a more subjective task than distinction, but allows us to dive deeper into the inner-workings of the interpretability methods and obtain complementary insights (e.g., if our participants are not able to successfully perform the first task, this allows us to analyze *why*). To evaluate the agreement task, we quantify which parts of an explanation a participant agrees and disagrees with, and measure the participants' level of confidence in the model's prediction.

3.2. Evaluating distinction

Our key goal is to develop a framework for quantitative evaluation that is applicable and extensible across a range of interpretability methods. In doing so, we consider four quite distinct interpretability methods: GradCAM [50], BagNet [9], ProtoPNet [12] and ProtoTree [41]. We briefly

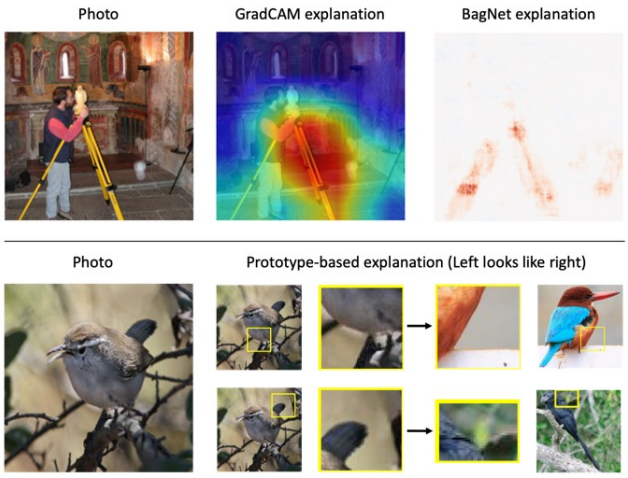

Figure 2. **Forms of explanation.** Top: Heatmap explanations by GradCAM [50] and BagNet [9] highlight decision-relevant image regions. Bottom: Prototype-based explanations by ProtoPNet [12] and ProtoTree [41] match image regions to canonical prototypes.

describe the four methods along with minor adaptations to fit within the framework. This adaptation is unavoidable given the range of the methods, but we prioritize two goals: (1) present each method as favorably as possible, (2) compare between methods fairly via performance metrics.

GradCAM [50]. GradCAM is a post-hoc explanation method that produces a heatmap for a given prediction. This heatmap highlights important regions in the image that contribute to the model's prediction. Note that these heatmaps can be generated for any class, not only for the model's predicted class or the ground-truth class. See Fig. 2 (top center) for an example GradCAM heatmap. In that example, just from looking at the heatmap, the user may infer that the model is recognizing the object "tripod" or (perhaps less likely) the action "taking a photo."

In the distinction task, the participants see an image and four heatmaps that serve as explanations for four potential class predictions. They are asked to (1) identify which heatmap corresponds to the *correct* prediction, and (2) identify which heatmap corresponds to the model's *output prediction*. A simplified UI is shown in Fig. 3 (left); the full version is in supp. mat.

We run two versions of this study: one with the semantic class labels provided, and one with the class labels replaced with "class 1," "class 2," etc. The task without the labels is significantly harder (and at times ambiguous); however, having the class labels may implicitly bias participants to value heatmaps with better localization properties and rely more on their prior knowledge. Thus, we run both studies on the same set of images and compare the results.

BagNet [9]. The BagNet model is similar in output to GradCAM, but recognizes an object by collecting class evidence from small regions of an image. For each class, BagNet creates a heatmap where higher values, shown as darker red in our visualizations, imply higher evidence for the class. BagNet then sums the values in the heatmap and chooses the class with the highest value as its prediction. See Fig. 2 (top right) for an example BagNet heatmap. We follow the same evaluation protocol for BagNet as for GradCAM.

ProtoPNet [12]. The next two methods reason with *prototypes* rather than heatmaps. Prototypes are small regions learned from the training set that these models deem as representative for certain classes. Given a test image, ProtoPNet [12] compares it to the set of learned prototypes, finding regions in the test image that are closest to each prototype. The model computes a similarity score between each prototype and region pair, then predicts the class with the highest weighted sum of the similarity scores. See Fig. 2 (bottom) for a schematic of prototype-based explanations.

ProtoPNet's approach to interpretability is completely different from GradCAM or BagNet, but it nevertheless fits into our framework. Given the model's explanations from the top four predicted classes of an image, participants select the class they think is correct; see Fig. 3 (center) for a simplfied UI. As a sub-task, for each image, participants rate how similar each prototype is to the matched region. While the region-prototype pairs in ProtoPNet explanations are presented in order from highest to lowest similarity, we randomize the pairs as to not skew the participant's rating. The design of this task balances simulating the basic concepts underlying ProtoPNet's decision making while abstracting away many complexities such the similarity scores and class weights used by the model to make its prediction.

ProtoTree [41]. Finally, the ProtoTree model learns a tree structure along with the prototypes. Each node in the tree contains a prototype which comes from a training image. At each node, the model compares a given test image to the node's prototype and produces a similarity score. If the score is above a threshold, the model judges that the prototype is present in the image and absent if not. The model then proceeds to the next node and repeats this process until it reaches a leaf node, which corresponds to a class.

The full ProtoTree explanation can have over hundreds of internal nodes, and our initial (even internal) studies revealed that is too overwhelming for any user. Thus, in our study we significantly simplify the decision process. Participants are shown the model's decisions until the penultimate decision node, and then are asked to make decisions for only the final two nodes of the tree by judging whether the prototype in each node is "present" or "absent" in the image. Afterwards, participants select which of the four (2^2) classes their decision leads to. One additional challenge is that participants may not be familiar with decision trees and thus may have trouble understanding the explanation. To help mitigate this, we introduce a simple decision tree model for

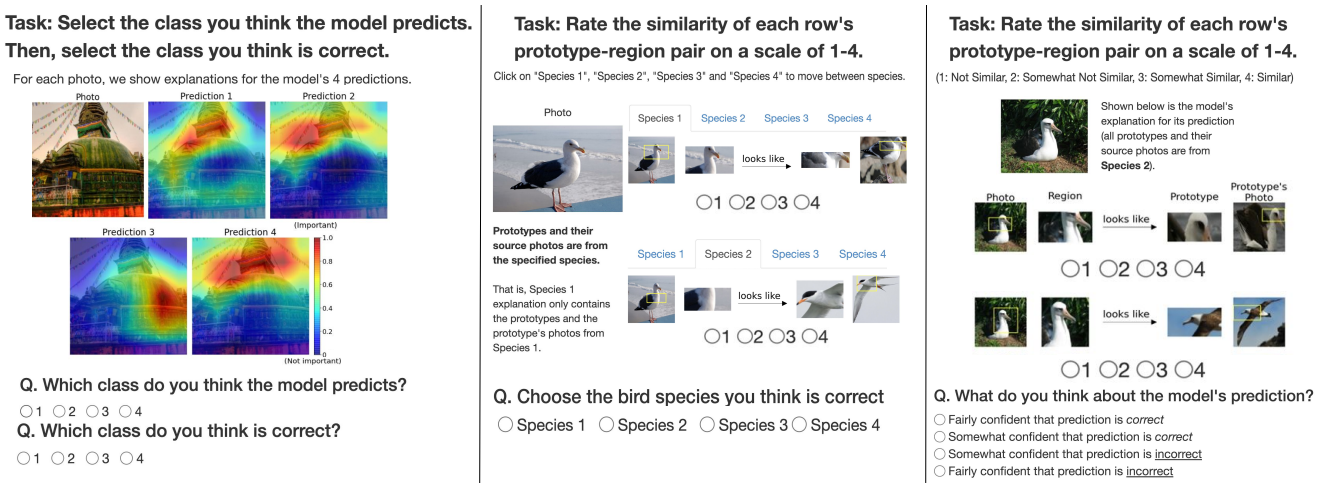

Figure 3. **User Interface (UI) snapshots.** We evaluate methods on two tasks: *distinction* (can you distinguish between correct and incorrect explanations?) and *agreement* (do you agree with the explanation?). We show the distinction task for GradCAM (left) and ProtoPNet (center) and the agreement task for ProtoPNet (right).

fruit classification before introducing ProtoTree. We provide an example and present two warm up exercises so that participants can get familiar with this type of model. See supp. mat. for more information.

3.3. Evaluating agreement

The agreement task is designed following similar protocols as the distinction task. In each task, participants receive 5 correct and 5 incorrect predictions of the studied model. Given that participants do not know which of the 10 samples correspond to a correct or incorrect prediction, they rate their confidence in the model's correctness of each prediction. We do not evaluate GradCAM and BagNet on this task because we are not trying to test whether their explanation match human intuition, but rather whether it is understandable to humans. One interpretability objective of heatmaps is to faithfully explain model behavior. The agreement test is in conflict with that goal; thus, we only evaluate GradCAM and BagNet on the distinction task. For ProtoPNet and ProtoTree, participants rate the similarity of each prototype-region pair. These sub-tasks encourage and ensure that participants look through each component of the model's decision process. We show a simplified agreement task UI we used to evaluate ProtoPNet in Fig. 3 (right).

4. Study design

Having presented the evaluation tasks, we next describe our study design. As our study involves humans, we received approval from the Institutional Review Board (IRB) prior to conducting the study. We recommend that future researchers follow similar protocols.

Introduction. For each participant, we first briefly introduce the study and receive their informed consent. To help future researchers calibrate our results and do proper comparison, we request optional demographic data regarding gender identity, race and ethnicity, and the participant's experience with machine learning.

To mitigate the difficulty of understanding complex visual explanations, we explain the model interpretability method in simple terms, actively avoiding using technical jargon and even replacing terms such as "image" and "training set" to "photo" and "previously-seen photos." To encourage participants to carefully read the method explanation, we also show a preview of the task they will complete. Additionally, we provide example explanations for one correct and one incorrect prediction made by the model to give the participant appropriate references. The participant can access the method explanation at any point during the task.

Tasks. For GradCAM and BagNet, participants are shown 10 sample images: 5 samples where the model made a correct prediction and 5 where the model made an incorrect prediction. Each sample image is associated with four heatmaps, as described in Sec. 3.2. On samples where the classification model makes correct predictions, the four heatmaps we show correspond to the top-4 predicted classes, in random order. For the incorrectly predicted samples, we show heatmaps for the top-3 predicted classes as well as the heatmap of the ground-truth class. For ProtoPNet and ProtoTree, we run two tasks each, one for distinction, and another for agreement. For the agreement tasks, participants are shown 10 sample images, similar to the studies run on GradCAM and BagNet, along with the prototype explanations for each. For the distinction tasks, the set up is as follows. For ProtoPNet the entire study consists of 4 images on which the model has made a correct prediction, along with 4 prototype explanations for each photo.

For ProtoTree, participants are shown 10 photos on which the model has made a correct prediction. Their decisions on the two final nodes lead to 4 different classes.

Subjective evaluation. In addition to the quantitative tasks, we also ask a subjective evaluation question three times. We ask the participant to self-rate their level of understanding of the model before and after completing the task, to investigate if the participant's self-rated level of understanding undergoes any changes during the task. After the task completion, we disclose the participant's performance on the task and ask the question one last time.

Interpretability-accuracy tradeoff. We also investigate the *interpretability-accuracy tradeoff* the participant is willing to make when comparing the interpretable method against a baseline model that doesn't come with any explanation. While interpretability methods offer useful insight into a model's decision, their explanations often come at the cost of lower model accuracy. In high-risk scenarios a user may prefer to maximize model performance over interpretability for improved accuracy. However, another user may prefer to prioritize interpretability in such settings so that there would be mechanisms for examine the model's prediction. To gain insight into the interpretability-accuracy tradeoff users are willing to make, we present three scenarios to the participant: low-risk (e.g., scientific or educational purposes), medium-risk (e.g., object recognition for automatic grocery checkout), and high-risk (e.g., scene understanding for self-driving cars). For each scenario, we ask the participant to input the minimum accuracy of a baseline model that would convince them to use the baseline model over the model that comes with explanations.

Rating scales. Participants rate the confidence in their answer or their understanding of the model on a 5-point Likert scale [38] (1: very poor, 2: poor, 3: fair, 4: good, 5: very good) allowing for degrees of opinion to be measured. For ProtoPNet and ProtoTree, participants rate the similarity of prototype-region pairs using a 4-point Likert scale (1: not similar, 2: somewhat not similar, 3: somewhat similar, 4: similar) to ensure participants form an opinion. In the ProtoPNet and ProtoTree agreement task, participants rate their confidence in the given prediction on a 4-point scale (1: confident prediction is incorrect, 2: somewhat confident prediction is incorrect, 3: somewhat confident prediction is correct, 4: confident prediction is correct).

5. Experiments

5.1. Experimental details

Human studies. We ran our study through Human Intelligence Tasks (HITs) deployed on Amazon Mechanical Turk (AMT). We recruited participants who are US-based, have done over 1000 HITs, and have prior approval rate of at least 98%. The demographic distribution was: woman 46%, man 51%, no gender reported 1%, non-binary 2%; White 75%, no race/ethnicity reported 7%, Black/African American 7%, Asian 8%, American Indian/Alaska Native 1%, Hispanic/Latino/Spanish Origin of any race 3%, Native Hawaiian/Other Pacific Islander 0%. The self-reported machine learning experience was 2.6 ± 1.1, between "2: have heard about..." and "3: know the basics..." No personally identifiable information was collected.

For each study, we deployed 10 HITs, each with a different set of input images. To reduce the variance with respect to the input, we had 5 participants complete each HIT, so each study had 50 participants. The average study duration was 7 and 7.4 minutes for GradCAM and BagNet, 10.9 and 16.4 minutes for ProtoPNet distinction and agreement, 10.8 and 10.1 minutes for ProtoTree distinction and agreement. Participants were compensated based on the state-level minimum wage of $12/hr.

Datasets and Models. We evaluate all four methods on classification tasks and use images from the ImageNet [47] validation set for GradCAM and BagNet and the CUB [53] test set for ProtoPNet and ProtoTree. For ProtoPNet, we use the model trained by Hoffman et al. [25]; for the other interpretable-by-design models, we use models provided by their respective authors. Specifically, we evaluate BagNet33, a ResNet34-based ProtoPNet model, a ResNet50-based ProtoTree model, and GradCAM explanations for ResNet50. We also made a few, small, modifications to the presentation of original explanations to make them more suitable for our tasks—e.g., we asked participants to rate the similarity of prototypes on a scale from 1-4 instead of suggesting a precise similarity score (the tasks were too challenging in initial studies without these modifications). See supp. mat. for details.

Statistical analysis. For each study, we report the mean accuracy and standard deviation. We also compare the study result to random chance and compute the p-value from a 1-sample t-test. When comparing results between two groups, we compute the p-value from a 2-sample t-test. Results are deemed statistically significant under $p < 0.001$ conditions.

5.2. Objective assessment

We summarize our key results in Tab. 1.

Participants struggle to distinguish between correct and incorrect predictions. In the distinction task (middle column in Tab. 1), participants achieve a mean accuracy of 27.8% on GradCAM, 42.0% on BagNet, 54.5% on ProtoPNet and 33.8% on ProtoTree. While participants' performance is statistically significantly above random chance for BagNet, ProtoPNet, and ProtoTree, these numbers are far from perfect accuracy and suggest that there is a gap to be filled for these interpretability methods to be reliably useful

Method	Distinction	Output prediction
GradCAM [50]	*27.8 ± 16.3*	34.6 ± 19.0
BagNet [9]	**42.0 ± 15.7**	**56.0 ± 16.4**
Method	Distinction	Agreement
ProtoPNet [12]	**54.5 ± 30.3**	**60.0 ± 18.3**
ProtoTree [41]	33.8 ± 15.9	*53.6 ± 15.2*

Table 1. **Key results.** These results suggest that interpretability methods should be improved (to be closer to 100% accuracy) before they can be reliably used to aid decision making. For each study, we report the mean accuracy and standard deviation (random chance is 25% for distinction and output prediction and 50% for agreement). *Italics* denotes methods that do not statistically significantly outperform random chance ($p > 0.001$); **bold** denotes the highest performing method.

in AI-assisted decision making scenarios.

For GradCAM and BagNet, participants are more likely to identify the class the model predicts than the ground-truth class. For GradCAM, the participants achieve 34.6% accuracy is identifying the model's output prediction but only 27.8% in identifying the correct class. For BagNet, the participants achieve 56.0% accuracy in identifying the model's output prediction but only 42.0% in identifying the correct class, which is statistically significantly different. These results are not surprising as these methods were designed to give insights into how the model arrives at its final prediction, rather than flagging difficult samples on which the model will likely fail.

Labels vs. no labels for BagNet and GradCAM. In addition to running the GradCAM and BagNet studies without providing class labels, we run studies in which we provide class labels in order to evaluate whether the heatmaps alone are sufficient or if class labels are needed.[2] We find that without class labels, participants perform much better on the task for BagNet compared to GradCAM: 42.0% for BagNet vs 27.8% for GradCAM on the distinction task and 56.0% for BagNet vs 34.6% for GradCAM on output prediction. Surprisingly, we observed the *opposite* trend when participants have class labels. On both tasks, participants perform better with GradCAM explanations (45.4% over 32.0% on distinction, 53.2% over 36.8% on output prediction). The performance across the two methods are statistically significantly different. This suggests that when participants *know* what they are looking for, GradCAM provides a more intuitive heatmap of the model's reasoning process; however, in the absence of this additional information, BagNet's heatmap is actually more informative because, by design, it conveys the model's confidence for a given class.

Participants perform relatively poorly on ProtoTree, but they understand how a decision tree works. While the previously described ProtoTree agreement study does not take into account the model's inherent tree structure, we run another version of the agreement study where, instead of asking participants to rate each prototype's similarity, we ask them to select the first step they disagree with in the model's explanation. Performance on this task (52.8 ± 19.9) is similar to that of the original study (53.6 ± 15.2); in both cases, we cannot conclude that participants outperform 50% random chance ($p = 0.33$, $p = 0.10$). To ensure participants understand how decision trees work, we provided a simple decision tree example and subsequent questions asking participants if the decision tree example makes a correct or incorrect prediction. Participants achieved 86.5% performance on this task, implying that the low agreement accuracy is not due to a lack of comprehension of decision trees.

Participants are consistent in their similarity ratings and decisions. When examining ProtoPNet and ProtoTree explanations, on average, participants assign higher similarity ratings to prototypes of the species they select to be correct (2.9 out of 4 for both ProtoPNet agreement and distinction tasks, 2.4 for ProtoTree agreement) and lower similarity ratings to prototypes of the species they select to be incorrect (2.0 and 2.1 for ProtoPNet agreement and distinction, 2.0 for ProtoTree agreement). The similarity ratings between the two groups are statistically significantly different in all studies. This suggests that participants understand how the model reasons (i.e., they predict the bird species whose prototypes appear most similar to the given photo).

A gap exists between similarity ratings of ProtoPNet & ProtoTree and those of users. Prior work [25, 41] has pointed out that prototype-based models' notion of similarity may not align with that of users. We empirically confirm this observation. For ProtoTree, we calculate the correlation between the participants' prototype similarity ratings and the model similarity scores. The Pearson correlation coefficient is 0.06, suggesting little to no relationship between the model and participants' notion of similarity.

The ProtoPNet's prototype similarity scores are not nor-

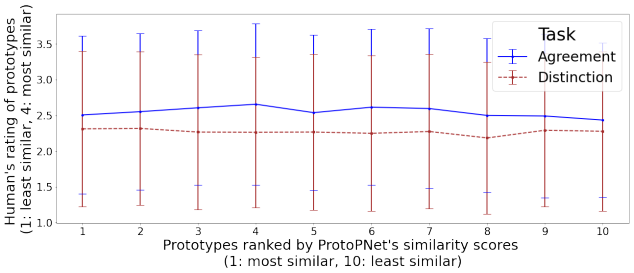

Figure 4. **Participant vs. ProtoPNet prototype similarity rating.** There exists a gap between ProtoPNet's similarity scores and human judgments of similarity (Spearman's $\rho = -0.25$, $p = 0.49$ for distinction; $\rho = -0.52$, $p = 0.12$ for agreement).

[2]Recall that for these models we are using the ImageNet data, with common objects that the participants may be familiar with, rather than CUB data with bird species that require domain expertise to recognize.

malized across images, so we compute the Spearman's rank correlation coefficient ($\rho = -0.25, p = 0.49$ for distinction and $\rho = -0.52$, $p = 0.12$ for agreement) and compare the rankings of prototypes when ordered by participants' mean ratings vs. by ProtoPNet's similarity scores (see Fig. 4). There is no significant negative correlation between the two for either the agreement or distinction tasks suggesting a gap between ProtoPNet and human judgments of similarity.

5.3. Subjective evaluation

When given an explanation for a model prediction, participants often lean towards believing that the prediction is correct. Participants indicate agreement on a scale of 1-4 on a 4-point Likert scale; these ratings are then binarized with 1-2 indicating disagreement and 3-4 signaling agreement. Participants achieve a mean accuracy of 60% on ProtoPNet and 54.5% on ProtoTree, with only ProtoPNet's result being statistically significantly different from the 50% random chance. Interestingly, participants' confidence that the model produces a correct prediction is statistically different from the neutral 2.5 rating, with a mean confidence on all samples of 2.7 for ProtoPNet and ProtoTree. For ProtoPNet, the mean confidence is 3.0 for correct and 2.4 for incorrect predictions; for ProtoTree, it is 2.8 for correct and similarly 2.7 for incorrect predictions. All mean ratings are statistically different from the neutral rating except the 2.4 for incorrect ProtoPNet predictions. This suggests a small but significant confirmation bias, as participants tend to believe that a prediction is correct when given an explanation.

To prefer a baseline model over model with explanations, participants require the baseline model to have higher accuracy and by a greater margin in higher-risk settings. In the final part of our studies, we asked the participants for the minimum accuracy of a baseline model that they would require in order to prefer using it over a model with explanations for its predictions. Across all studies, participants require the baseline model to have a higher accuracy than the evaluated model and by a greater margin for higher-risk settings. See Fig. 5 for full results and supp. mat. for the participants' self-described reasons.

Participants' self-rated level of model understanding decreases after they see their task performance. We asked the participants to self-rate their level of model understanding three times. The average ratings are 3.8 ± 0.9 after the method explanation, 3.8 ± 0.9 after the task, and 3.6 ± 1.0 after seeing their task performance, which all lie between the fair (3) and good (4) ratings. Interestingly, the rating tends to *decrease* after the participants see their task performance ($p = 0.006$). This trend suggests that participants find their performance to be lower than what they expected, which in turn leads them to lower their reported level of understanding. Full results can be found in supp. mat.

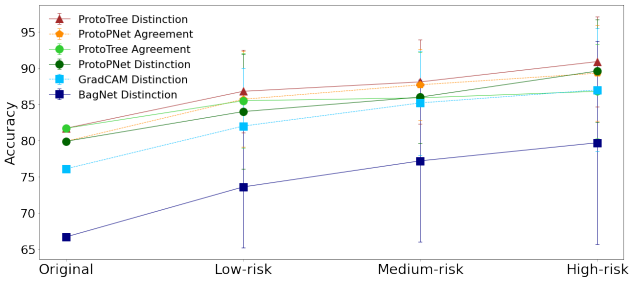

Figure 5. **Interpretability-accuracy tradeoff.** We show the evaluated model's original accuracy and the minimum accuracy of a baseline model that participants require in order to use it over the model with explanations under different risk settings. This plot shows that participants desire higher accuracies for the baseline model, especially in higher-risk settings.

6. Conclusion

In summary, we propose HIVE, a novel human evaluation framework for evaluating interpretability methods for computer vision. Our framework is composed of two tasks, the *distinction* and *agreement* tasks, which we use to evaluate four methods: Grad-CAM, BagNet, ProtoPNet, and ProtoTree. We also open-source our evaluation framework so future researchers can use it to evaluate their own methods.

There are a few limitations and potential negative impacts of our work: First, we use a relatively small sample size of 50 participants for each study, due to the costs associated with human studies and our desire to evaluate four methods. Second, our participants have limited background in machine learning; this is slightly at odds with the ideal scenario of studying how ML practitioners and/or domain experts make decisions about using one method over another (e.g., how would bird experts evaluate interpretable models trained for bird species recognition). Third, future applications of our work could misuse our framework (even unintentionally) and design experiments favorable for their method. Thus, these studies must be designed and conducted with care and relevant approval (e.g., IRB).

Nonetheless, we hope this work facilitates more user studies and encourages the field to take a more human-centered approach in interpretability research, as our evaluation reveal several key insights about the field. In particular, we find that users generally believe presented explanations are correct. Humans are naturally susceptible to confirmation bias; thus, interpretable explanations will likely engender trust from humans, even if they are incorrect. Our work underscores the need for *falsifiable* explanations and an evaluation framework that evaluates the desiderata of interpretable methods fairly. We hope our framework for human evaluation helps shift the field's objective from focusing on method development to also prioritizing the development of high-quality evaluation metrics.

Positionality statement. All five authors primarily work in the fields of computer vision and machine learning. SK, NM, VR, and OR have not previously conducted interpretability research. RF has, but we did not evaluate RF's works for impartial comparison of the studied methods.

Acknowledgments. This work is supported by the National Science Foundation Grant No. 1763642 to OR, the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award to OR, and the Princeton SEAS and ECE Senior Thesis Funding to NM. We thank the authors of [9, 12, 25, 41, 50] for open-sourcing their code and the authors of [9,24,25,41] for sharing their trained models. We also thank the AMT workers who participated in our studies, as well as the Princeton Visual AI Lab members (Dora Zhao, Kaiyu Yang, Angelina Wang, and others) who tested our user interface and provided helpful feedback.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018. 1, 2

[2] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In *NeurIPS*, 2020. 2

[3] Chirag Agarwal and Sara Hooker. Estimating example difficulty using variance of gradients. In *ICML Workshop on Human Interpretability in Machine Learning*, 2020. 1

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 2020. 2

[5] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 2015. 2

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017. 1, 2

[7] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *PNAS*, 2020. 2

[8] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a GAN cannot generate. In *ICCV*, 2019. 1, 2

[9] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *ICLR*, 2019. 1, 2, 3, 4, 7, 9, 12

[10] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward trustworthy AI development: Mechanisms for supporting verifiable claims, 2020. 2

[11] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *TPAMI*, 2019. 2

[12] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*, 2019. 1, 2, 3, 4, 7, 9, 12, 13

[13] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Towards connecting use cases and methods in interpretable machine learning. In *ICML Workshop on Human Interpretability in Machine Learning*, 2021. 2

[14] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. 2

[15] Ruth Fong. *Understanding convolutional neural networks*. PhD thesis, University of Oxford, 2020. 2

[16] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, 2019. 1, 2

[17] Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017. 1, 2

[18] Ruth Fong and Andrea Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *CVPR*, 2018. 1, 2

[19] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 12

[20] Leilani H. Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *DSAA*, 2018. 2

[21] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *ICML*, 2019. 2

[22] David Gunning and David Aha. Darpa's explainable artificial intelligence (XAI) program. *AI Magazine*, 2019. 2

[23] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *ACL*, 2020. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 9, 12

[25] Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. In *ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*, 2021. 1, 2, 3, 6, 7, 9, 12

[26] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, 2019. 1, 2

[27] Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. In *NeurIPS*, 2020. 2

[28] Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C. Mozer. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, 2021. 2

[29] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, 2017. 2

[30] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, 2020. 1, 2, 3

[31] Maya Krishnan. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 2020. 2

[32] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *CHI*, 2019. 3

[33] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *HCOMP*, 2019. 3

[34] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. *NeurIPS*, 2018. 3

[35] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *FAccT*, 2019. 3

[36] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *KDD*, 2016. 3

[37] Matthew L. Leavitt and Ari S. Morcos. Towards falsifiable interpretability research. In *NeurIPS Workshop on ML Retrospectives, Surveys & Meta-Analyses*, 2020. 1

[38] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. 6

[39] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 2018. 2

[40] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? In *ICLR Workshop on Responsible AI*, 2021. 1, 3

[41] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *CVPR*, 2021. 1, 2, 3, 4, 7, 9, 12, 13

[42] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *BMVC*, 2018. 1, 2

[43] Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In *CVPR Workshop on Responsible Computer Vision*, 2021. 1, 2

[44] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *CHI*, 2021. 2, 3

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*, 2016. 2

[46] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. In *Statistics Surveys*, 2021. 2

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 6, 12

[48] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer, 2019. 2

[49] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your ai: Expertise and explanations. In *IUI*, 2019. 3

[50] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 1, 2, 3, 4, 7, 9, 12

[51] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. One explanation is not enough: Structured attention graphs for image classification. In *NeurIPS*, 2021. 2, 3

[52] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014. 1, 2

[53] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 3, 6, 12

[54] Pei Wang and Nuno Vasconcelos. Towards realistic predictors. In *ECCV*, 2018. 1

[55] Mengjiao Yang and Been Kim. Benchmarking Attribution Methods with Relative Feature Importance, 2019. 1, 2

[56] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *NeurIPS*, 2018. 2

[57] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *CHI*, 2019. 3

[58] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1, 2

[59] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016. 1, 2

[60] Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. Predicting failures of vision systems. In *CVPR*, 2014. 1

[61] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *FAccT*, 2020. 3

[62] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect on confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *FAccT*, 2020. 3

[63] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 2

[64] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *ECCV*, 2018. 2, 3

[65] Sharon Zhou, Mitchell L Gordon, Ranjay Krishna, Austin Narcomey, Li Fei-Fei, and Michael S Bernstein. HYPE: A benchmark for human eye perceptual evaluation of generative models. In *NeurIPS*, 2019. 3

Appendix

In this supplementary document, we provide additional details on certain sections of the main paper.

Section A: We provide additional details on the four studied methods and the modifications we made to their original explanation form.

Section B: We provide the full subjective evaluation results.

Section C: We describe the simple decision tree model for fruit classification we introduced in the ProtoTree studies.

Section D: We show snapshots of our user interfaces.

A. Details on the studied methods

GradCAM [50]. For GradCAM, we used the ResNet50 [24] model in the `torchvision` library which achieves 76.1% accuracy on ImageNet [47] classification. We used the code by Gildenblat et al. [19] to generate GradCAM visualizations.[3] For a given image, we generate GradCAM heatmaps for all 1000 classes, identify the global minimum and maximum, and then normalize the heatmaps into the range [0, 1]. This way, we preserve the intensity difference between heatmaps for different predictions. See Fig. A1 for an example set of GradCAM explanations we show to participants.

BagNet [9]. We used the BagNet33 model trained by the original authors which achieves 66.7% accuracy on ImageNet classification. To generate explanations, we used the authors' code with one small modification.[4] The authors' visualization code normalizes each heatmap individually by clipping the values above the 99th percentile. On the other hand, we normalize all 1000 heatmaps for a given image together so that we preserve the intensity difference. See Fig. A2 for an example set of BagNet explanations we show to participants.

ProtoPNet [12]. For ProtoPNet, we used the ResNet34-based model trained by Hoffmann et al. [25]. For more interpretability, we pruned 331 prototypes from this model. The resulting model has 1669 prototypes and achieves 79.9% accuracy on CUB [53] bird image classification. For generating explanations, we used the code by the original authors with some modifications.[5] In our studies, given an explanation, participants are asked to rate the similarity of each prototype-region pair, then either rate the level of confidence in the prediction's correctness *(agreement)* or select the correct class *(distinction)*. To make ProtoPNet's explanations more suitable for these tasks, we made the following modifications to the original explanation form.

[3] https://github.com/jacobgil/pytorch-grad-cam

[4] https://github.com/wielandbrendel/bag-of-local-features-models

[5] https://github.com/cfchen-duke/ProtoPNet

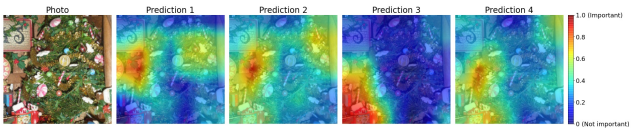

Figure A1. **GradCAM explanations.**

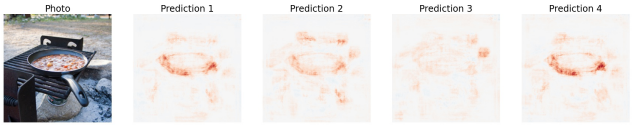

Figure A2. **BagNet explanations.**

- The ProtoPNet model calculates evidence for all classes using the learned prototypes, then predicts the class with the highest evidence. However, we deemed it is unrealistic to ask users to review explanations for all 200 bird classes in CUB. Hence, we only present explanations for one (*agreement*) or four (*distinction*) classes and ask users to examine them.
- The original explanation (Fig. A3 left) shows activation maps, similarity scores, class connection weights, and the total class evidence. In our version (Fig. A3 right) we remove them as we seek to investigate what participants rate as similar and not.
- In the original explanation, prototypes are presented in the order of highest to lowest similarity. In ours, we randomly shuffle the order of prototypes because we don't want to skew their ratings.
- In our explanations, we replaced the bird species name (e.g., Gray Catbird, Cardinal) with Species 1, Species 45, etc. so that participants are not influenced by their prior knowledge.

ProtoTree [41]. For ProtoTree, we used the model trained by the original authors which achieves 81.7% accuracy on CUB bird image classification. This model is a pruned tree of depth 10 and 511 nodes. We used the authors' code to generate explanations with some modifications.[6]

- Same as what we did for ProtoPNet explanations, we removed the similarity scores as we seek to investigate what participants rate as similar and not. We also replaced the bird species name (e.g., Gray Catbird, Cardinal) with Species 1, Species 45, etc. so that participants are not influenced by their prior knowledge.
- For the local explanation, we converted the original horizontal explanation (Fig. A4) into a vertical one (Fig. A5). A vertical explanation is more faithful to how the model reasons, as it starts from the root node and proceeds down the tree until it reaches one of the bottom leaves. Further, it is easier for the participants to examine the explanation by scrolling up and down.

[6] https://github.com/M-Nauta/ProtoTree

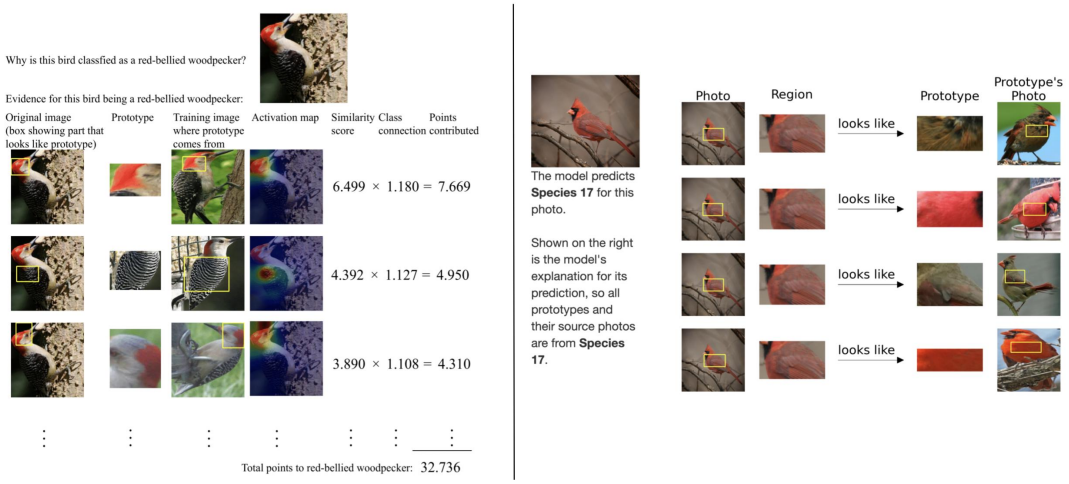

Figure A3. **ProtoPNet original and modified explanations.** The original explanation (left) displayed in Fig. 3 of the original paper [12] contains details such as activation maps, similarity scores, and class connection weights. In our version (right), we remove these to abstract away the complexities and have the participants focus on the main task.

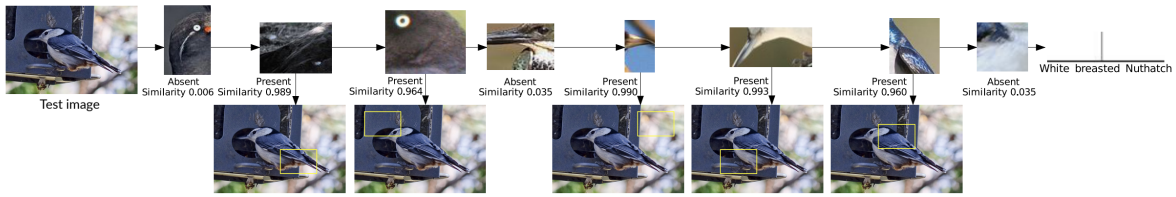

Figure A4. **ProtoTree original explanation.** We show the original explanation displayed in Fig. 9 of the original paper [41]. See Fig. A5 for our modified explanation.

B. Subjective evaluation results

In Sec. 5.3 of the main paper, we discussed our subjective evaluation results. Here we provide the full results.

In Tab. A1 we report the participants' self-rated level of understanding of the given model's reasoning process. Overall, the participants rated their level of understanding between 3 (fair) and 4 (good). As discussed in the main paper, we find that the rating tends to decrease after the participants see their task performance.

In Tab. A2 we show the numbers corresponding to Fig. 5 in the main paper. We show the evaluated model's original accuracy and the minimum accuracy of a baseline model that participants require in order to use it over the model with explanations under different risk settings. Across all studies, we find that participants require the baseline model to have higher accuracy than the evaluated model, and input a higher accuracy requirement for higher-risk settings.

C. Simple decision tree model

One additional challenge of evaluating the ProtoTree model is that participants may not be familiar with decision trees. To mitigate this challenge, we introduce a simple decision tree model for fruit classification before introducing

Study	Post-intro	Post-task	Post-results
GradCAM *distinction*	3.8 ± 0.7	3.9 ± 0.8	3.7 ± 0.9
BagNet *distinction*	3.8 ± 0.8	3.7 ± 0.9	3.5 ± 1.0
ProtoPNet *agreement*	3.9 ± 0.8	4.0 ± 0.8	3.7 ± 0.8
ProtoPNet *distinction*	4.1 ± 0.8	3.9 ± 0.8	3.7 ± 1.1
ProtoTree *agreement*	3.7 ± 0.8	3.7 ± 1.0	3.4 ± 0.8
ProtoTree *distinction*	3.4 ± 1.0	3.6 ± 1.1	3.3 ± 1.2

Table A1. **Subjective evaluation.** We report the mean and standard deviation of the participants' self-rating of their understanding of the given model. Participants provide ratings three times: after reading about the method, after completing the task, and after learning about their task performance.

Study	Orig	Low-risk	Med-risk	High-risk
GradCAM *d*	76.1	82.0 ± 8.0	85.2 ± 7.2	87.0 ± 8.5
BagNet *d*	66.7	73.6 ± 8.4	77.2 ± 11.2	79.7 ± 14.0
ProtoPNet *a*	79.9	85.7 ± 6.6	87.7 ± 4.9	89.3 ± 6.6
ProtoPNet *d*	79.9	84.0 ± 7.9	86.0 ± 6.4	89.6 ± 7.1
ProtoTree *a*	81.7	85.5 ± 6.5	85.9 ± 6.3	86.8 ± 6.5
ProtoTree *d*	81.7	86.8 ± 5.7	88.1 ± 5.8	90.9 ± 6.2

Table A2. **Interpretability-accuracy tradeoff.** We show the evaluated model's original accuracy and the minimum accuracy of a baseline model that participants require in order to use it over the model with explanations under different risk settings. See Fig. 5 in the main paper for a visualization of the results.

Figure A5. **ProtoTree modified explanation.** See Fig. A4 for the original explanation.

ProtoTree. This simple decision tree model takes in an input image and makes an output classification (Class A, B, C, D, E) based on three decision nodes. We first walk through the participants through an example. We then present two warm-up exercises so that the participants can become more familiar with decision trees. When the participants submit their answers, we also provide the correct answer and the reason for it. See Fig. A6 for the UI.

D. User interface

As described in Sec. 1, UI design for evaluating interpretability methods is non-trivial. In addition to outlining our study in Sec. 4, we provide snapshots of our study UIs in the following order.

1. Study introduction. For each participant, we first briefly introduce the study and receive their informed consent. This consent form was approved by the IRB and acknowledges that participation is voluntary, refusal to participate will involve no penalty or loss of benefits, etc. See Fig. A7.

2. Demographics and background. To help future researchers calibrate our results and do proper comparison, we request optional demographic data regarding gender identity, race and ethnicity. We also ask the participant's experience with machine learning. See Fig. A8.

3. Method introduction. We introduce each interpretability method/model in simple terms. See Fig. A9.

4. Task preview and first subjective evaluation. To encourage participants to carefully read the method explanation, we show a preview of the task they will complete along with a correct and incorrect prediction. Participants then answer their first subjective evaluation question. In Fig. A10 we shown an example from the ProtoPNet agreement study.

5. Task. Participants then proceed onto the main task. We show the UI for the following 6 studies:

- GradCAM distinction (Fig. A11)
- Bagnet distinction (Fig. A12)
- ProtoPNet distinction (Fig. A13)
- ProtoPNet agreement (Fig. A14)
- ProtoTree distinction (Fig. A15)
- ProtoTree agreement (Fig. A16)

6. Second and third subjective evaluation. After the task, participants complete their second subjective evaluation question. We then disclose their task performance and ask the third subjective evaluation question. These questions allow us to investigate if the participants' self-rated level of understanding undergoes any changes throughout the study. See Fig. A17.

7. Interpretability-accuracy tradeoff. Finally, we investigate the interpretability-accuracy tradeoff participants are willing to make when comparing the interpretable method against a baseline model that doesn't come with any explanation. We present three scenarios to the participants: low-risk (e.g., scientific or educational purposes), medium-risk (e.g., object recognition for automatic grocery checkout), and high-risk (e.g., scene understanding for self-driving cars). We then ask them to input the minimum accuracy of a baseline model that would convince them to use the baseline model over the model that comes with explanations and briefly describe their reasoning. See Fig. A18.

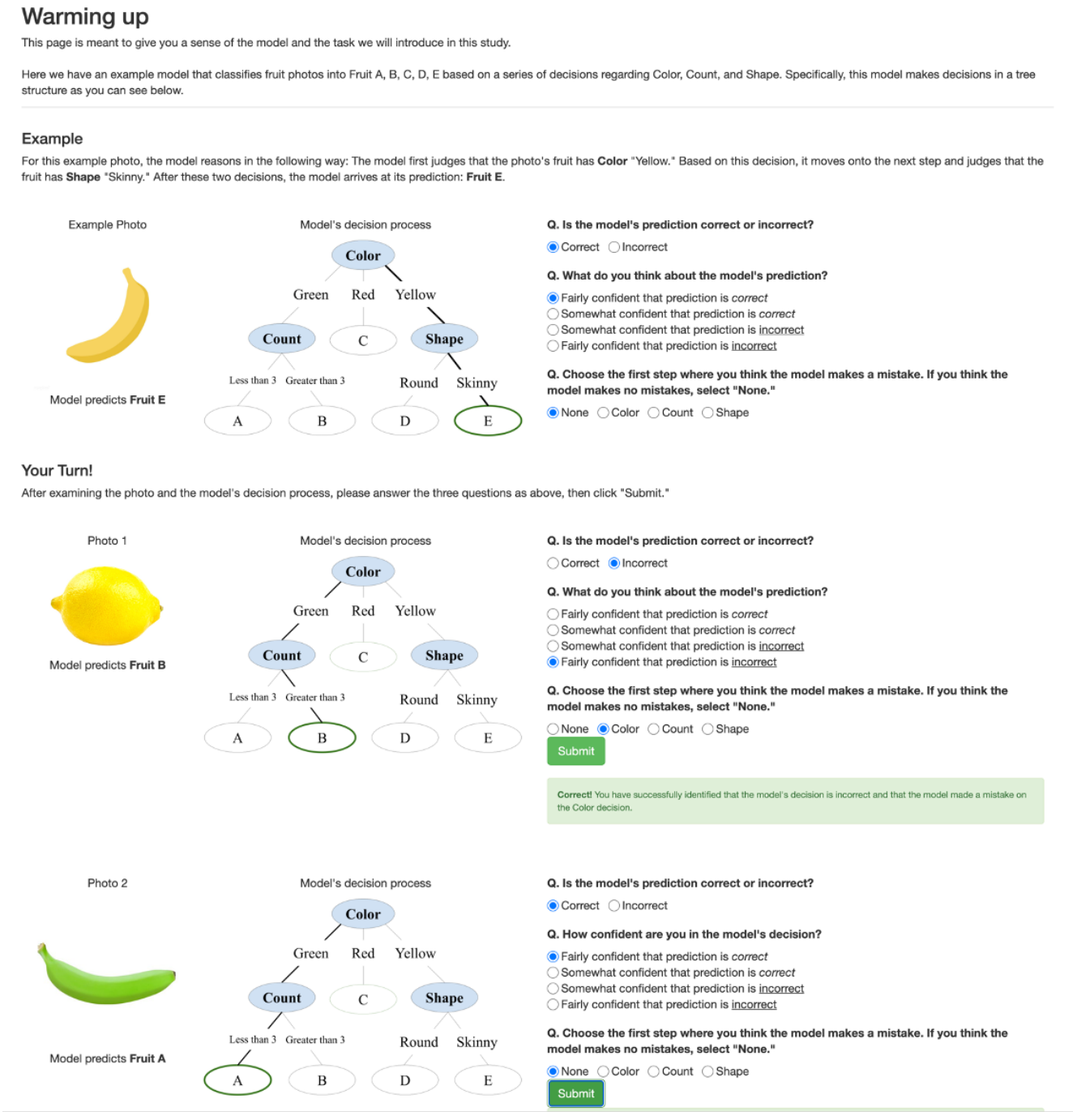

Figure A6. A simple decision tree model for fruit classification. We use this model to introduce participants to decision trees before explaining the more complex ProtoTree model. See Section C for details.

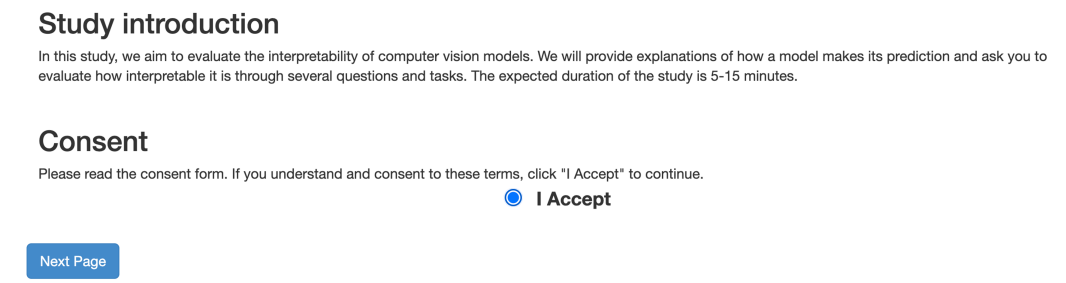

Figure A7. **1. Study introduction.**

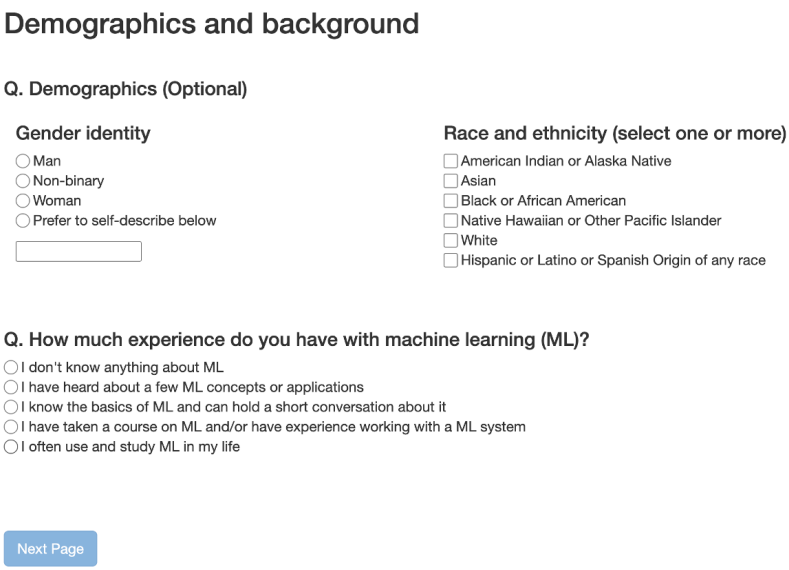

Figure A8. **2. Demographics and background.**

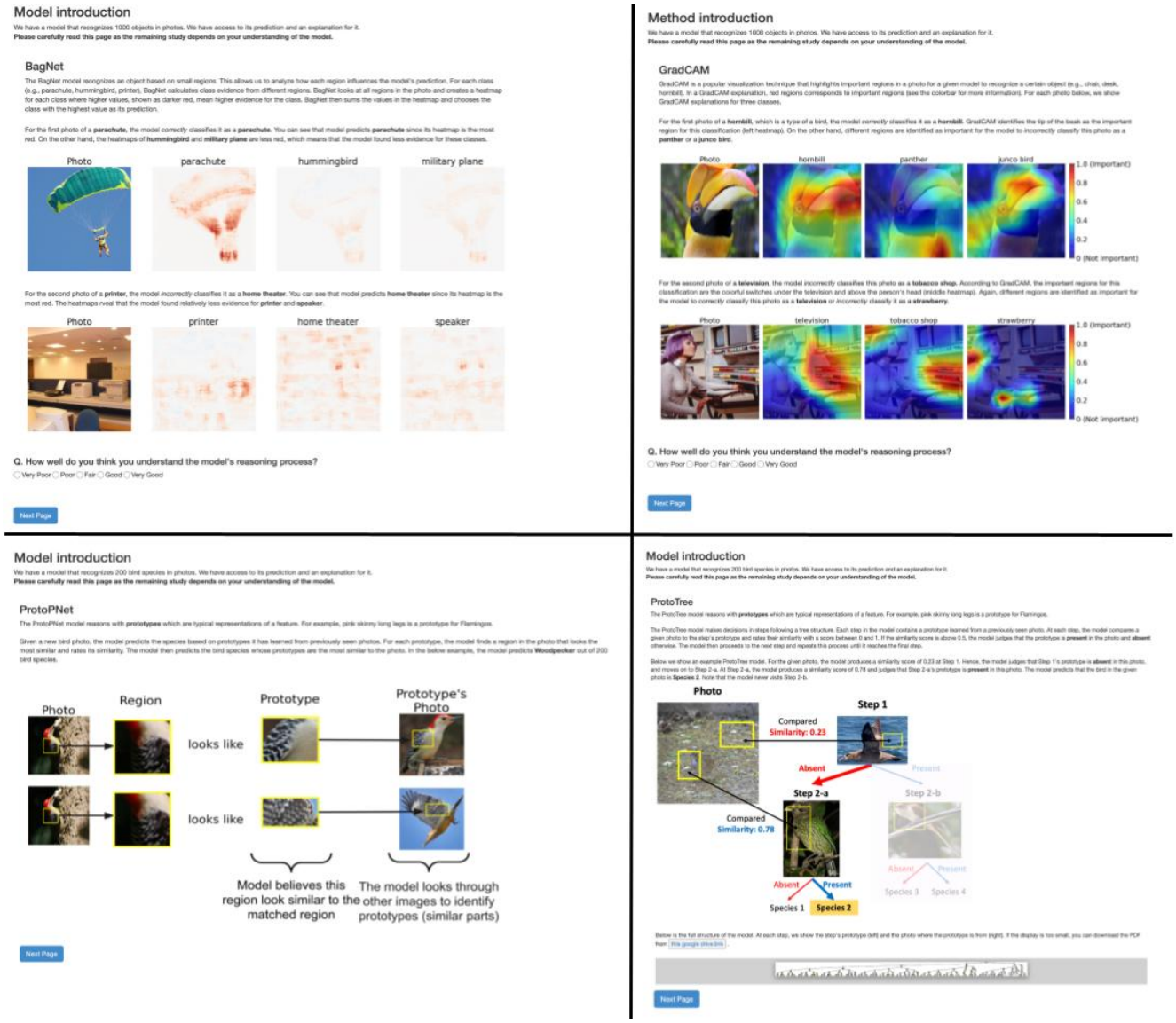

Figure A9. **3. Method introduction.** GradCAM (top left), BagNet (top right), ProtoPNet (bottom left), ProtoTree (bottom right).

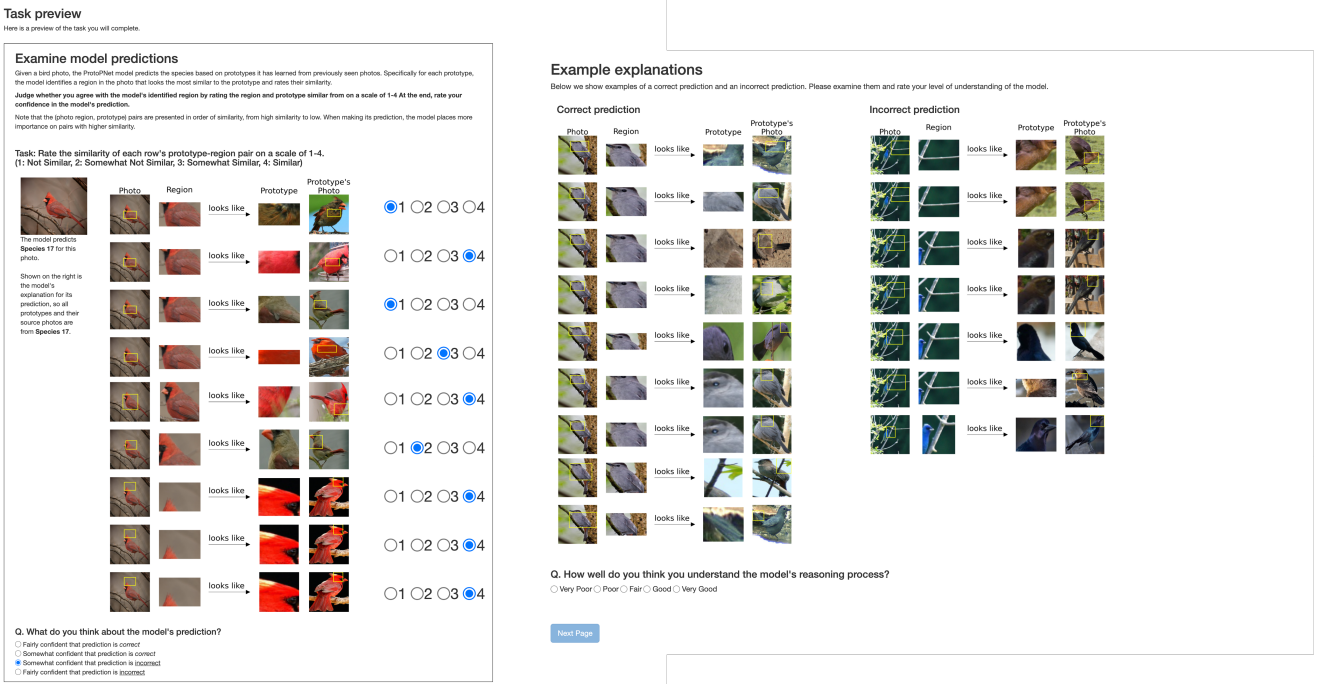

Figure A10. **4. Task preview and first subjective evaluation.**

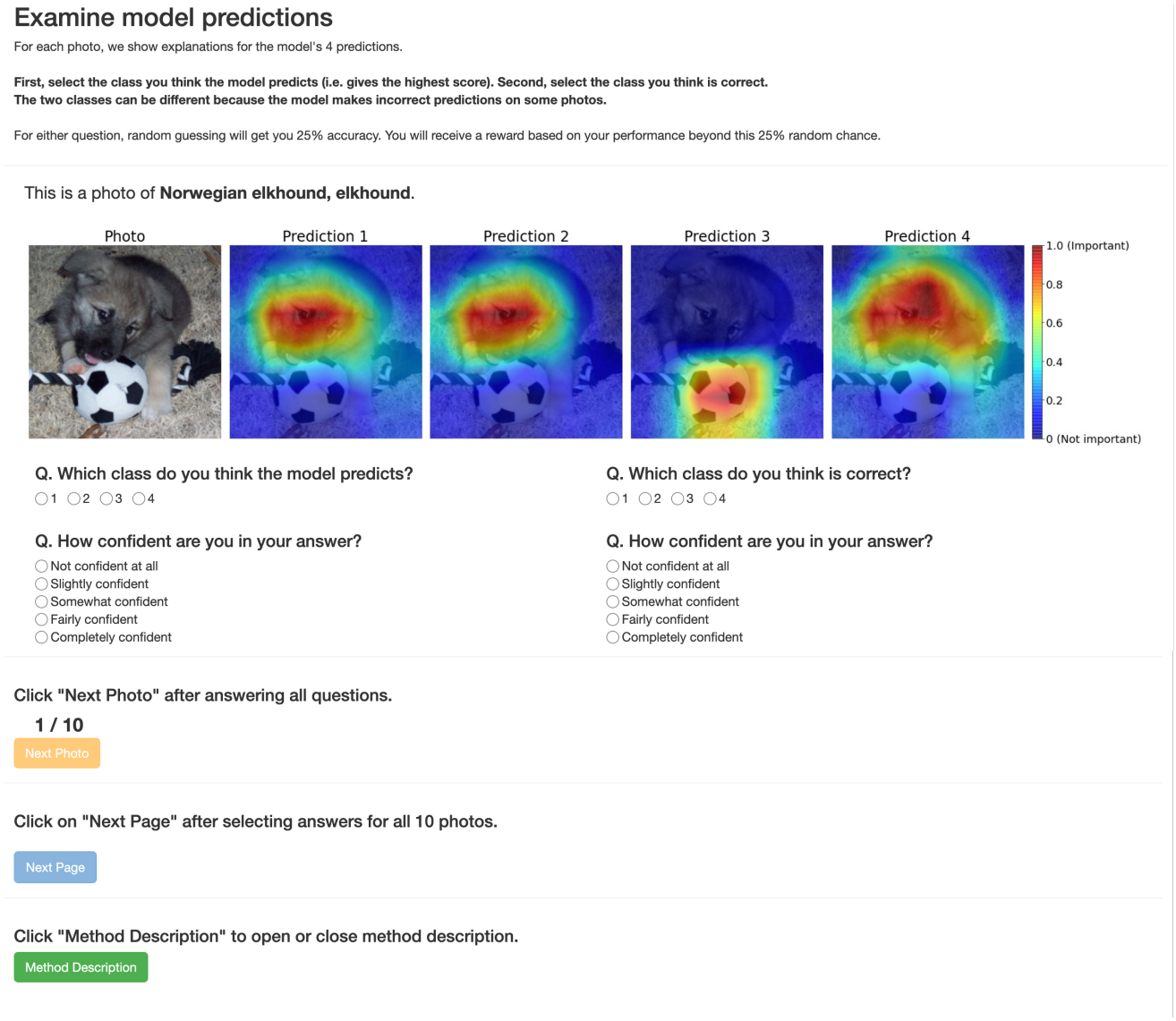

Figure A11. **5. Task: GradCAM distinction.**

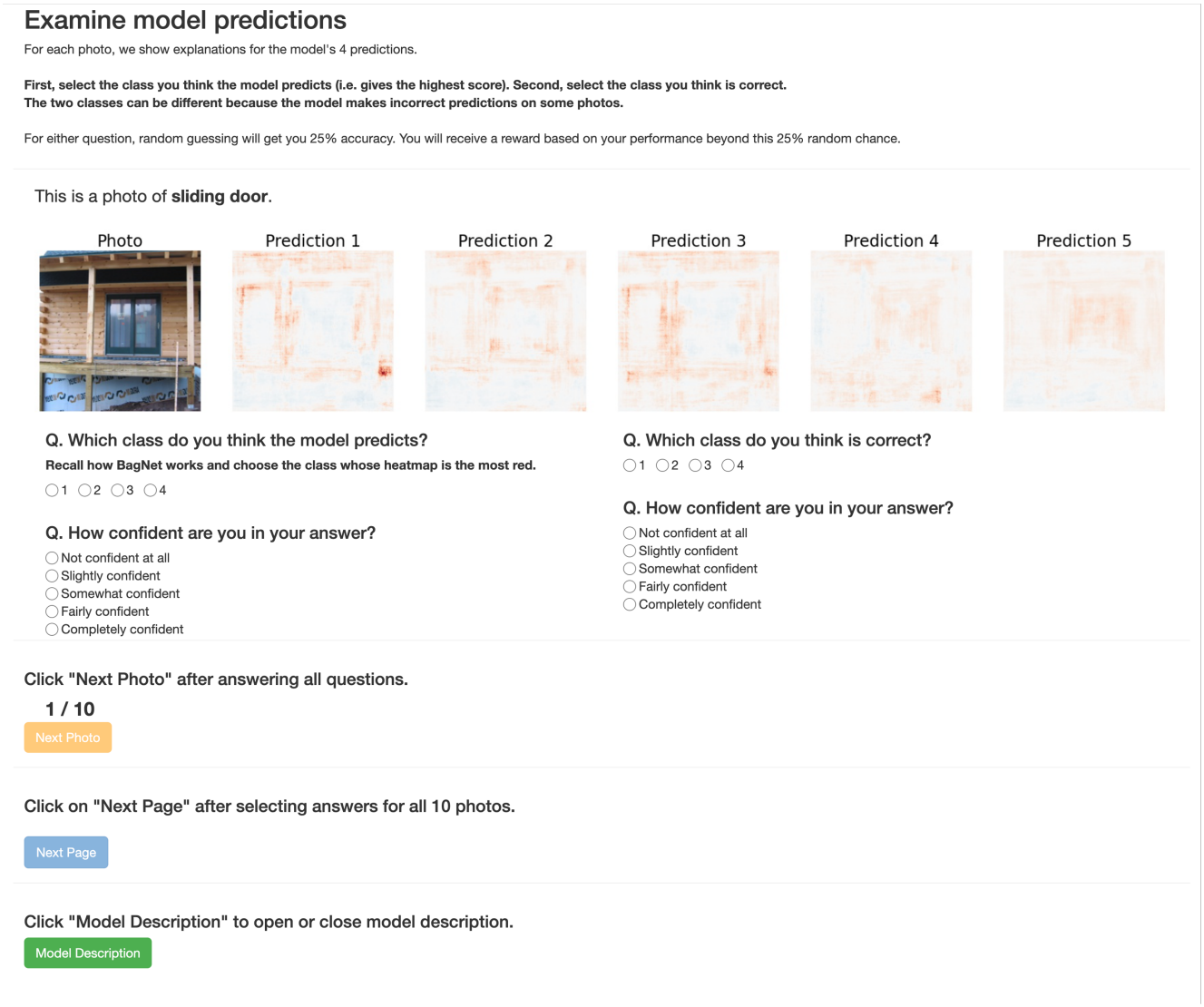

Figure A12. **5. Task: BagNet distinction.**

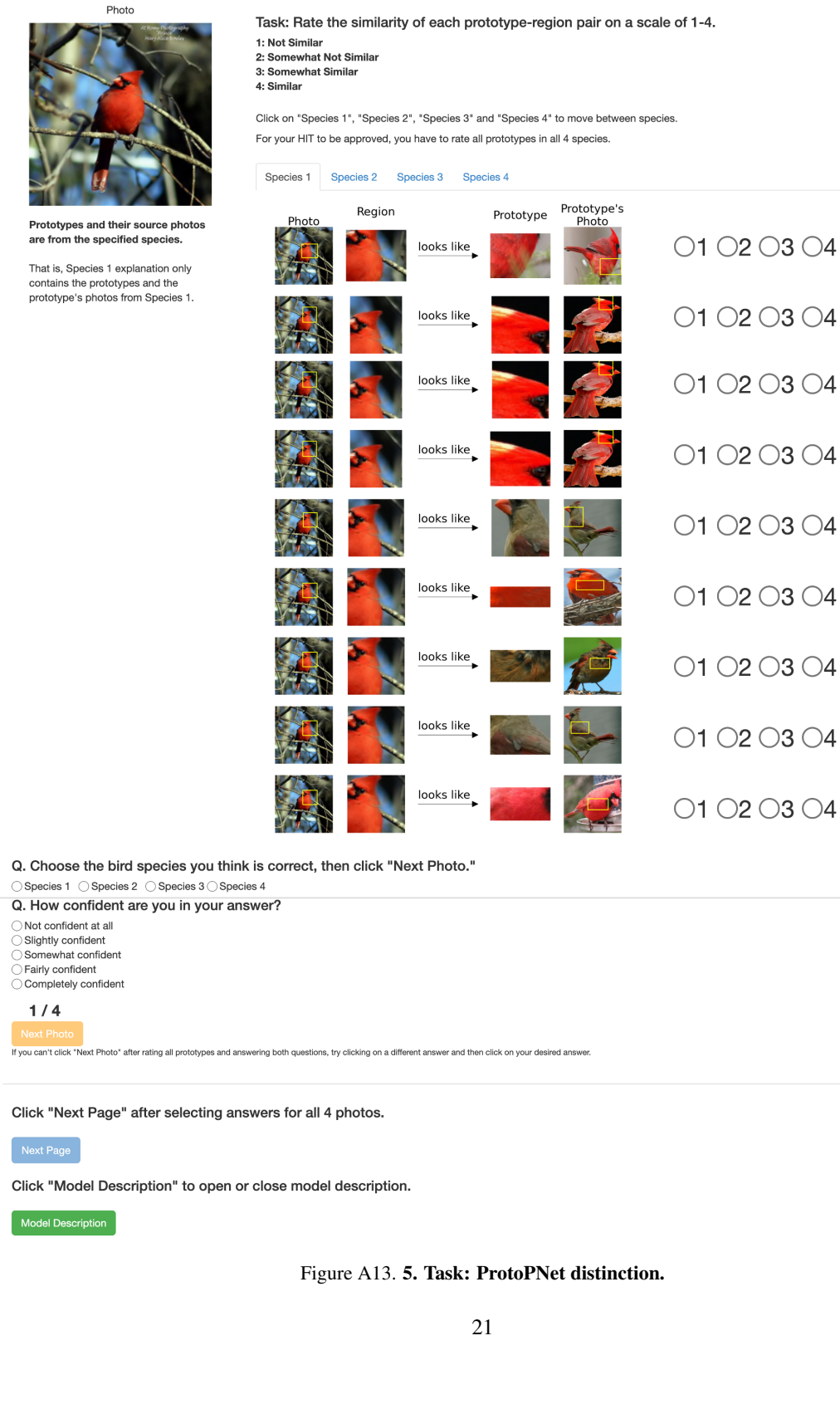

Figure A13. **5. Task: ProtoPNet distinction.**

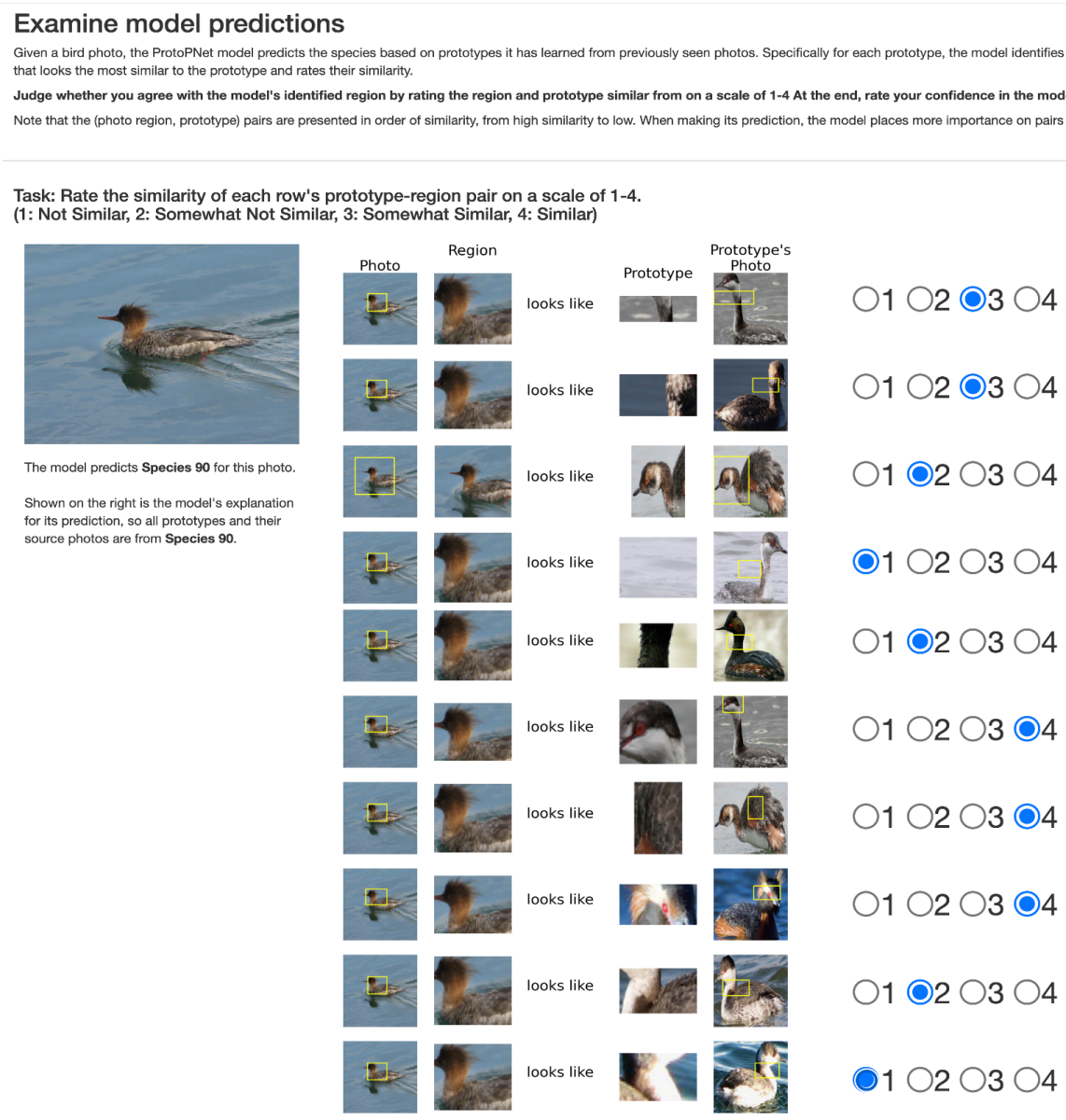

Figure A14. **5. Task: ProtoPNet agreement.**

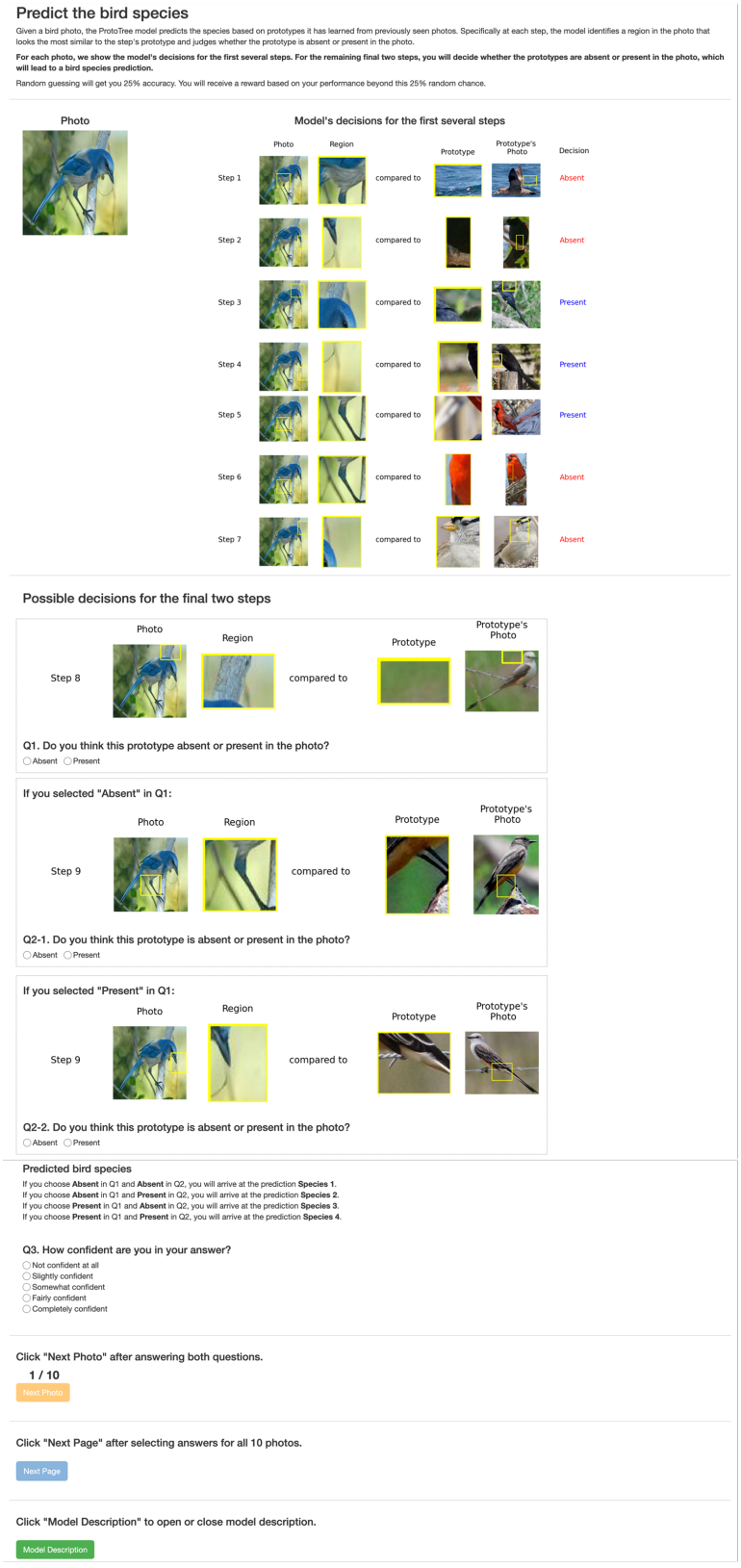

Figure A15. **5. Task: ProtoTree distinction.**

Figure A16. **5. Task: ProtoTree agreement.**

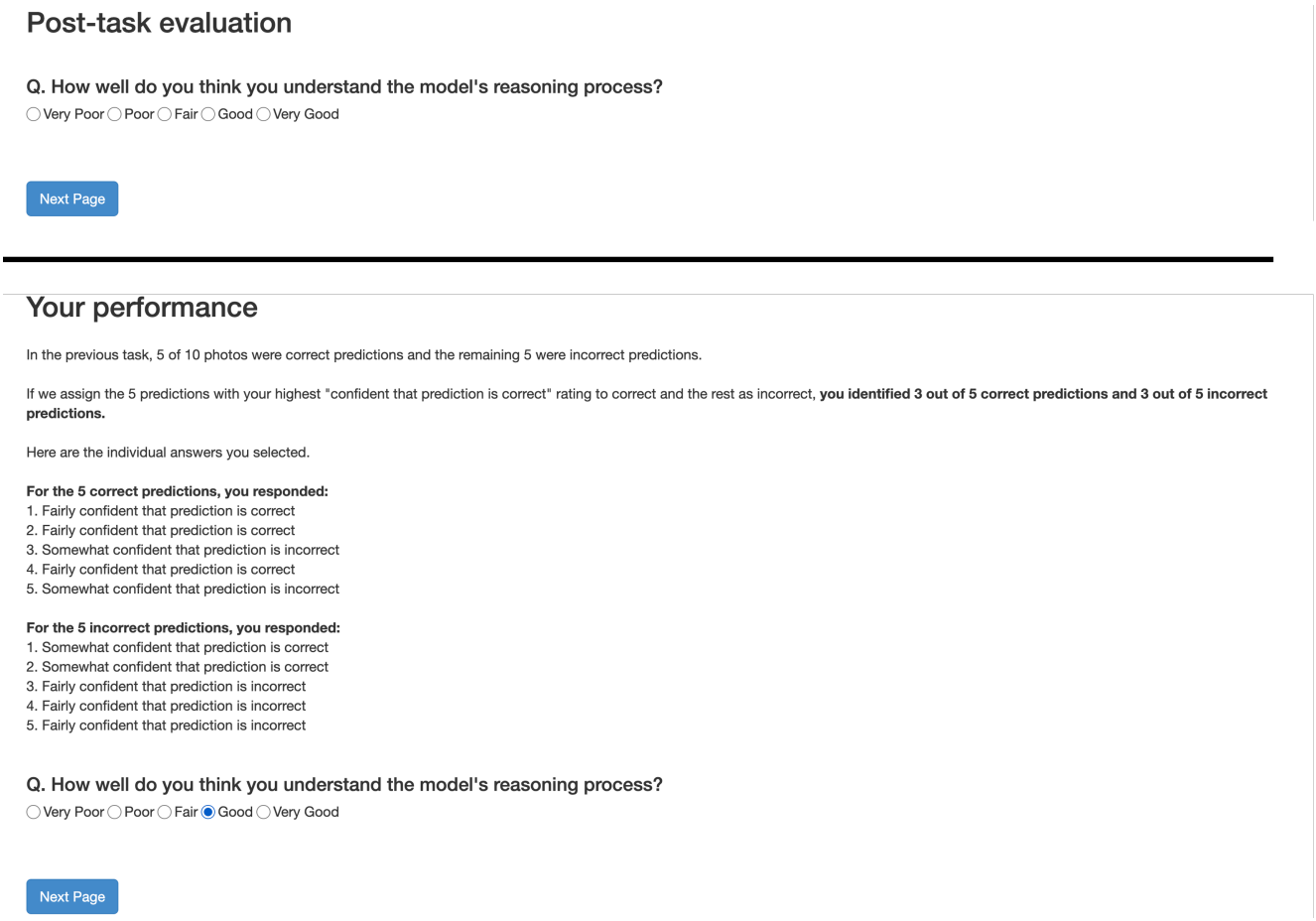

Figure A17. **6. Second and third subjective evaluation.**

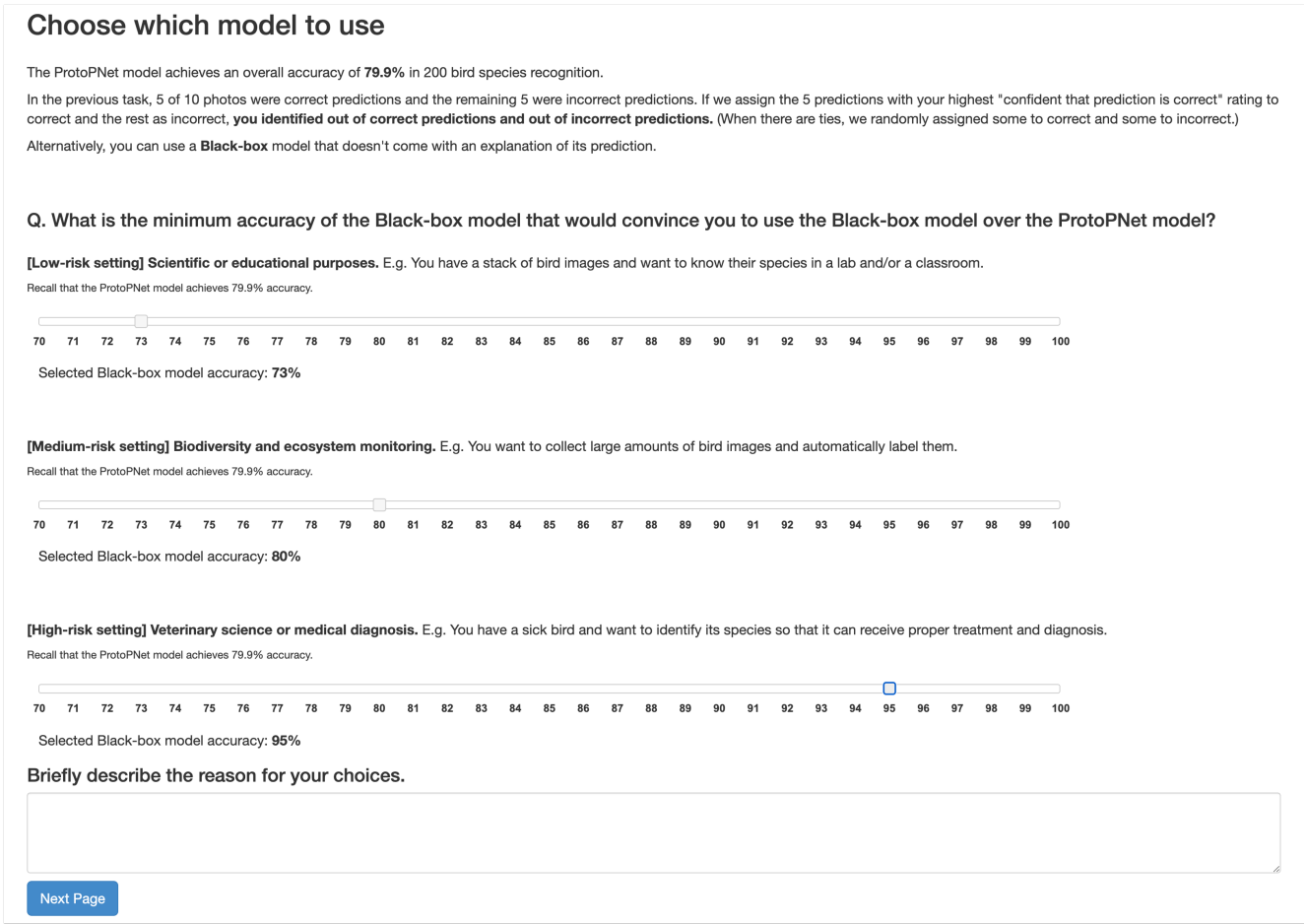

Figure A18. **7. Interpretability-accuracy tradeoff.**