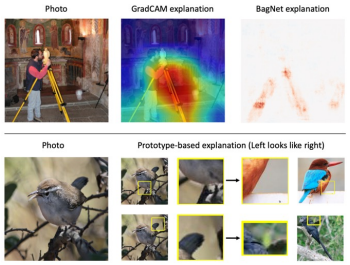


**Figure 1: Different forms of explanations present challenges to standardized evaluation of interpretability methods.** (Top) Heatmap explanations by GradCAM [24] and BagNet [2] highlight decision-relevant image regions. (Bottom) Prototype-based explanations by ProtoPNet [3] and ProtoTree [18] match image regions to canonical prototypes. This schematic is a much simplified version of the actual explanations that consist of multiple prototype-region pairs and their similarity scores, among other information (see Figs. 4 and 7).

# HIVE: Evaluating the Human Interpretability of Visual Explanations

**Sunnie S. Y. Kim**
Department of Computer Science, Princeton University
Princeton, NJ, USA
suhk@cs.princeton.edu

**Ruth Fong**
Department of Computer Science, Princeton University
Princeton, NJ, USA
ruthfong@cs.princeton.edu

**Nicole Meister**
Department of Electrical and Computer Engineering, Princeton University
Princeton, NJ, USA
nmeister@princeton.edu

**Olga Russakovsky**
Department of Computer Science, Princeton University
Princeton, NJ, USA
olgarus@cs.princeton.edu

**Vikram V. Ramaswamy**
Department of Computer Science, Princeton University
Princeton, NJ, USA
vr23@cs.princeton.edu


*CHI'22 Workshop on Human-Centered Explainable AI (HCXAI)*, May 12–13, 2022.

## 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 HIVE (Human Interpretability of Visual Explanations), a novel human evaluation framework for visual interpretability methods that allows for falsifiable hypothesis testing, cross-method comparison, and human-centered evaluation. To the best of our knowledge, this is the first work of its kind. Using HIVE, we conduct IRB-approved human studies with nearly 1000 participants and evaluate four methods that represent the diversity of computer vision interpretability works: GradCAM, BagNet, ProtoPNet, and ProtoTree. Our results suggest that explanations engender human trust, even for incorrect predictions, yet are not distinct enough for users to distinguish between correct and incorrect predictions. We open-source HIVE to enable future studies and to encourage more human-centered approaches to interpretability research.[1]

## Author Keywords

Interpretability; Explainable AI; Interpretability evaluation; Human evaluation; Visual explanations

[1] The full paper and code for the user interface (UI) can be found at https://princetonvisualai.github.io/HIVE.

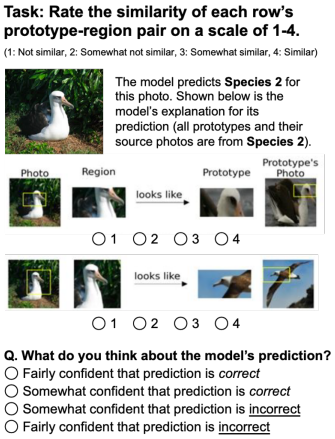


**Figure 2: Simplified UI for evaluating ProtoPNet [3] on the *agreement* task.**

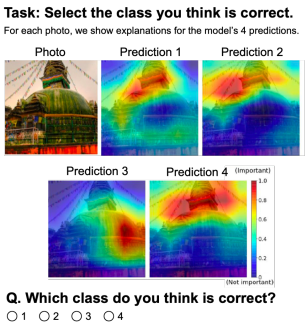


**Figure 3: Simplified UI for evaluating GradCAM [24] on the *distinction* task.** The heatmap titled "Prediction 1" highlights regions that contribute to the model's prediction of Class 1.

## Introduction

Despite the recent growth of the interpretability research field, *interpretability evaluation* remains a key challenge. The interpretability of a proposed method is often argued through a few exemplar explanations that highlight how it is more interpretable than a baseline method. Some evaluate their proposed method with quantitative metrics [6, 11, 21, 22, 27, 28], however, these metrics are often disconnected from downstream use cases of explanations and typically only apply to one explanation form (e.g., only heatmaps). Further, recent works suggest that some methods are not as interpretable as originally imagined [1, 10, 17, 19, 25] and that explanations may have an unintended negative effect of engendering over-trust in automated systems [4, 8]. These works caution against intuition-based justifications and raise awareness for the need of proper evaluation and falsifiable hypothesis testing [15] in interpretability research.

As more diverse interpretability methods are being proposed, it is more important than ever to develop a standardized evaluation framework. To this end, we propose HIVE (Human Interpretability of Visual Explanations), a novel human evaluation framework for visual interpretability methods that operationalizes the amount of utility an explanation provides to human users. Through careful design, HIVE allows for *falsifiable hypothesis testing* regarding the utility of explanations for identifying model errors, *cross-method comparison* between different interpretability techniques, and *human-centered evaluation* for understanding the practical effectiveness of interpretability.

Using HIVE, we conduct IRB-approved human studies with nearly 1000 participants and evaluate four existing methods that represent different streams of interpretability work (e.g., post-hoc explanations, interpretable-by-design models, heatmaps, and prototype-based explanations): GradCAM [24], BagNet [2], ProtoPNet [3], ProtoTree [18]. See Fig. 1 for visualizations of the methods. In this extended abstract, we give an overview of HIVE and share some of the insights we obtained through our human studies.[2]

## Overview of HIVE

In this work, we focus on AI-assisted decision making scenarios where humans use an AI (image classification) model and an interpretability method to make decisions about whether the model prediction is correct or more generally about whether to use the model and/or interpretability method. We evaluate how useful a given interpretability method is in these scenarios through the following tasks.

First, we evaluate interpretability methods on a simple *agreement* task, where we present users with a single model prediction-explanation pair for a given image and ask how confident they are in the prediction (see Fig. 2). This task simulates a common decision making setting and is close to existing evaluation schemes that consider a model's top-1 prediction and an explanation for it [24].

However, it has been previously observed that users tend to believe in model predictions when given explanations for them [16, 23]. Hence, we evaluate methods on a *distinction* task to mitigate the effect of such *confirmation bias* in interpretability evaluation. Here we simultaneously show four prediction-explanation pairs and ask users to identify the correct prediction based on the provided explanations (see

[2]We join a growing group of works that evaluate interpretability methods with human studies. However, different from prior works that evaluate explanations for models trained on tabular datasets [12, 13, 14, 23, 29] and/or explanations of similar forms (e.g., heatmaps) [5, 19, 25], HIVE enables comparison of interpretability techniques that produce different explanation forms. We are also the first to conduct a human evaluation of the studied interpretable-by-design models [2, 3, 18]. Please see the full paper for a more detailed discussion of related work.

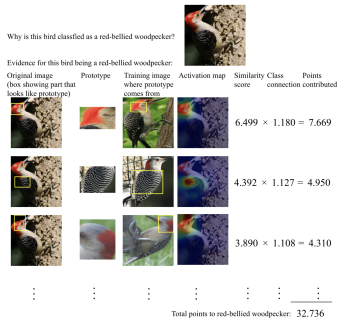


**Figure 4: ProtoPNet [3] original explanation for a single prediction.** The original explanation consists of up to 10 rows and shows prototypes and their source images, image regions matched to each prototype, heatmaps and scores conveying the similarity between the matched prototype-region pairs, and weights multiplied to the similarity scores.

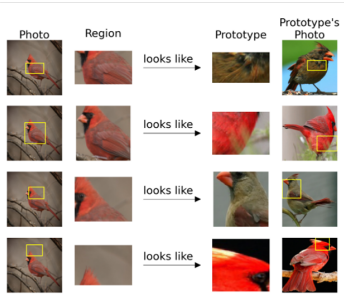


**Figure 5: ProtoPNet [3] modified explanation for a single prediction.** We abstract away the complex details and focus on showing which image region has been matched to each prototype.

Fig. 3). This task measures how well explanations can help users distinguish between correct and incorrect predictions.

In summary, HIVE consists of the following steps: We first introduce the study and the interpretability method to be evaluated. Next, we show a preview of the evaluation task and provide example explanations for one correct and one incorrect model prediction to give participants appropriate references. Afterwards, participants complete the evaluation task. Throughout the study, we also ask subjective evaluation and user preference questions to make the most out of the human studies. Our study design was approved by our institution's Institutional Review Board (IRB).

**Addressing practical challenges in human evaluation.** We also tackle a number of practical challenges to evaluate diverse interpretability methods. Evaluating all methods on a common task is only part of the answer. Since different methods are developed for different scenarios, we create an individual evaluation UI that respects each method's characteristics (e.g., its explanation form, dataset used for model training). We also make adaptations to the explanation forms of the more complex methods [3, 18] to ensure that users can complete the evaluation task.

For example, the ProtoPNet [3] model is a deep, prototype-based, interpretable-by-design model with a complex explanation form (see Fig. 4). In our study, we abstract away the details and focus on showing which image region has been matched to each prototype (see Fig. 5). The ProtoTree [18] model is also a prototype-based model, but one that takes the form of a decision tree which is commonly regarded as an interpretable model type. However, it again produces a complex explanation (see Fig. 6) and our initial, internal studies revealed that is too overwhelming for participants. Thus, in addition to simplifying the explanation (see Fig. 7), we ask participants to focus on the final two decision steps. We also walk through a simplified decision tree model and present warm-up exercises before introducing ProtoTree.

Going further, we explain the interpretability methods in simple terms, actively avoiding using technical jargon and even replacing terms such as "image" and "training set" to "photo" and "previously-seen photos." We also went through multiple iterations of UI design to present visual explanations in digestible bits to not overwhelm users.

## Experiments & Key findings

We recruited study participants through Amazon Mechanical Turk and evaluated GradCAM [24], BagNet [2], ProtoPNet [3], and ProtoTree [18] in the context-of fine-grained bird image classification [26].[3] Each study had 50 participants, and participants were compensated $12/hr.

**Participants tend to believe that a model prediction is correct, when given an explanation for it.** In the *agreement* task, participants found 72.4% of correct predictions convincing for GradCAM, 75.6% for BagNet, 73.2% for ProtoPNet, and 66.0% ProtoTree. However, they also thought 67.2% of incorrect predictions were correct for GradCAM, 57.6% for BagNet, 53.6% for ProtoPNet, and 62.8% for ProtoTree. These results reveal an issue of *confirmation bias*. Prior works have made similar observations for non-visual interpretability methods [16, 23]; we substantiate them for visual explanations, and more importantly, quantify the degree to which participants believe in explanations.

**Participants struggle to identify the correct prediction based on explanations.** On correctly predicted samples

[3] In the full paper, we present additional results for GradCAM and BagNet evaluated on another task of model output prediction [5, 9, 20] and in another setting of coarse-grained object classification.

**Figure 6: ProtoTree [18] explanation of the full model.** The ProtoTree model is an interpretable decision tree, but it decision process, consisting of 511 decision nodes and up to 10 decision steps, is likely overwhelming to human users.

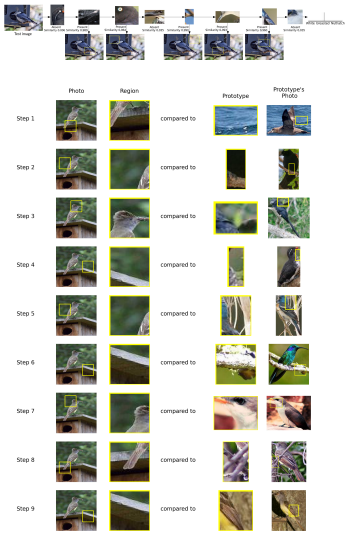


**Figure 7: ProtoTree [18] original (top) and modified (bottom) explanations for a single prediction.** We convert the horizontal explanation into a vertical one and focus on showing which image region has been matched to each prototype.

in the *distinction* task, participants achieve a mean accuracy of 71.2% on GradCAM, 45.6% on BagNet, 54.5% on ProtoPNet and 33.8% on ProtoTree, all above 25% random chance. On incorrectly predicted samples, however, the accuracy drops from 71.2% to 26.4% for GradCAM and from 45.6% to 32.0% for BagNet. (We evaluate ProtoPNet and ProtoTree only on correctly predicted samples due to the complexity of their explanations.) This result suggests that explanations for correct predictions may be more coherent and convincing than those for incorrect predictions. Nonetheless, all results are far from 100% accuracy, indicating that these interpretability methods cannot yet be reliably used for identifying model errors.

**A gap exists between prototype-region similarity ratings of ProtoPNet & ProtoTree and those of humans.** For ProtoPNet and ProtoTree, we ask participants to rate the similarity of prototype-image pairs (see Fig. 2) and empirically confirm prior work's [10, 18] anecdotal observation that prototype-based models' notion of similarity sometimes doesn't align with that of humans. This observation is noteworthy because the interpretability of these "interpretable-by-design" models stems from the property that their reasoning process consists of interpretable units (prototype-image pairs). If users disagree with the model's judgment for these units, they may not find the model interpretable.

**To prefer a baseline model over a model that comes with explanations, participants require the baseline model to have higher accuracy in higher-risk settings.** Finally, we study the *interpretability-accuracy tradeoff* participants are willing to make under different risk settings. On average, participants require the baseline model to have $+6.2\%$ higher accuracy for low-risk (e.g., scientific or educational purposes), $+8.2\%$ for medium-risk (e.g., biodiversity and ecosystem monitoring), and $+10.9\%$ for high-risk (e.g., veterinary science or medical diagnosis) settings, to use it over a model that comes with explanations.

## Discussion & Conclusion

To the best of our knowledge, HIVE is the first evaluation framework that was developed to enable falsifiable hypothesis testing, cross-method comparison, and human-centered evaluation of computer vision interpretability works. However, there are a few limitations 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, some under multiple conditions. Second, while HIVE takes a step towards use case driven evaluation, our setup is still far from real-world uses of interpretability methods. In future work, we hope to conduct a more contextually situated evaluation with domain experts and/or end-users of a real-world computer vision application (e.g., examine how bird experts choose to use one method over another when given multiple interpretability methods for a bird species recognition model).

Nonetheless, our human studies reveal several key insights about the field. In particular, we find that users generally believe model predictions are correct when given explanations for them. This result suggests explanations may have an unintended negative effect of engendering trust in AI models, even if they are incorrect or flawed in their reasoning process. Our findings underscore the need for careful examination of explanations for both correct and incorrect model predictions, as well as evaluation methods that capture how human users perceive and use interpretability methods. We hope this work helps pave the way towards human evaluation becoming commonplace, by presenting and analyzing a human study design, demonstrating its effectiveness and informativeness for interpretability evaluation, and open-sourcing the code to enable future work.

**Positionality statement.** All 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 in HIVE for impartial comparison of the studied methods.

**Acknowledgments.** This material is based upon work partially supported by the National Science Foundation (NSF) under Grant No. 1763642. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. We also acknowledge support from the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award to OR, Princeton SEAS Project X Fund to RF and OR, and the Princeton SEAS and ECE Senior Thesis Funding to NM. We thank the authors of [2, 3, 10, 18, 24] for open-sourcing their code and the authors of [2, 7, 10, 18] for sharing their trained models. We also thank our study participants, anonymous reviewers, and the Princeton Visual AI Lab members (Dora Zhao, Kaiyu Yang, Angelina Wang, and others) who tested our UI and provided helpful feedback.

## REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *NeurIPS*.

[2] Wieland Brendel and Matthias Bethge. 2019. Approximating CNNs with Bag-of-local-Features Models Works Surprisingly Well on ImageNet. In *ICLR*.

[3] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *NeurIPS*.

[4] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The Role of Trust in Automation Reliance. In *IJHCS*.

[5] Thomas Fel, Julien Colin, Rémi Cadène, and Thomas Serre. 2021. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. `arXiv:2112.04417`.

[6] Ruth Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *ICCV*.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[8] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *CSCW*.

[9] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. `arXiv:1812.04608`.

[10] Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. 2021. 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*.

[11] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. In *NeurIPS*.

[12] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human Evaluation of Models Built for Interpretability. In *HCOMP*.

[13] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2018. Human-in-the-Loop Interpretability Prior. In *NeurIPS*.

[14] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *KDD*.

[15] Matthew L. Leavitt and Ari S. Morcos. 2020. Towards Falsifiable Interpretability Research. In *NeurIPS Workshop on ML Retrospectives, Surveys & Meta-Analyses*.

[16] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. In *Queue*.

[17] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. 2021. Do Concept Bottleneck Models Learn as Intended?. In *ICLR Workshop on Responsible AI*.

[18] Meike Nauta, Ron van Bree, and Christin Seifert. 2021. Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. In *CVPR*.

[19] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The Effectiveness of Feature Attribution Methods and Its Correlation with Automatic Evaluation Scores. In *NeurIPS*.

[20] Mahsan Nourani, Chiradeep Roy, Tahrima Rahman, Eric D. Ragan, Nicholas Ruozzi, and Vibhav Gogate. 2020. Don't Explain without Verifying Veracity: An Evaluation of Explainable AI with Video Activity Recognition. `arXiv:2005.02335`.

[21] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *BMVC*.

[22] Samuele Poppi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2021. Revisiting the Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis. In *CVPR Workshop on Responsible Computer Vision*.

[23] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *CHI*.

[24] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *ICCV*.

[25] Hua Shen and Ting-Hao Kenneth Huang. 2020. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. In *HCOMP*.

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

[27] Mengjiao Yang and Been Kim. 2019. Benchmarking Attribution Methods with Relative Feature Importance. `arXiv:1907.09701`.

[28] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2016. Top-Down Neural Attention by Excitation Backprop. In *ECCV*.

[29] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect on Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *FAccT*.