# Allowing humans to interactively guide machines where to look does *not* always improve human-AI team's classification accuracy

Giang Nguyen[1, *], Mohammad Reza Taesiri[2, *], Sunnie S. Y. Kim[3], Anh Totti Nguyen[1]
[1]Auburn University, [2]University of Alberta, [3]Princeton University
[1]nguyengiangbkhn@gmail.com, [2]mtaesiri@gmail.com, [3]suhk@princeton.edu, [1]anh.ng8@gmail.com
[*]Equal contributions

## Abstract

*Via thousands of papers in Explainable AI (XAI), attention maps [50] and feature importance maps [9] have been established as a common means for finding how important each input feature is to an AI's decisions. It is an interesting, unexplored question whether allowing users to edit the feature importance at test time would improve a human-AI team's accuracy on downstream tasks. In this paper, we address this question by leveraging CHM-Corr, a state-of-the-art, ante-hoc explainable classifier [48] that first predicts patch-wise correspondences between the input and training-set images, and then bases on them to make classification decisions. We build CHM-Corr++, an interactive interface for CHM-Corr, enabling users to edit the feature importance map provided by CHM-Corr and observe updated model decisions. Via CHM-Corr++, users can gain insights into if, when, and how the model changes its outputs, improving their understanding beyond static explanations. However, our study with 18 expert users who performed 1,400 decisions finds no statistical significance that our interactive approach improves user accuracy on CUB-200 bird image classification over static explanations. This challenges the hypothesis that interactivity can boost human-AI team accuracy [15, 29, 31, 33, 42, 44, 46, 47] and raises needs for future research. We open-source CHM-Corr++, an interactive tool for editing image classifier attention (see an interactive demo here). We release code and data on github.*

## 1. Introduction

Despite much attention from the community, the practical utility of Explainable AI (XAI) tools in downstream applications (e.g., image classification [19, 23, 37, 43]) remains limited, hindering human-AI collaboration in real-world settings. A major **limitation** is that there is no interface for humans to provide feedback to the model so that it can update its decisions, which could change users' thoughts and final decisions. For example, feature importance maps [9, 36] and example-based explanations [37] are among the most popular XAI methods that offer insights into "what a model is looking at" and which real examples support a model decision, respectively. However, they only offer a **static, one-time** explanation of the input. Here, we test allowing users to edit the "attention" input to the model and observe updated model decisions iteratively until they are ready to make the final decisions (Fig. 1).

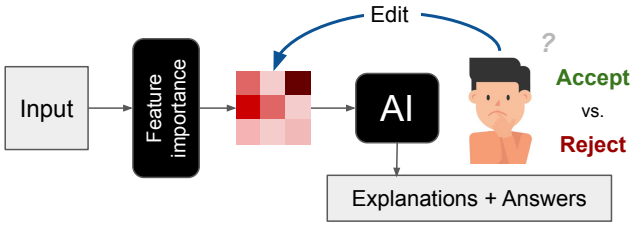



Figure 1. Users provide **feedback** to the explainable AI by editing the feature importance map (like an "attention map" [50]).

We perform our study on CHM-Corr, a recent explainable CUB-200 bird classifier [48] that combines the best of both worlds by first finding, at the patch level, how the input image is similar to the nearest training-set examples, and then using these patch-wise correspondences to predict the image label (see Fig. 2–**b** and Supp. A3 for examples of CHM-Corr explanations). CHM-Corr explanations enabled users to achieve **state-of-the-art human-AI team accuracy** in bird identification on CUB-200 [48].

We build an interactive interface called CHM-Corr++ for CHM-Corr, allowing users to manipulate the "attention" of the CHM-Corr classifier by selecting the set of patches that the classifier uses in its decision-making step (Fig. 1). By iteratively telling the model where to look, and observing the changes in the output space (see Fig. 2), users could better understand the AI model and make more informed and accurate decisions.

Via a user study of 1400 decisions, surprisingly, we did not find the interactivity to help improve users' decision-making accuracy (Tab. 1). This finding is intriguing and in stark contrast to the common hypothesis that interactive explanations might improve human-AI collaboration effectiveness and therefore human-AI team accuracy [29, 33, 42, 44].

## 2. Related Work

Fine-grained visual classification is a domain with active XAI research. Numerous explanations methods have been proposed, producing explanations of various forms. Representative forms include heatmaps [30, 40], examples [20, 38, 48, 49], concepts [25, 41, 54], and prototypes [14, 17, 34]. Regardless of the form, however, most explanations are *static*. They are presented to users in a unidirectional manner without opportunities for follow-up interactions. In this work, we explore *interactive* XAI, following growing calls from the AI and HCI communities [1, 21, 27, 29, 32, 55].

Prior work has demonstrated the needs and benefits of interactive XAI. Notably, Lakkaraju *et al*. [29] found practitioners strongly prefer interactive interfaces when making decisions with AI systems. Hohman *et al*. [21] found that interactivity was fundamental for data scientists in interpreting and comparing AI systems. Kulesza *et al*. [27] found that interactivity increased users' understanding of the AI system and ability to correct its mistakes.

However, there is a lack of interactive XAI tools that help users gain a better understanding of *computer vision* models through direct interaction with the models. Many existing tools are proprietary (e.g., AIFinnity [12], Symphony [10]), not applicable to computer vision models (e.g., Gamut [21], EluciDebug [27], TalkToModel [45], AVTALER [56]), or support different functionalities (e.g., Shared Interest [11], ActiVis [22], CNN Explainer [53]). In contrast, our interactive tool, CHM-Corr++, enable users to directly control an image classification model's attention to particular regions of the input and observe changes to its outputs (see Fig. 2). We expect CHM-Corr++ to help users build an understanding of if, when, and how the model changes its outputs, on top of the understanding provided by static explanations.

Finally, our work builds on and contributes to research on human-AI collaboration [5–8, 13, 26, 28, 35, 39, 52] that explores how humans work together with AI systems to achieve shared goals. Particularly relevant is work that studied explanations' role in human-AI decision making, especially in the context of fine-grained visual recognition [16, 23, 24, 37, 38, 48]. However, most if not all explanations studied in prior work are *static*. In this work, we further the field's understanding by exploring the role of *dynamic* explanations in human-AI collaboration.

## 3. Method

### 3.1. CHM-Corr++: An interactive interface for controlling model attention

For interactive human-AI collaboration, we developed an interactive interface that enables users to control model attention. Our interface is built on CHM-Corr [48], a visual correspondence-based image classification model that produces *static* explanations of its outputs. Given an input image, CHM-Corr employs a kNN approach to extract the $N = 50$ most similar candidate images from the training set. It then divides the input image into $7 \times 7$ non-overlapping patches and compares them with corresponding patches in each candidate image. Based on the cosine similarity among patches, CHM-Corr makes the prediction.

However, CHM-Corr is completely automatic and sometimes focuses on image patches that are not semantically meaningful to users (e.g., background or indiscriminative features – Figs. A1a & A3a). Therefore, we built an interactive interface named CHM-Corr++ that enables users to select image patches that the model should focus on, or in other words, control the model's attention. At a high level, users are presented with the new attention, support samples, and model outputs. The interface was developed using Python and Gradio [2]. We utilized Gradio with a custom HTML component that enables users to control model attention in a $7 \times 7$ grid. Compared with CHM-Corr, CHM-Corr++ enables a more interactive and user-centric image classification process, accompanied by *dynamic* explanations of the model's outputs.

### 3.2. User study

We next explore the effectiveness of static and dynamic explanations with a user study. Our problem setup is as follows (see Fig. 2): Given an input image (e.g., Cardinal), the model predicts its class **(c)** and provides an explanation **(b)** for its prediction. The user's task is to accept or reject **(d)** the model's original prediction (Summer Tanager) based on the provided explanation.

**Static vs. dynamic explanations.** In the *static* explanation setting, the model provides five support samples from the predicted class, along with patch annotations highlighting the most similar corresponding patch pairs between the input and the five support samples (see Fig. A1a).

In the *dynamic* explanation setting, the model provides the same type of explanations. However, users can also control the model's attention by selecting input image patches the model should focus on. Based on the selections, the model makes a prediction again and produces corresponding explanations (shown in Fig. A1b). Note that the new prediction can be same as the original prediction. This process enables users to explore if, when, and how the model's prediction changes based on their selections (e.g., users can

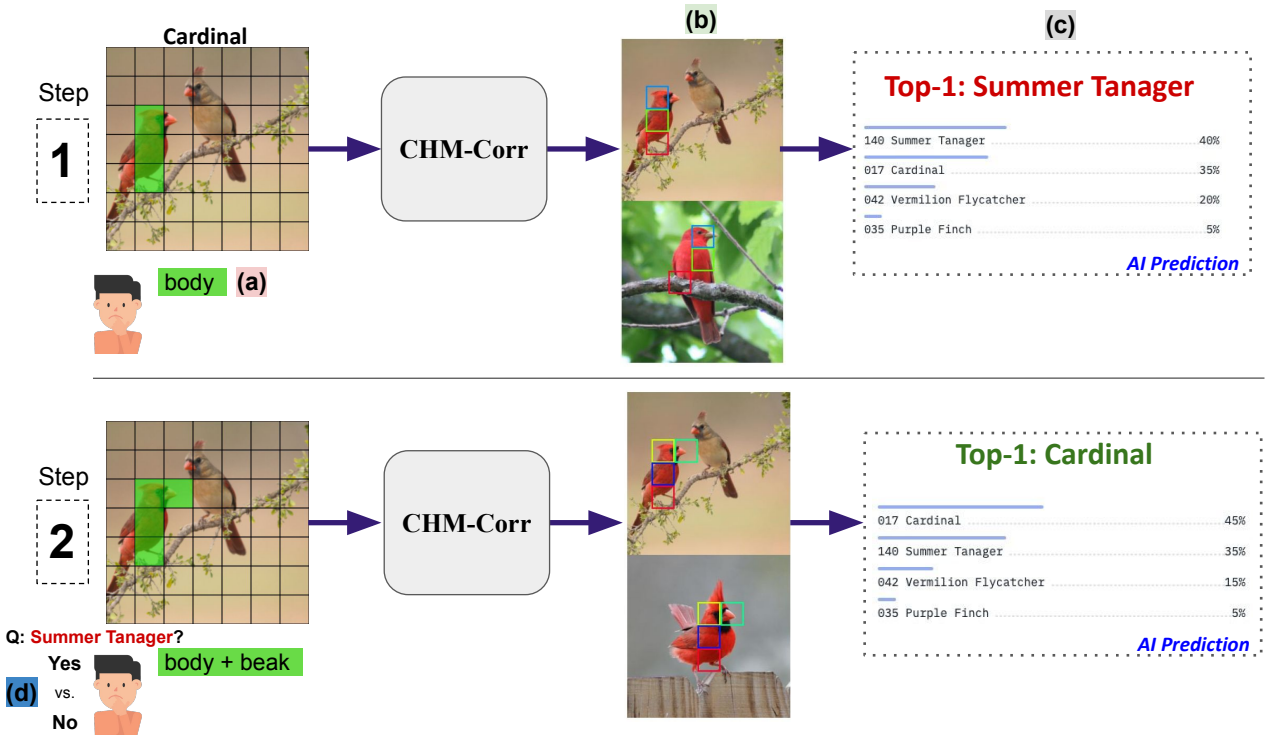



Figure 2. Our CHM-Corr++ interactive interface. We let users interact with the image classification model (here CHM-Corr [48]) via controlling the attention (selecting patches) the model should look at (**a**). Based on the user-guided attention, the model compares the input image (GT class: Cardinal) with candidate, training examples to simultaneously generate visual-correspondence explanations (**b**) and predictions (**c**). The user **iteratively** observes the dynamic explanations (**b**) and predictions (**c**) to understand the image classification model to accept or reject (**d**) the *original* top-1 predicted label (here **Summer Tanager**).

generate counterfactual explanations via observing the answer for "what if?" questions). For example, Fig. 2 demonstrates how the model's predictions change upon human interaction. Initially, using patches that only includes the bird's body, the model (CHM-Corr [48]) incorrectly predicts Summer Tanager (Step 1). However, with user-guided attention on patches that include both the body and the beak, the model correctly predicts Cardinal (Step 2).

The key difference between static and dynamic explanations lies in the level of interactivity and user involvement. Static explanations provide a fixed set of supporting information to help users make a decision, while dynamic explanations allow users to actively explore and influence the model's behavior, leading to a more engaging and informative decision-making process.

**Study materials.** Following prior works [16, 37, 48], we balance the number of correct and incorrect model predictions. From the test set of CUB-200 [51], we select 600 samples, consisting of 300 correctly classified and 300 misclassified by CHM-Corr [48], resulting in a random-chance accuracy of 50% for the task. We input each image to the model to obtain the model's predictions and explanations.

**Data collection and participants.** Due to the complexity of our interface, we decided to pilot the study with ML experts who are knowledgeable about XAI. We recruited 18 participants, most of whom were Master's and Ph.D. students in ML. Each participant completed one to several submissions, where each submission consisted of 20 decisions on whether to accept or reject the model's original prediction (e.g., Summer Tanager in Fig. 2). In total, we collected data on 1400 decisions.

## 4. Results

In our study, participants in the dynamic explanation setting "used" (i.e., controlling the model's attention and seeing a new prediction) the interactive interface 1.93 times after seeing the model's original prediction. That is, participants in the dynamic explanation setting saw around $3 \approx 1$ (original)$+1.93$ (new) model predictions on average. In this section, we explore the effect of this interactivity.

### 4.1. Interactivity did not improve decision accuracy

To our surprise, interactivity did not improve participants' decision accuracy. Dynamic explanations provided little to

Table 1. User study results. We report per-user mean decision accuracy ($\mu$) and standard deviation ($\sigma$) over a study of 18 machine learning experts who generated in total of 70 submissions (each with 20 decisions).

| Explanation type | Static (CHM-Corr) | | Dynamic (CHM-Corr++) | |
|---|---|---|---|---|
| $\mu \pm \sigma$ | Overall | | Overall | |
| | $72.68 \pm 12.36$ | | $73.57 \pm 10.42$ | |
| | AI originally correct | AI originally incorrect | AI originally correct | AI originally incorrect |
| | $85.21 \pm 11.82$ | $60.13 \pm 18.66$ | $86.79 \pm 13.16$ | $59.39 \pm 15.51$ |
| # of decisions | 283 | 277 | 443 | 397 |
| # of submissions | 28 | | 42 | |

no benefit over static explanations for participants in assessing the correctness of the model's original prediction. The overall decision accuracy is 72.68% with static and 73.57% with dynamic explanations (Tab. 1). Both are higher than the random-chance accuracy (50%) but still far short of where we want to be (100%). This result suggests that interactivity does not always benefit users, contrary to common belief that interactivity inherently boosts user understanding and task performance [29, 55].

### 4.2. Participants struggled to reject incorrect model predictions

Next, to understand where participants struggled most, we separately analyze results on instances where the model's original prediction was correct and instances where it was incorrect (Tab. 1). We find participants' decision accuracy on correct instances is much higher than that on incorrect ones for both types of explanations: 85.21% vs. 60.13% with static, 86.79% vs. 59.39% with dynamic. This result is consistent with prior findings that users tend to accept AI predictions as correct even when they are incorrect [18, 23, 37, 48], highlighting the need for tools that help users detect and reject AI errors [3, 4].

### 4.3. The usefulness of interactivity depended on the interaction outcomes

Finally, to better understand the effect of interactivity, we break down participants' decision accuracy with dynamic explanations based on the interaction outcomes (Tab. 2). Here, "consistent" refers to the model maintaining its original prediction even after the user controlled its attention.

**When the model is originally correct (i, ii)**, we find that participants' decision accuracy is higher when the model is consistent than not (90.80% vs. 75.21%). This result is in line with our expectations. When the model maintains its prediction after attention control, participants may gain higher confidence in the prediction and accept it as correct (see Supp. Figs. A1 and A2).

Table 2. Participants' decision accuracy (%) with dynamic explanations under different settings.

| **AI model correctness w.r.t. human interaction** | **Acc (%)** |
|---|---|
| (i) Originally correct and consistent (always correct) | 90.80 |
| (ii) Originally correct and inconsistent (becomes incorrect) | 75.21 |
| (iii) Originally incorrect and consistent (always incorrect) | 52.55 |
| (iv) Originally incorrect and inconsistent (always incorrect) | 62.11 |
| (v) Originally incorrect and inconsistent (becomes correct) | 65.43 |

**When the model is originally incorrect (iii, iv, v)**, participants' decision accuracy is lower when the model is consistent than not (52.55% vs. 62.11 $\rightarrow$ 65.43%). Again, this result is as expected. When the model maintains its prediction, even when it is incorrect, participants may gain higher confidence in the prediction and accept the prediction as correct. What happens when the model is inconsistent as shown in Supp. Fig. A4? We find that when the model's new predictions are always incorrect, participants' decision accuracy is 62.11%. But when the model eventually becomes correct, participants' decision accuracy goes up to 65.43%. That is, the interactive interface is most helpful when users' attention control changes the model's prediction from incorrect to correct (Supp. Fig. A3). As such, understanding when users can and cannot help the model be more accurate, and aiding users in the process, would be important directions for future research.

## 5. Discussion

We assume two leading hypotheses for why dynamic explanations do not surpass static explanations in improving human decision accuracy. First, regarding the nature of the task, in most instances, AI attention is already sufficient, as the birds are well-centered and clearly visible. Changing the task domain, for example, to include complex scenes where AI struggles to focus on the correct pixels, would likely enhance the effectiveness of CHM-Corr++. Second, we mentioned in Sec. 4.3 that CHM-Corr++ is especially helpful when the base CHM-Corr model can classify correctly. Yet, this base classifier has shortcomings (see Supp. A2) and inherently makes CHM-Corr++ ineffective in many cases (e.g., Supp. Figs. A2 & A4). We hope our open-source tool and investigation of dynamic explanations stimulates further research towards enabling effective human-AI interaction in computer vision.

## Acknowledgments

We are grateful for the participation of volunteers who spent their time and efforts in our human studies. We also thank the anonymous reviewers of XAI4CV workshop for their helpful feedback. We also thank Travis Thompson, Pooyan Rahmanzadehgervi, and Tin Nguyen from Auburn University for their helpful feedback on our early results. AN is supported by NaphCare Foundations, Adobe gifts, and NSF grant no. 2145767.

## References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–18, New York, NY, USA, 2018. Association for Computing Machinery. 2

[2] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. 2

[3] Julius Adebayo. *Towards Effective Tools for Debugging Machine Learning Models*. PhD thesis, Massachusetts Institute of Technology, 2022. 4

[4] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 700–712, 2020. 4

[5] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudré-Mauroux. OpenCrowd: A human-AI collaborative approach for finding social influencers via open-ended answers aggregation. In *Proceedings of The Web Conference 2020*, page 1851–1862, New York, NY, USA, 2020. Association for Computing Machinery. 2

[6] Zahra Ashktorab, Q. Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. Human-AI collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), 2020.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-AI team performance. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2019.

[8] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2021. Association for Computing Machinery. 2

[9] Naman Bansal, Chirag Agarwal, and Anh Nguyen. SAM: The sensitivity of attribution methods to hyperparameters. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8673–8683, 2020. 1

[10] Alex Bäuerle, Ángel Alexander Cabrera, Fred Hohman, Megan Maher, David Koski, Xavier Suau, Titus Barik, and Dominik Moritz. Symphony: Composing interactive interfaces for machine learning. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2022. Association for Computing Machinery. 2

[11] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2022. Association for Computing Machinery. 2

[12] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Robert Deline, Adam Perer, and Steven M. Drucker. What did my AI learn? how data scientists make sense of model behavior. *ACM Trans. Comput.-Hum. Interact.*, 30 (1), 2023. 2

[13] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "Hello AI"': Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), 2019. 2

[14] 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 *Neural Information Processing Systems (NeurIPS)*, 2019. 2

[15] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019. 1

[16] Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems*, 35: 2832–2845, 2022. 2, 3

[17] Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. Deformable ProtoPNet: An interpretable image classifier using deformable prototypes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[18] Raymond Fok and Daniel S Weld. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *arXiv preprint arXiv:2305.07722*, 2023. 4

[19] Courtney Ford and Mark T Keane. Explaining classifications to non experts: An XAI user study of post hoc explanations for a classifier when people lack expertise. *arXiv preprint arXiv:2212.09342*, 2022. 1

[20] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019. 2

[21] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In

*Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019. 2

[22] Minsuk Kahng, Pierre Yves Andrews, Aditya Kalro, and Duen Horng Chau. ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics*, 24:88–97, 2017. 2

[23] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. HIVE: Evaluating the Human Interpretability of Visual Explanations. In *European Conference on Computer Vision*, pages 280–298. Springer, 2022. 1, 2, 4

[24] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "Help Me Help the AI"': Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2023. Association for Computing Machinery. 2

[25] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, 2020. 2

[26] Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39):e2115730119, 2022. 2

[27] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, page 126–137, New York, NY, USA, 2015. Association for Computing Machinery. 2

[28] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-AI collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2022. Association for Computing Machinery. 2

[29] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875*, 2022. 1, 2, 4

[30] Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. Scouter: Slot attention-based classifier for explainable image recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[31] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–45, 2021. 1

[32] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. 2

[33] Dimitry Mindlin, Fabian Beer, Leonie Nora Sieger, Stefan Heindorf, Philipp Cimiano, Elena Esposito, and Axel-Cyrille Ngonga-Ngomo. Beyond one-shot explanations: A systematic literature review of dialogue-based xai approaches. 2024. 1, 2

[34] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[35] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, page 189–199, New York, NY, USA, 2018. Association for Computing Machinery. 2

[36] Giang Nguyen, Shuan Chen, Tae Joon Jun, and Daeyoung Kim. Explaining how deep neural networks forget by deep visualization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, pages 162–173. Springer, 2021. 1

[37] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 34:26422–26436, 2021. 1, 2, 3, 4

[38] Giang Nguyen, Valerie Chen, Mohammad Reza Taesiri, and Anh Totti Nguyen. Pcnn: Probable-class nearest-neighbor explanations improve fine-grained image classification accuracy for ais and humans, 2023. 2

[39] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, 64(5):904–938, 2022. PMID: 33092417. 2

[40] Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[41] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, and Olga Russakovsky. ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features, 2022. 2

[42] Hua Shen. Towards useful AI interpretability for humans via interactive ai explanations. 2024. 1, 2

[43] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. One explanation is not enough: structured attention graphs for image classification. *Advances in Neural Information Processing Systems*, 34:11352–11363, 2021. 1

[44] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024. 1, 2

[45] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, 2023. 2

[46] Kacper Sokol and Peter Flach. One explanation does not fit all: The promise of interactive explanations for machine

learning transparency. *KI-Künstliche Intelligenz*, 34(2):235–250, 2020. 1

[47] Yuan Sun and S Shyam Sundar. Exploring the effects of interactive dialogue in improving user control for explainable online symptom checkers. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022. 1

[48] Mohammad Reza Taesiri, Giang Nguyen, and Anh Nguyen. Visual correspondence-based explanations improve ai robustness and human-AI team accuracy. *Advances in Neural Information Processing Systems*, 35:34287–34301, 2022. 1, 2, 3, 4

[49] Simon Vandenhende, Dhruv Mahajan, Filip Radenovic, and Deepti Ghadiyaram. Making heads or tails: Towards semantically consistent visual counterfactuals. In *ECCV 2022*, 2022. 2

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[51] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. 2011. 3

[52] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), 2019. 2

[53] Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Chau. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2020. 2

[54] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[55] Tong Zhang, X Jessie Yang, and Boyang Li. May I Ask a Follow-up Question? Understanding the Benefits of Conversations in Neural Network Explainability. *arXiv preprint arXiv:2309.13965*, 2023. 2, 4, 1

[56] Joyce Zhou, Elena Glassman, and Daniel S Weld. An interactive UI to support sensemaking over collections of parallel texts. *arXiv preprint arXiv:2303.06264*, 2023. 2

# Supplemental Materials

## A1. The significance t-test for comparing two groups

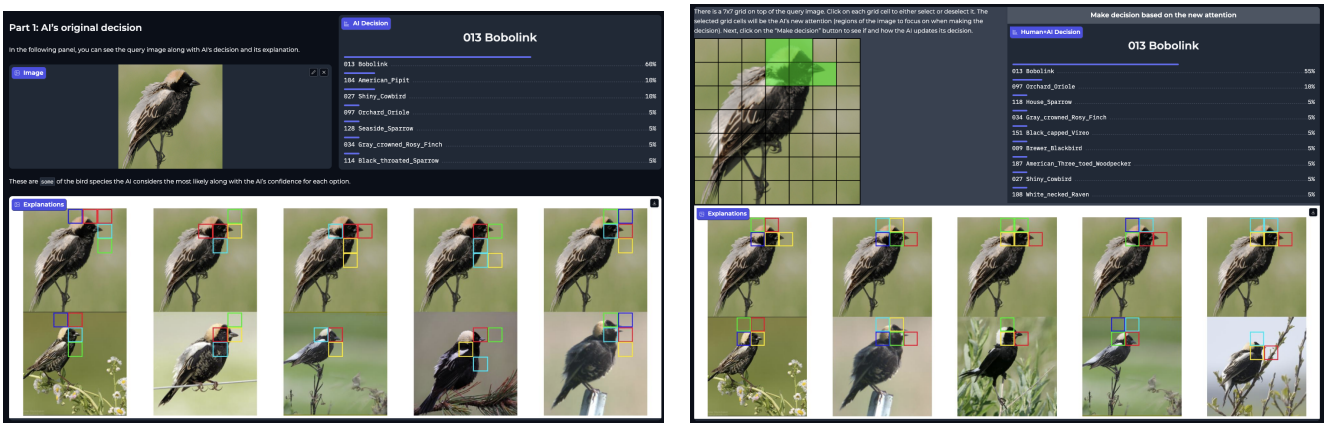
With data collected from the human study, we compute the per-user average accuracy over our controlled, balanced image sets (i.e., the ratio of samples where AI correctly classified and misclassified is approx. 50/50) and report in Table 1. Yet, we find that dynamic explanations (CHM-Corr++ users score 73.57%), which one might naturally assume to be superior, do not show a significant improvement over static ones (CHM-Corr users score 72.68%) in helping users making more accurate decisions.

The findings, with a t-statistic of -0.321 and a p-value of 0.749, indicate that the average accuracy levels for users exposed to both types of explanations are not significantly different. This challenges the common belief that dynamic, interactive content inherently boosts user comprehension or performance [29, 55].

## A2. The shortcomings of CHM-Corr classifier

For some samples, AI fails to classify the input image correctly regardless of how the attention is directed towards the image (see Fig. A4). This indicates that improving attention alone does not suffice for classifiers to accurately classify these samples, suggesting the need for insights into developing new models that focus on more than just improving attention mechanisms. Moreover, the underlying nature of the classifier contributes to this issue. Given that the classifier employs a k-Nearest Neighbors (kNN) algorithm to retrieve a set of candidate samples, there is a possibility that the ground-truth class may not appear within the candidate pool. Consequently, no matter how the CHM-Corr model re-ranks these candidates, it may never correctly identify the top-1 class.

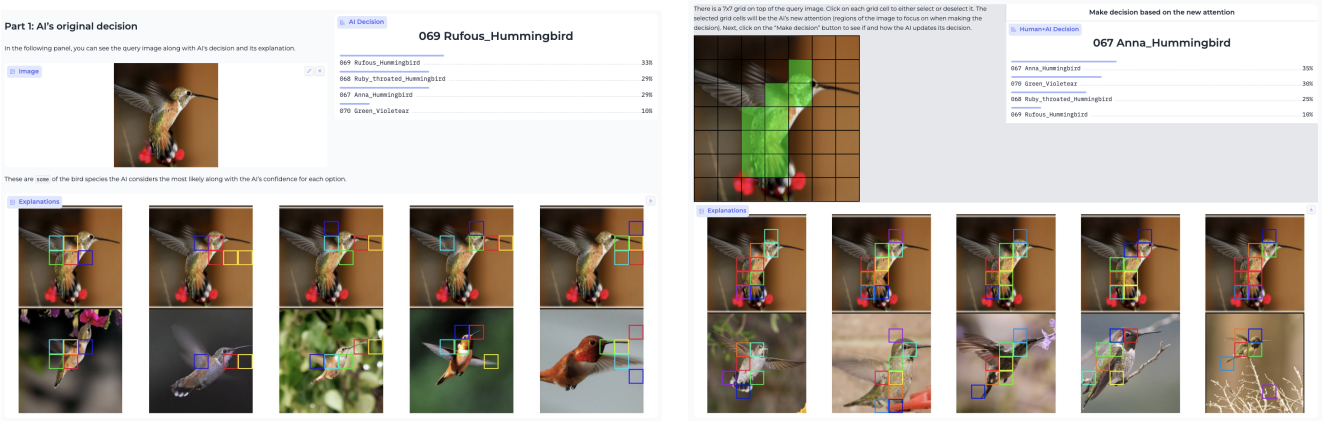
## A3. How users interact with explanations




(a) Static explanation for Bobolink.

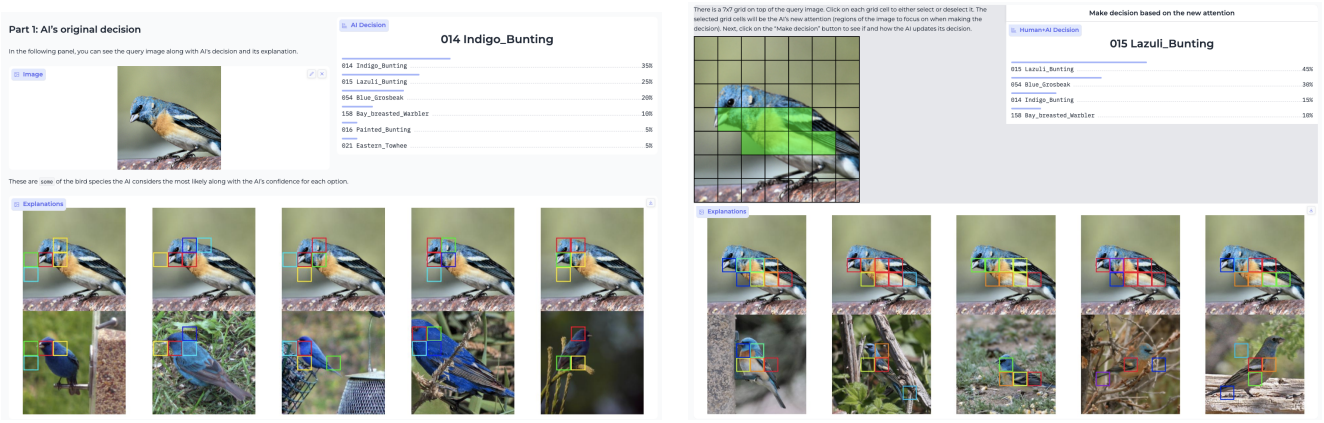
(b) Human-prompt, correspondence-based explanation for Bobolink.

Figure A1. Both dynamic and static explanations enable human users to verify that the AI is predicting the top-1 label correctly.




(a) Static explanation for Rufous Hummingbird. (b) Human-prompt, corr-based explanation for Anna Hummingbird.

Figure A2. Human intervention changes the top-1 label from Rufous Hummingbird → Anna Hummingbird that makes users more likely to reject the original, correct label.




(a) Static explanation for Indigo Bunting. (b) Human-prompt, correspondence-based explanation for Lazuli Bunting.

Figure A3. AI initially makes the wrong classification Indigo Bunting on the input image. Human intervention changes the top-1 label from Indigo Bunting → Lazuli Bunting, a more similar-looking class, encouraging users to reject the original, predicted label.

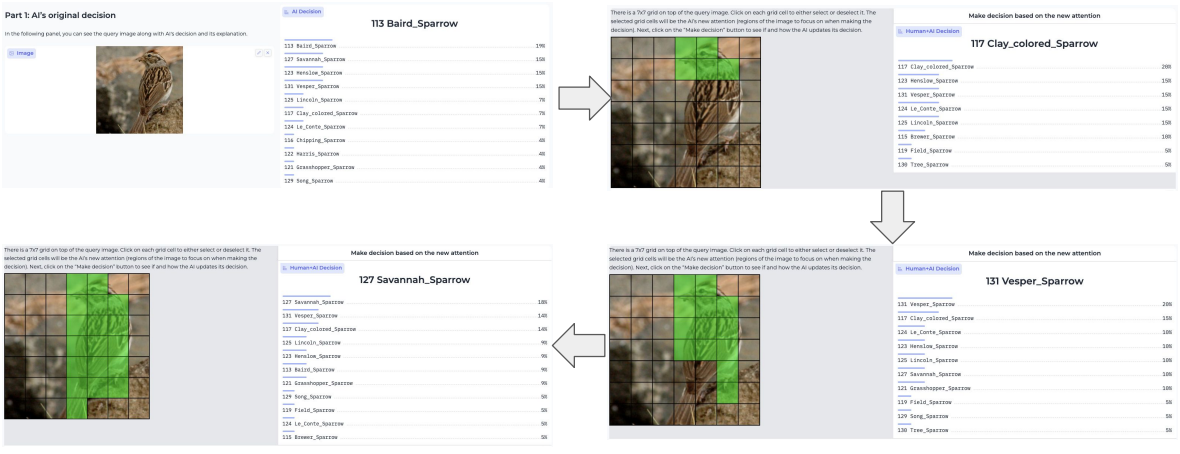



Figure A4. A sample that is unclassifiable for the classifier CHM-Corr. The ground-truth label is Chipping Sparrow. Initially, AI makes a wrong classification of Baird Sparrow. With user-guided attention, the top-1 label evolves from Baird Sparrow → Clay-colored Sparrow → Vesper Sparrow → Savannah Sparrow but none of them matches the groundtruth, making users unable to make decisions given the explanations.