Received: 21 July 2023 | Accepted: 22 May 2024

DOI: 10.1111/bjop.12714

the british psychological society
promoting excellence in psychology

ARTICLE

Explanation strategies in humans versus current explainable artificial intelligence: Insights from image classification

Ruoxi Qi[1] | **Yueyuan Zheng**[1,2] | **Yi Yang**[2] | **Caleb Chen Cao**[2,3] | **Janet H. Hsiao**[4]

[1]Department of Psychology, University of Hong Kong, Hong Kong SAR, China

[2]Huawei Research Hong Kong, Hong Kong SAR, China

[3]Big Data Institute, Hong Kong University of Science and Technology, Hong Kong SAR, China

[4]Division of Social Science, Hong Kong University of Science and Technology, Hong Kong SAR, China

Correspondence
Janet H. Hsiao, Room 3374, Division of Social Science, Hong Kong University of Science & Technology, Clearwater Bay, Kowloon, Hong Kong SAR.
Email: jhhsiao@ust.hk

Funding information
Huawei Technologies; Research Grants Council of Hong Kong (RGC), Grant/Award Number: C7129-20G

Abstract

Explainable AI (XAI) methods provide explanations of AI models, but our understanding of how they compare with human explanations remains limited. Here, we examined human participants' attention strategies when classifying images and when explaining how they classified the images through eye-tracking and compared their attention strategies with saliency-based explanations from current XAI methods. We found that humans adopted more explorative attention strategies for the explanation task than the classification task itself. Two representative explanation strategies were identified through clustering: One involved focused visual scanning on foreground objects with more conceptual explanations, which contained more specific information for inferring class labels, whereas the other involved explorative scanning with more visual explanations, which were rated higher in effectiveness for early category learning. Interestingly, XAI saliency map explanations had the highest similarity to the explorative attention strategy in humans, and explanations highlighting discriminative features from invoking observable causality through perturbation had higher similarity to human strategies than those highlighting internal features associated with higher class score. Thus, humans use both visual and conceptual information during explanation, which serve different purposes, and

Ruoxi Qi and Yueyuan Zheng contributed equally and share co-first authorship.

Some of the data in the manuscript were used in a conference paper (Yang et al., 2022) presented at the 10th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2022) and a conference paper (Qi et al., 2023) presented at the 45th Annual Conference of the Cognitive Science Society (CogSci 2023).

XAI methods that highlight features informing observable causality match better with human explanations, potentially more accessible to users.

KEYWORDS
EMHMM, explainable AI, explanation, eye movements, image classification, text analysis

BACKGROUND

To ensure good use of AI by humans, researchers have long recognized the importance of explanation to enhance human-AI interaction, including the development of Expert Systems in 1980s and Knowledge-Based Tutors in 1990s and early 2000s (Mueller et al., 2019; see Gunning et al., 2019 for an overview). Around the mid-2010s, a new generation of explainable AI (XAI) emerged due to the advance of deep learning methods, whose decision-making processes are often obscured to both users and developers. As compared with previous explanation solutions, these XAI methods use better visualization techniques (Goyal et al., 2016) or directly make the classifiers themselves more explainable (Akata, 2013). However, similar to previous efforts, they remain focusing on using more AI to explain AI without much consideration of users' mental processes (Hoffman et al., 2018; Hsiao, Ngai, et al., 2021). This differs significantly from how humans provide explanations. For example, Kaufman and Kirsh (2022) found that in visual explanations, human explanations typically involve directing attention to relevant details following a sequence of visual reasoning processes, in contrast to XAI methods that simply highlight features used by AI classifiers without temporal information. They also suggested that human explanations often consider explainees' prior knowledge and qualitative reasoning styles, which are typically missing in current XAI methods. Indeed, effective human explanations often involve causal reasoning based on observed regularities in the world (Bender, 2020; Einhorn & Hogarth, 1986; Holzinger et al., 2019; Maxwell, 2004). When providing explanations about others, people use more observable behaviour (e.g. facial expressions, gestures, etc.) in contrast to the unobservable behaviour (e.g. thoughts, feelings, desires, etc.; Malle & Knobe, 1997). They also prefer explanations that invoke causality (Zemla et al., 2017). In particular, rather than listing all possible causes of an event in an explanation, people tend to provide contrastive explanations that focus on why the current event occurs instead of other non-occurring events (Chin-Parker & Cantelon, 2017; Miller, 2021; Van Fraassen, 1980). Knowledge about how humans give explanations provides important insights into ways to make explanations from XAI more accessible to humans.

Despite these initial efforts, our current understanding of how humans provide explanations on tasks that are commonly performed by AI remains very limited, especially for those involving making decisions based on complex perceptual processes that are often automatic and unconscious in humans such as image classification. Image classification has been a heated topic in computer vision, and the advance of deep learning methods in recent years has significantly increased automated image classification accuracy (Rawat & Wang, 2017). A common XAI method for image classification has been using saliency maps that highlight regions of the input image according to their importance to the AI model's output (Li et al., 2021). For example, to explain how an AI model classifies an image as 'horse' (Figure 1), a saliency map highlighting pixels around the head and the thigh of the horse in the image would suggest these are the most important visual features used by the model for this classification. Two major saliency map-based approaches are perturbation-based and backpropagation-based methods. Perturbation-based methods, such as RISE (Petsiuk et al., 2018), perturb the input image and place more weights on the pixels that affect the output class probability relative to other classes when occluded. In contrast, backpropagation-based methods, such as GradCAM (Selvaraju et al., 2020), calculate the gradient of the score for the target class in a particular layer as the class relevance of each pixel.

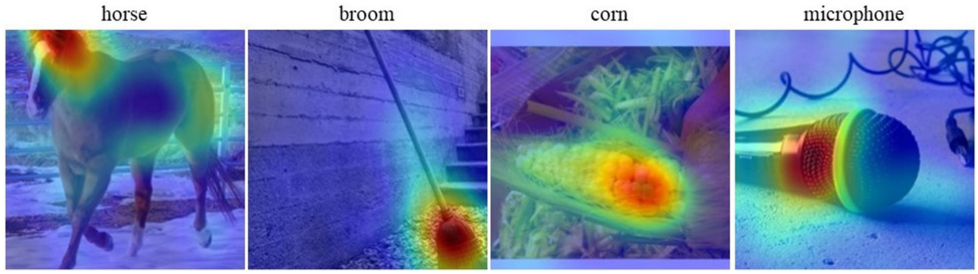


FIGURE 1 Examples of saliency maps generated by an XAI method (i.e. PCB-RISE) to explain AI model's outputs.

Saliency-based explanations are often evaluated using different computational and cognitive metrics including faithfulness (Samek et al., 2017), reliability (Kindermans et al., 2019) and plausibility. Since humans typically attend to features important for their judgements during image classification (Hwu et al., 2021; Lai et al., 2020), human attention is often considered a good benchmark for plausibility (Karim et al., 2022; Lanfredi et al., 2021; Mohseni et al., 2021; Yang et al., 2022).

Note, however, that human attention when viewing an image has been shown to be task-driven and thus may differ significantly when the task demand changes (Hsiao, An, et al., 2021; Kanan et al., 2015; Hsiao & Chan, 2023). On one hand, saliency map-based XAI highlights image regions that contribute to the classifier output, and thus should be compared with human attention when performing image classification tasks (Lai et al., 2020). On the other hand, the purpose of the saliency maps is to provide explanations, and thus they may be better compared with human attention when they explain image classification (Yang et al., 2022). It remains unclear how human attention differs between image classification and explanation tasks. Providing explanations on how to perform a task involves metacognitive skills to evaluate thought processes through self-awareness (Balcikanli, 2011; Jiang et al., 2016), and thus explanation strategies for the task may differ from the strategies for performing the task itself. Also, during image classification, humans only need to attend to sufficient information for making a decision (Hsiao & Cottrell, 2008; Smith & Ratcliff, 2004), whereas during explanation, they may attend to all relevant information to provide a comprehensive explanation (Gelman et al., 1998). Thus, human attention during image classification and explanation may differ significantly. Understanding whether saliency maps generated using the current XAI methods are better matched with human attention during image classification or explanation will provide important insights on what information these XAI salience maps reflect and how human users should interpret them.

Another factor to consider is the substantial individual differences in human attention during cognitive tasks as demonstrated through eye-tracking studies (Chuk et al., 2014; Chuk, Crookes, et al., 2017; Hsiao, Chan, et al., 2021; Hsiao, Lan, et al., 2021; Peterson & Eckstein, 2013), and these individual differences are often associated with differences in task performance and cognitive abilities (An & Hsiao, 2021; Hsiao, Lan, et al., 2021). Since individuals can differ significantly in both cognitive and metacognitive abilities across domains (Rouault et al., 2018), substantial individual differences in explanation strategies are expected. It remains unclear how individual differences in explanation strategy are compared with current XAI methods. More specifically, human object category representations involve both visual and abstract conceptual features (Martin et al., 2018), both of which can be used in explanations. Thus, individuals may differ in their reliance on visual or conceptual information when providing explanations. Previous research has suggested that humans use more visual information than conceptual information to form perceptually rich representations when learning novel categories (Kloos & Sloutsky, 2008), suggesting the importance of visual information during early category learning. Since XAI saliency maps provide visual feature explanations, they may match better with human attention strategies associated with more use of visual information. If so, they may be particularly effective for early category learning (Fisher & Sloutsky, 2005). Also, since humans prefer explanations that involve causality and contrastive explanations (Miller, 2021;

Zemla et al., 2017), saliency maps generated through perturbation methods, which use observable causality between input perturbation and consequent change in output class probability, may show a better match with human attention during explanation than gradient-based methods. The comparisons between different human and XAI explanation strategies will provide important insights on the explanation processes of both humans and XAI methods and ways to enhance XAI explanations to facilitate human-AI interaction.

Thus, here, we aimed to fill these research gaps by examining the following research questions:

1. What are the individual differences in human attention strategies when performing image classification and explanation tasks? Are they associated with task performance, in particular the differences related to the use of visual vs. conceptual information during explanation?
2. How do human attention strategies during image classification and explanation differ?
3. How are XAI saliency maps compared with human attention strategies? In particular, are XAI saliency maps more similar to a particular human attention strategy during classification or explanation? Which XAI method matches better with human attention strategies?

To examine participants' use of visual and conceptual information in the explanations, we created three different measures: (1) *Effectiveness* for novel category learning rated by domain experts (computer vision scientists), which is relevant to the use of visual information; (2) *Diagnosticity*, that is how specific and informative the explanation is for identifying the class label, which is relevant to the use of conceptual information; and (3) *explanation text characteristics analysis* to directly assess the amount of visual and conceptual information conveyed in the explanation text.

We used eye tracking to directly measure human attention, in contrast to indirect measures such as the annotation or pointing approaches that are typically used in previous studies (Gelman et al., 1998; Mohseni et al., 2021). To quantify individual differences in visual scanning behaviour during the tasks and to discover representative attention strategies in humans, we adopted a machine learning model-based approach, Eye Movement analysis with Hidden Markov Models (EMHMM; Chuk et al., 2014) with co-clustering (Hsiao, Lan, et al., 2021). In this approach, an individual's eye movement behaviour in viewing a stimulus is summarized in a hidden Markov model (HMM) in terms of person-specific regions of interest (ROIs) and transition probabilities among the ROIs. The co-clustering algorithm is then used to discover participant groups where group members adopt similar strategies to one another across stimuli, with each group forming a representative attention strategy. Similarities among individual strategies then can be quantitatively assessed using their data log-likelihoods given the representative strategy models. Consistency of a strategy can be assessed using entropy of the HMM (Cover & Thomas, 2006; higher entropy indicates lower consistency). Thus, adopting this approach allows us to take both spatial (ROI choice) and temporal information (the order of the ROIs visited) into account when quantifying individual differences in attention strategies. This approach has been applied to a variety of research fields and led to novel findings not discoverable by traditional methods (e.g. summary statistics of eye movement in predefined ROIs or fixation heatmaps; Barton et al., 2006; Caldara & Miellet, 2011), including psychology (An & Hsiao, 2021; Hsiao et al., 2022; Hsiao, Chan, et al., 2021), mental health (Chan et al., 2020; Zhang et al., 2019) and education (Zheng et al., 2022). We compared human attention strategies with saliency maps generated by a representative perturbation-based method, RISE (Petsiuk et al., 2018) and a representative backpropagation-based method, Grad-CAM (Selvaraju et al., 2020) for an image classification AI model ResNet50, which has excellent classification performance (He et al., 2016). Since RISE can be affected by the pixel distribution of the random masks used to generate saliency maps, we included a pixel coverage bias (PCB) corrected version of RISE (Xie et al., 2022).

We hypothesized that (1) in both image classification and explanation tasks, we would discover different attention strategies associated with different task performance, and attention strategies during explanation may be associated with different reliance on visual or conceptual information; (2) human attention strategies when explaining image classification results would cover more relevant features

than those during image classification itself; and (3) Human attention maps from those who rely more on visual information would show higher similarity to XAI saliency maps, and perturbation-based explanations such as RISE, which highlights discriminative features by invoking causality through perturbation, may show higher similarity to these human attention maps.

METHODS

Participants

We recruited 62 participants (52 females[1]), aged 18–37 years ($M = 22.5$, $SD = 3.8$) from a local university. They had normal or corrected-to-normal vision. The participants included 7 native speakers of English. For the non-native English speakers, they started to learn English at an average age of 5.2 ($SD = 2.4$). The participants had a mean score of 71.20% ($SD = 12.79\%$) on the Lexical Test for Advanced Learners of English (LexTALE; Lemhöfer & Broersma, 2012).[2] Here, we examined the difference between two participant groups using different eye movement patterns in classification and explanation performance. A power analysis of independent sample t-test based on a similar study comparing eye movement pattern groups on face recognition performance (Chuk, Crookes, et al., 2017; $d = 2.18$) suggested that a sample size of 52 was sufficient ($d = 0.8$, $\alpha = .05$, $\beta = .2$). In addition, we examined whether eye movement patterns can predict the participants' performance in the two tasks. A power analysis of linear multiple regression indicated that 55 participants were required assuming a medium effect size ($f^2 = 0.15$, $\alpha = .05$, $\beta = .2$) and one tested predictor (i.e. eye movement pattern). The informed consent was obtained from all participants.

Materials and apparatus

The stimuli included 160 images in 20 categories, with 8 images in each category. The 20 image categories contained 9 natural categories, including ant, corn, horse, jellyfish, lemon, lion, mushroom, snail and zebra, and 11 artificial categories, including broom, cell phone, fountain, harp, laptop, microphone, pizza, shovel, sofa, tennis ball and umbrella. These image classes were selected from human basic level categories (Markman & Wisniewski, 1997; Wang et al., 2015) and were also commonly used as output categories of image classification AI models (Russakovsky et al., 2015).

The 16 images for the categories horse and sofa were obtained from PASCAL VOC (Everingham et al., 2010), while the rest 144 images were from ImageNet (Deng et al., 2009). The images together constituted a representative set, including different levels of foreground object complexity and background saliency. All images were resized to fit into a 400×520 pixel frame on a blank canvas. Since the original images differed in their aspect ratios, white edges were added so that the images had the same size without any distortion.

The experiment was conducted using E-Prime 3.0 with the extensions for EyeLink (Psychology Software Tools) on a 255 mm × 195 mm laptop with a resolution of 1024×768 pixels. Each image spanned $9.68° \times 12.32°$ of visual angle at a viewing distance of 60 cm. The dominant eyes of participants were tracked with an EyeLink Portable Duo eye tracker (SR Research), and a chinrest was used to minimize head movement. A nine-point calibration and validation procedure was performed at the beginning of the classification and explanation task, and recalibration took place whenever drift check error was over 1° of visual angle.

[1]There was no gender difference in eye-movement patterns during image classification, $t(59) = 0.47$, $p = .642$, or explanation, $t(60) = 0.43$, $p = .671$. Female and male participants also did not differ in classification performance (accuracy: $t(60) = 0.46$, $p = .645$; RT, $t(53) = 1.09$, $p = .279$) or explanation performance (effectiveness: $t(60) = 0.63$, $p = .529$; diagnosticity: $t(60) = 0.59$, $p = .559$).

[2]Note that one limitation of our study was the majority of the participants were non-native English speakers.

Design

In the current study, we aimed to compare explanation strategies in humans and current XAI in image classification. Specifically, we examined:

1. Human attention strategies during image classification and explanation and their associations with performance: We examined participants' attention strategies as reflected in eye movement behaviour when classifying images (the classification task) and when explaining image classification (the explanation task). We also examined the relationship between these attention strategies and task performance, including accuracy and RT in the classification task, and explanation effectiveness, diagnosticity and text characteristics in the explanation task (see Data analysis section for details of how these were measured). For each task, EMHMM was used to discover two representative attention strategies in the participants through clustering and to quantify individual participants' attention strategies along the dimension contrasting the two representative strategies. ANCOVA was used to examine whether participants using the two attention strategies differed in performance measures with participants' English proficiency (measured by LexTALE; see English proficiency test section) as a covariate. Correlation analyses were used to examine the relationship between attention strategy and performance measures. Hierarchical regression analyses were used to investigate whether attention strategy could predict performance after cognitive abilities (see Cognitive tasks and English proficiency test section for how they were measured) and English proficiency were controlled. These results were presented in Image classification task section and Explanation task section.
2. Comparisons between human attention strategies during image classification and explanation: To directly compare participants' attention strategies across classification and explanation tasks, in a separate analysis we mixed the eye movement data from the two tasks together and used EMHMM to discover two representative attention strategies across the two tasks through clustering. We then quantified participants' attention strategies in the two tasks along the dimension contrasting the two discovered strategies. We used paired sample *t*-test to examine whether participants' attention strategies in the two tasks significantly differed. The results were presented in Comparison of the two tasks section.
3. Comparisons between human attention strategies and XAI saliency maps: To compare human attention strategies with XAI saliency maps, we used a 2 × 2 × 3 by-item ANOVA to examine how participants' task (classification vs. explanation), attention strategy (the two representative strategies) and XAI method for saliency-based explanations (RISE vs. PCB corrected RISE vs. Grad-CAM) affected the similarity between XAI saliency maps and human attention maps for image classification. This ANOVA allowed us to examine whether this similarity measure would differ when the human attention maps were obtained from different tasks or participants using different attention strategies, or when XAI saliency maps were obtained using different methods. The results were presented in Comparison with XAI saliency maps Section.

   Please refer to section Data analysis for the details of the measures and analyses.

Procedures

Participants completed two main tasks (i.e. the classification task and explanation task) with eye tracking, followed by four cognitive tasks and an English proficiency test (LexTALE).

Classification task

In the classification task, the participants were instructed to assign a class label to 160 images one at a time based on the 20 labels shown at the beginning of the experiment (Figure 2a). Each trial started

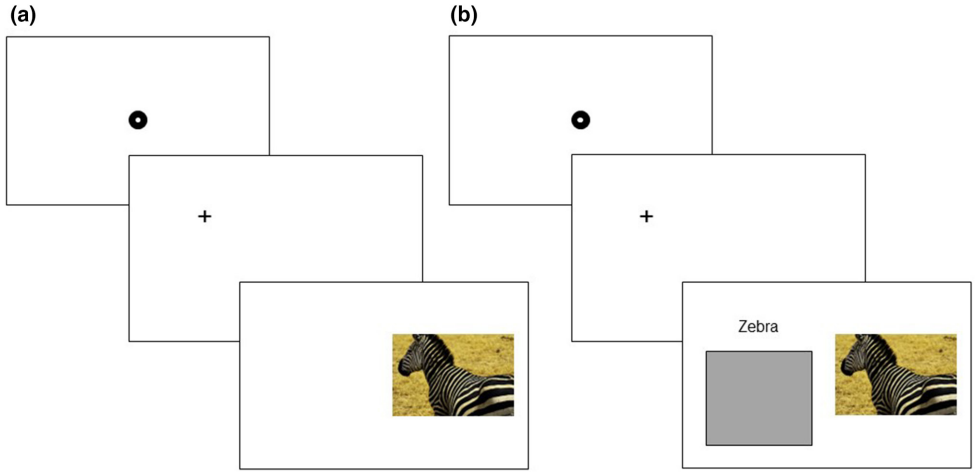


FIGURE 2 Experimental procedure of the (a) classification task and the (b) explanation task.

with a drift check at the centre of the screen. After a stable fixation at the centre was detected, a fixation cross was shown at the upper left corner of the screen. The image appeared once the participant fixated on the cross for more than 250 ms to ensure that the first fixation on the image was planned by the participants. The participant's fixation was consistently directed to the left and the images were consistently placed on the right to match the reading direction of English (Spalek & Hammad, 2005). After viewing the image, the participants named the class label aloud as quickly as they recognized it. Their reaction time (RT) was recorded by a microphone through the voice key of the Chronos response box (Psychology Software Tools).

Explanation task

In the explanation task, the participants were shown the same 160 images one at a time along with the correct label and were asked to provide an explanation in a textbox about why the label should be assigned to the image based on how they classified the same image in the previous task (Figure 2b). They were told to imagine explaining to someone without any prior knowledge of the visual categories, such as a very young child, and to include sufficient information from the image to help the person learn how to identify the categories. Similar to the classification task, each trial started with a drift check at the centre of the screen and a fixation cross directed the participant's fixation to the upper left corner, where the class label appeared.

Before starting the explanation task, the participants were given an example explanation for an image of an elephant: 'a long trunk below the eye next to a white pointy object that looks like a tusk; a big triangular ear'. They also completed three practice trials with images from three different classes not used in the experiment to ensure that they understood the task instruction. They were reminded of the instruction and the example when their explanation did not have sufficient information from the image as the example.

Cognitive tasks and English proficiency test

Verbal and visuospatial two-back tasks

The two-back tasks (Lau et al., 2010) were used to measure the participants' working memory capacity. In the verbal two-back task, numbers were presented at the centre of the screen one at a time, and the participants judged whether it was the same as the one shown two trials back. In the visuospatial two-back task, different symbols appeared at different locations one at a time, and the participants judged

whether they appeared at the same location as the one shown two trials back. There were two blocks in total, with 28 trials each. Accuracy and RT were recorded.

The flanker task

The flanker task (Ridderinkhof et al., 1999) tested the participants' selective attention. In each trial, the participants were presented with five arrows and instructed to judge the direction of the central arrow. The other four arrows, or the flankers, pointed to the same direction as the central arrow in the congruent condition and pointed to the opposite direction in the incongruent condition. In the neutral condition, the flankers were non-directional. There were 20 trials for each of the three conditions. Flanker effect in accuracy and RT was measured as $\frac{\text{Congruent} - \text{Incongruent}}{\text{Congruent} + \text{Incongruent}}$.

Multitasking test

The multitasking test (Stoet et al., 2013) assessed the participants' task-switching ability. The participants were shown four types of figures differentiated by outer shape (diamond or square) and inner filling (two dots or three dots). Each figure appeared either at the upper section or the lower section of a rectangular box. The participants judged the shape of the figure if it appeared at the upper section and the filling if it appeared at the lower section. The task included three blocks, with 32 trials each. The figures always appeared at the top in the first block and always appeared at the bottom in the second block (no-switching tasks). In the third block, the figures appeared randomly either at the top or at the bottom (dual task). Their task-switching ability was measured as the accuracy/RT in the dual task minus the average accuracy/RT in the two no-switching tasks.

Tower of London task

The Tower of London task (Phillips et al., 2001) assessed participants' executive function and planning ability. In each trial, participants were presented with one target board and one move board, each with three balls that were randomly distributed on three sticks. The participants were instructed to move the balls on the move board to make it look exactly the same as the target board with the fewest possible moves. There were 10 trials. Accuracy, averaged number of moves, planning time and execution time were measured.

English proficiency test

The LexTALE (Lemhöfer & Broersma, 2012) was used to assess the participants' English lexical knowledge. This test included 60 trials, where the participants saw a string of letters and judged whether it was an existing English word on each trial.

Data analysis

Classification task

Task performance analysis

The participants' performance in the classification task was measured by accuracy and average RT. When the participant's response did not match the class label exactly, it was considered correct if it was a synonym (e.g. mobile phone for cell phone) or a closely related word (e.g. tennis for tennis ball).

Eye movement analysis

EMHMM (Chuk et al., 2014) with co-clustering (Hsiao, Lan, et al., 2021; see http://visal.cs.cityu.edu.hk/research/emhmm/) was used to model and quantify the participants' eye movement patterns in the classification task, with both spatial (fixation locations) and temporal (transitions between the locations) dimensions taken into account. Eye movement data from 61 participants on 160 image stimuli were used

to perform this analysis.[3] Only fixations on the image area were included in the analysis. Outlier fixations that were more than three standard deviations from the mean fixation location of the specific image on either the vertical or the horizontal dimension were removed. Trials where the participant answered incorrectly were excluded.

Each participant's eye movements in viewing one of the 160 image stimuli were summarized with one HMM, which included person-specific ROIs and transition probabilities among these ROIs. A variational Bayesian approach (Coviello et al., 2014) was used to determine the optimal number of ROIs for each individual HMM with a preset range of possible number of ROIs from 1 to 10. Each HMM with a specific number of ROIs was trained for 200 times, and the HMM with the highest log-likelihood was chosen. The co-clustering method was used to cluster the participants into two groups such that participants in the same group had similar eye movement patterns across the stimuli, and a representative HMM was generated for each group for each stimulus, with the number of ROIs set to be the median number of ROIs of the individual HMMs. The co-clustering procedure was repeated for 200 times to select the result with the highest log-likelihood.

The participants' eye movement patterns were quantified using the A-B scale, which was defined as $\frac{L_A - L_B}{L_A + L_B}$, where L_A and L_B represent the log-likelihoods of a participant's eye movement data being classified as belonging to Pattern Group A and Pattern Group B, respectively (Chan et al., 2018; Liao et al., 2022; Zheng & Hsiao, 2023). A higher A-B scale indicates higher similarity to Pattern Group A. In addition, L_A and L_B were used to evaluate whether the two representative patterns differ significantly from each other: If the two groups indeed differed significantly, it was expected that Pattern Group A participants should have significantly higher L_A than L_B, and vice versa for participants from Pattern Group B (Chuk et al., 2014; Hsiao, Lan, et al., 2021). Eye movement consistency was assessed by calculating the entropy for each HMM and summing over all the stimuli (Cover & Thomas, 2006).

Explanation task

Task performance analysis

The participants' performance in the explanation task was based on the quality of the explanation text they provided. Two measures were used:

1. *Effectiveness* for teaching someone without prior category knowledge how to classify the image, measured as subjective ratings from two computer vision experts. More specifically, two raters were asked to rate on a scale from 1 to 7, where 1 indicated very low effectiveness and 7 indicated very high effectiveness. The instructions that the participants received were first summarized to the raters. It was emphasized that ratings should be based on whether the explanation could effectively teach someone without prior category knowledge how to classify the image using the visual features or characteristics of the image. The raters could see the corresponding images when rating the explanations. Two data scientists with expertise in computer vision were selected as the raters, since they had more experience in processing images in terms of visual features. The two raters had good inter-rater reliability, intraclass correlation (ICC) = .720, 95% CI [0.709, 0.731], calculated based on a mean-rating, consistency, two-way random-effects model. Average rating was used as the measure of effectiveness.
2. *Diagnosticity*, that is how specific and informative the explanation is for identifying the class label, as measured by a separate group of naïve observers' accuracy in inferring the class label directly from the explanation without seeing the image. More specifically, the explanations provided by the participants were presented to 124 naïve observers (88 females, aged 18–32 years, $M = 20.08$,

[3]One participant was excluded from this analysis due to suspected tracking error. In addition, six participants had inaccurately measured classification RT and were excluded from the subsequent analyses on this variable, but they were still included in the co-clustering analysis because their eye movement data were unaffected. Thus, the effects can be tested with sufficient power provided the finalized sample size (55 valid participants).

SD = 2.03) without the image. Each observer viewed 160 explanations and was asked to guess the category label for each explanation. Each explanation was presented to two observers, and none of the observers saw multiple explanations provided by the same participant for the same category. This group of naïve observers had an average score of 74.26% (*SD* = 14.29%) on the LexTALE. Responses that matched the original label word or any of its synonyms, hyponyms, or close hypernyms were counted as correct. Percent accuracy was used as the measure of diagnosticity.

Typos, misspelled words, misused words and grammatical errors that may impede understanding were fixed before the evaluations.

Explanation text characteristics analysis

The explanations were tokenized and lemmatized using spaCy (Honnibal et al., 2020). The characteristics of the explanation text were quantified by two measures:

1. *Visual strength* to assess reliance on visual information: The visual strength measure was retrieved from the Lancaster Sensorimotor Norms (Lynott et al., 2020), which include ratings for how much a word is experienced through different perceptual senses or through actions performed by different body parts. For instance, 'black and white stripes' is an explanation with high visual strength. It was computed for all words that had corresponding entries in the Lancaster Sensorimotor Norms.
2. *WordNet similarity to the class label* to reflect reliance on conceptual information: WordNet similarity was calculated with the NLTK interface (Bird et al., 2009) for WordNet (Miller, 1995), which organizes words into sets of synonyms and connects the sets with semantic relations. We used path similarity, which is based on the inverse of the shortest path between two words in the hypernym/hyponym taxonomy, to measure the similarity between each word in the explanations and the label word. For instance, 'chair' has high similarity to 'sofa'. Similarity was calculated for all nouns that were included in WordNet but not for words with other parts of speech since WordNet does not link words with different parts of speech, and the first sense, or meaning, was always used for words with multiple senses.

The two measures were obtained for each single word, and the mean scores were computed for each explanation.

Eye movement analysis

The same EMHMM with co-clustering procedure was used to analyse the 62 participants' eye movements on 160 images during the explanation task. Trials where the participants responded incorrectly on the classification task were excluded. In addition to the co-clustering analysis, the gaze preference of the image area and the textbox area were calculated, respectively, using the average percentage of fixations on the image/textbox area in each trial.

Comparison of the two tasks

Following a previous study (Hsiao, An, et al., 2021), the consistency between the participants' eye movement patterns across the two tasks was examined by analysing the correlation between the A-B scales. In addition, we performed another EMHMM with co-clustering with the eye movement data on the image area from both tasks to discover representative eye movement pattern groups across the two tasks in order to directly compare eye movements in the two tasks. More specifically, the analysis included 61 participants' eye movement data in the classification task and 62 participants' eye movement data in the explanation task, resulting in 123 participant-task combinations. One HMM was generated for each

participant-task combination and each image stimulus. The other parameters and procedures were the same as those used in the analyses of the two tasks separately.

Comparison with XAI saliency maps

In this study, we also aimed to compare human participants' attention strategies in the two tasks with the saliency map-based explanations generated by current XAI methods. We chose a pre-trained ResNet-50 from PyTorch (Paszke et al., 2019) as our image classification AI, which is a convolutional neural network that performs image classification with high accuracy. It extracted visual features from the images for classification without any other types of features (e.g. conceptual information) of the classes.

To compare with XAI saliency maps, we generated human attention maps (i.e. heatmaps). One heatmap was plotted for each task and eye movement pattern group combination, resulting in four heatmaps for each image. Only the fixations on the image area were used and the pattern group assignment was based on the analyses of the two tasks separately. Each heatmap was initialized with a zero matrix that matched the size of the image in pixels, and each fixation point was marked as 1 in the matrix according to its x- and y-coordinates. A Gaussian filter with a standard deviation of 21 pixels (0.5° of visual angle) was applied to the matrix.

We selected commonly used XAI methods, including RISE, PCB corrected RISE, and Grad-CAM, to generate XAI saliency maps. The XAI methods provided explanations for the output of a pre-trained ResNet-50.[4] The matrices of the saliency maps were normalized by setting the largest value to 1 and the smallest value to 0. Two similarity metrics were used to compare the saliency maps with the heatmaps: cosine similarity and KL divergence. Cosine similarity measures the similarity between two vectors in an inner product space by computing the cosine of the angle between them and was calculated using the following formula, where $\mathbf{X}$ and $\mathbf{Y}$ represent the vectors for two maps and $\|\mathbf{X}\|$ and $\|\mathbf{Y}\|$ represent the Euclidean norms of the two vectors:

$$\cos(\theta) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\|\|\mathbf{Y}\|}$$

KL divergence quantifies the difference between two probability distributions and was computed with the following formula, where P and Q^D represent the probability distributions of the two maps and ϵ represents a very small value:

$$\mathrm{KL}(P,\ Q^D) = \sum_i Q_i^D \log\left(\epsilon + \frac{Q_i^D}{\epsilon + P_i}\right)$$

RESULTS

Image classification task

EMHMM with co-clustering resulted in two representative pattern groups: Explorative (Group A, with larger ROIs, exploring a wider region of an image) vs. Focused (Group B, with smaller ROIs, focusing on a certain part of an image) Pattern Groups (Figure 3a). The two groups differed significantly based on KL divergence estimation (Chuk et al., 2014): $F(1, 59) = 187.15$, $p < .001$, $\eta_p^2 = .76$, 90% CI = [0.66, 0.81].[5] Explorative participants had a significantly larger number of fixations than

[4]Due to the input constraints of ResNet-50, all the images were adjusted to the size of 217 × 217 pixels before generating the heatmaps and the saliency maps.

[5]90% CI instead of 95% CI is reported for F-tests since F-tests are one-sided (Steiger, 2004).

(a) Classification Task

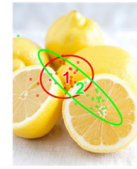
Group A: Explorative (N = 34)

Group A	To R	To G
Priors	.83	.17
From Red	.13	.87
From Green	.25	.75

Group A	To R	To G
Priors	.81	.19
From Red	.06	.94
From Green	.07	.93

Group B: Focused (N = 27)

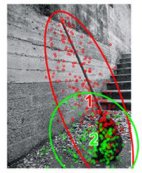
Group B	To R	To G
Priors	.92	.08
From Red	.16	.84
From Green	.36	.64



Group B	To R	To G
Priors	.71	.29
From Red	.04	.96
From Green	.10	.90

(b) Explanation Task

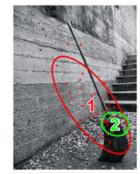
Group A: Explorative (N = 47)



Group A	To R	To G
Priors	.89	.11
From Red	.96	.04
From Green	.10	.90

Group A	To R	To G
Priors	.88	.12
From Red	1.0	.00
From Green	.00	1.0

Group B: Focused (N = 15)

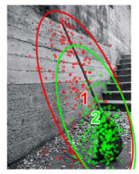


Group B	To R	To G
Priors	.91	.09
From Red	.96	.04
From Green	.08	.92

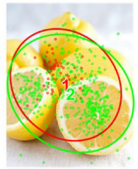
Group B	To R	To G
Priors	.78	.22
From Red	.96	.04
From Green	.08	.92

(c) Classification and Explanation Tasks Together

Group A: Explorative (N = 60)



Group A	To R	To G
Priors	1.0	.00
From Red	.42	.58
From Green	.47	.53



Group A	To R	To G
Priors	.99	.01
From Red	.24	.76
From Green	.08	.02

Group B: Focused (N = 63)

Group B	To R	To G
Priors	.94	.06
From Red	.32	.68
From Green	.38	.62

Group B	To R	To G
Priors	.91	.09
From Red	.25	.75
From Green	.39	.61

FIGURE 3 Examples of EMHMM co-clustering results of broom and lemon images for (a) the classification task only, (b) the explanation task only and (c) two tasks together. Ellipses show ROIs as 2-D Gaussian emissions. The table shows transition probabilities among the ROIs, and priors show the probabilities that a fixation sequence starts from the ellipse. In each pattern, the image on the right shows raw fixations and their ROI assignment.

focused participants, $t(59) = 3.49$, $p < .001$, $d = 0.90$, 95% CI [0.34, 1.44]. They did not differ in average fixation duration, $t(59) = -0.41$, $p = .685$, $d = -0.11$, 95% CI [−0.61, 0.40], or in eye movement entropy (consistency), $t(58.10) = -0.42$, $p = .674$, $d = -0.11$, 95% CI [−0.61, 0.40]. We quantified participants' eye movement pattern using A-B scale, which we referred to as EF scale (Explorative-Focused scale).

English proficiency, as measured by LexTALE, was correlated with both classification accuracy, $r(60) = .28$, $p = .026$, and RT, $r(55) = -.36$, $p = .007$. ANCOVA analyses examining the effect of eye movement pattern group on accuracy and RT with LexTALE as a covariate showed a significant effect of group in RT, $F(1, 52) = 6.66$, $p = .013$, $\eta_p^2 = .11$, 90% CI = [0.01, 0.25], but not in accuracy. Participants in the focused group had shorter RT, $M_{adj} = 945$, $SE_{adj} = 19.3$, than those in the explorative group, $M_{adj} = 1015$, $SE_{adj} = 18.9$, $t(52) = 2.58$, $p = .013$, $d = 0.70$, 95% CI [0.14, 1.25]. Hierarchical multiple regression predicting classification RT showed that, at stage one, cognitive ability measures and LexTALE jointly explained 30.6% of the variance; the regression model was not significant, $F(13, 40) = 1.36$, $p = .223$. Adding EF scale accounted for an additional 6.4% of the variance and this change was marginally significant, $F(1, 39) = 3.96$, $p = .054$. Together the results suggested that the focused attention strategy was associated with shorter image classification RT.

Explanation task

Similarly, we observed Explorative and Focused Pattern Groups (Figure 3b), which differed significantly from each other: $F(1, 60) = 122.77$, $p < .001$, $\eta_p^2 = .67$, 90% CI = [0.55, 0.74]. Explorative participants had significantly more fixations per trial, $t(57.78) = 6.57$, $p < .001$, $d = 1.53$, 95% CI [1.26, 2.62], longer average fixation duration, $t(60) = 2.34$, $p = .022$, $d = 0.70$, 95% CI [0.09, 1.29] and higher eye movement entropy (lower consistency), $t(37.01) = 8.34$, $p < .001$, $d = 2.20$, 95% CI [1.66, 3.27] than focused participants. In addition, explorative participants had more fixations on the image region, $t(60) = 4.02$, $p < .001$, $d = 1.19$, 95% CI [0.56, 1.82], but fewer fixations on the textbox region, $t(60) = 3.38$, $p = .001$, $d = 1.00$, 95% CI [0.38, 1.61], than focused participants.

For explanation performance, English proficiency was significantly correlated with effectiveness, $r(60) = .28$, $p = .029$, but not with diagnosticity, $r(60) = .10$, $p = .430$. ANCOVA analyses examining the effect of the eye movement pattern group on these two explanation performance measures with LexTale controlled showed significant differences in effectiveness, $F(1, 59) = 12.71$, $p < .001$, $\eta_p^2 = .18$, 90% CI [0.05, 0.31], and diagnosticity, $F(1, 59) = 16.74$, $p < .001$, $\eta_p^2 = .22$, 90% CI [0.08, 0.36]: explorative participants' explanations were rated higher for effectiveness, $M_{adj} = 4.4$, $SE_{adj} = 0.09$, than focused participants', $M_{adj} = 3.7$, $SE_{adj} = 0.17$, $t(59) = 3.57$, $p < .001$, $d = 1.06$, 95% CI [0.44, 1.69]. In contrast, focused participants' explanations had higher diagnosticity, $M_{adj} = 0.63$, $SE_{adj} = 0.02$, than explorative participants', $M_{adj} = 0.53$, $SE_{adj} = 0.01$, $t(59) = 4.09$, $p < .001$, $d = 1.22$, 95% CI [0.58, 1.85]. Consistent with these findings, EF scale was positively correlated with effectiveness, $r(60) = .45$, $p < .001$ (Figure 4a), and negatively correlated with diagnosticity, $r(60) = -.42$, $p < .001$ (Figure 4b).

Hierarchical multiple regression analyses predicting explanation effectiveness showed that, at stage one, the cognitive ability measures and LexTALE contributed significantly to the regression model, $\Delta R^2 = 39.2\%$, $F(13, 47) = 2.33$, $p = .017$, and at stage two EF scale significantly explained additional variations, $\Delta R^2 = 9.9\%$, $F(1, 46) = 8.98$, $p = .004$. For predicting diagnosticity, cognitive ability measures and LexTALE did not contribute significantly to the regression model at stage one, $\Delta R^2 = 28.2\%$, $F(13, 47) = 1.42$, $p = .186$, while EF scale significantly explained additional variance at stage two, $\Delta R^2 = 14.0\%$, $F(1, 46) = 11.16$, $p = .002$. Thus, after taking English proficiency and cognitive abilities into account, the explorative strategy was associated with explanations that were rated higher in effectiveness for teaching image classification, whereas the focused strategy was associated with explanations with higher diagnosticity for inferring class labels without seeing the image.

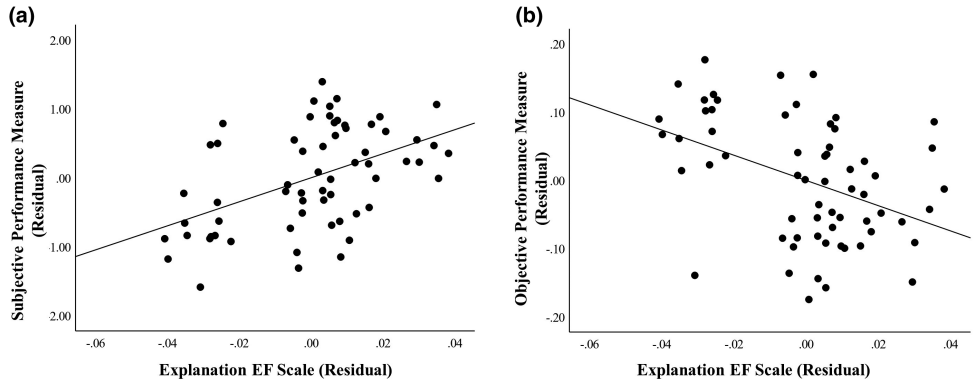


FIGURE 4 Correlation between EF scale in the explanation task and explanation performance as evaluated by (a) effectiveness and (b) diagnosticity.

Visual strength was positively correlated with EF scale, $r(60) = .44$, $p < .001$, and effectiveness, $r(60) = .50$, $p < .001$, but negatively correlated with diagnosticity, $r(60) = -.36$, $p = .004$. In contrast, WordNet similarity was negatively correlated with EF scale, $r(60) = -.29$, $p = .023$, and effectiveness, $r(60) = -.42$, $p < .001$, but positively correlated with diagnosticity, $r(60) = .31$, $p = .013$. These results suggested that explorative strategies were associated with more use of visual information and less use of conceptual information. In addition, explanations with more visual information tended to be rated higher for effectiveness, while those with more conceptual information tended to have higher diagnosticity.

Comparison of the two tasks

No significant correlation was found between participants' EF scale of the image classification task and the explanation task, $r(59) = .16$, $p = .210$, suggesting that participants did not use consistent attention strategies across the two tasks. To compare attention strategies in the two tasks directly, we used EMHMM with co-clustering on participants' eye movement patterns in both tasks together. The results showed similar Explorative and Focused Pattern Groups (Figure 3c), which differed significantly from each other: $F(1, 121) = 294.30$, $p < .001$, $\eta_p^2 = .71$, 90% CI = [0.64, 0.76]. Explorative Pattern Group had higher entropy (lower consistency) than Focused Pattern Group, $t(78.47) = 13.95$, $p < .001$, $d = 2.54$, 95% CI [1.98, 3.04]. Interestingly, participants' eye movement patterns were more explorative during explanation than during classification, $t(60) = 11.95$, $p < .001$, $d = 1.53$, 95% CI [1.16, 1.90].

Comparison with XAI saliency maps

Some examples of human attention maps and XAI saliency maps were presented in Figure 5. As shown in Table 1, a three-way interaction between task, strategy and XAI method was found in both similarity measures. When we split the data by XAI methods (Figure 6), main effects of task and strategy, and an interaction between task and strategy were consistently found across the XAI methods (ps < .001). XAI saliency maps had higher similarity to human attention maps during the explanation task, particularly for the explorative strategy, which was associated a higher reliance on visual information.

We then split the data by task and strategy to examine the effect of XAI method. In all combinations of task and strategy, we found a main effect of XAI method in both similarity measures (ps < .001): Saliency maps from PCB corrected RISE had the highest similarity to human attention maps, followed

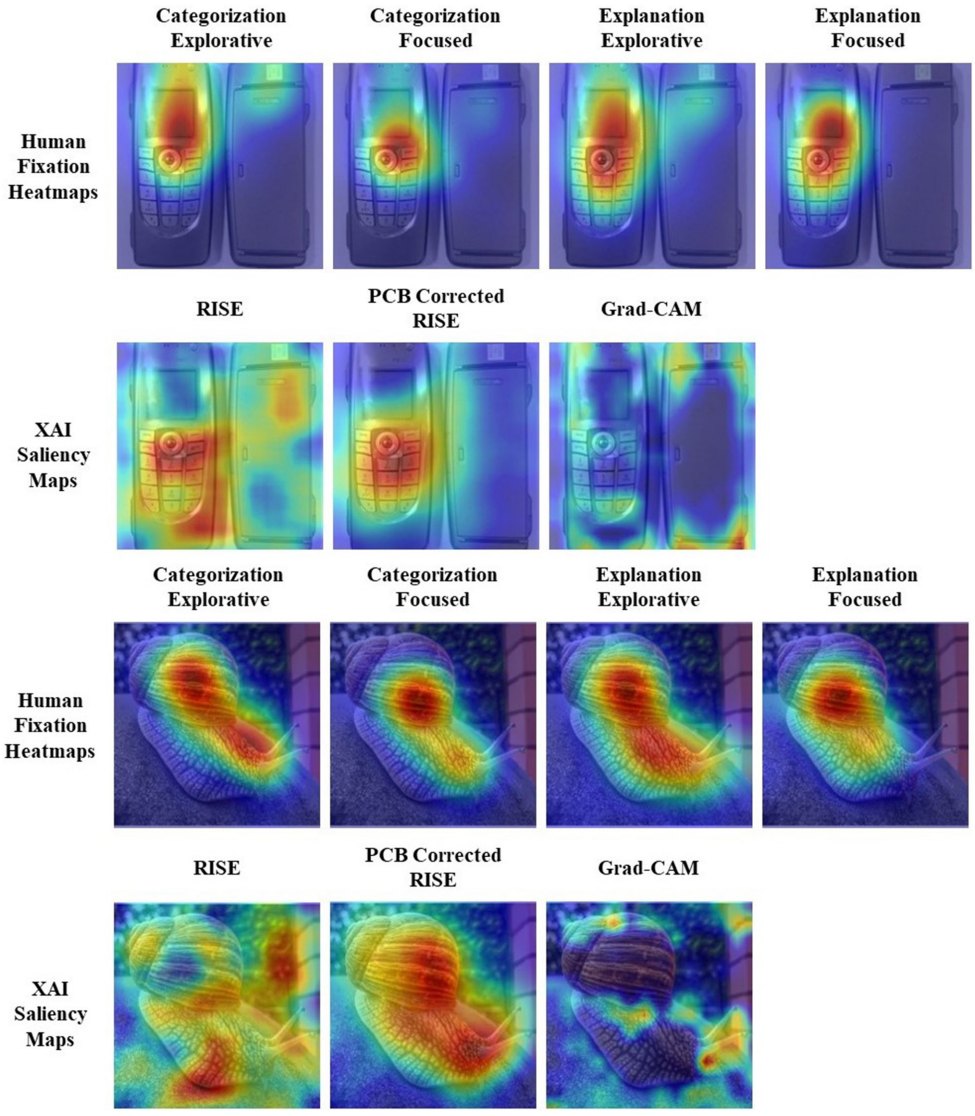


FIGURE 5 Example human attention maps and XAI saliency maps, with one image from an artificial category (cellphone) and one image from a natural category (snail).

by RISE, and then by Grad-CAM (Figure 7),[6] suggesting that saliency maps generated by perturbation-based XAI methods have higher similarity to human attention maps than those from gradient-based methods.

[6]In a separate analysis, we compared XAI maps with human-segmented object masks as a control condition. We took the human attention maps from those with explorative strategies during the explanation task, which had the highest similarity to XAI maps, and performed ANOVA to examine whether saliency maps from different XAI methods (Grad-CAM, RISE, PCB-corrected RISE and human-segmented object masks) differed in their similarity to the human attention maps. When using either cosine similarity or KL divergence as the similarity measure, we found that the similarity of human-segmented object masks to human attention maps was higher than those from Grad-CAM (cosine similarity: $t(159) = 13.28$, $p < .001$; KL divergence: $t(159) = 9.48$, $p < .001$), but was not significantly different from RISE (cosine similarity: $t(159) = 1.66$, $p = .347$; KL divergence: $t(159) = 2.19$, $p = .130$) and was lower than PCB corrected RISE (cosine similarity: $t(159) = 5.67$, $p < .001$; KL divergence: $t(159) = 7.18$, $p < .001$). This result suggested that PCB corrected RISE's higher similarity to human attention could not be completely accounted for by the object containing the saliency map/object segmentation. Indeed, human attention for image classification typically focuses on important features for the task, rather than simply following object segmentation.

TABLE 1 Results of the task × eye movement strategy × XAI method ANOVA on cosine similarity and KL divergence (*$p < .05$, **$p < .01$, ***$p < .001$).

	Cosine similarity			KL divergence		
Effect	F	p	η_p^2	F	p	η_p^2
Task	240.46	<.001***	.60	133.81	<.001***	.46
Strategy	296.74	<.001***	.65	350.14	<.001***	.60
XAI Method	265.87	<.001***	.63	176.61	<.001***	.53
Task × Strategy	243.68	<.001***	.61	124.38	<.001***	.44
Task × XAI Method	39.97	<.001***	.20	18.88	<.001***	.11
Strategy × XAI Method	2.80	.091	.02	66.15	<.001***	.29
Task × Strategy × XAI Method	76.44	<.001***	.32	11.79	<.001***	.07

Note: The Greenhouse–Geisser correction was used for all of the effects that involved XAI methods due to violations of the sphericity assumption.

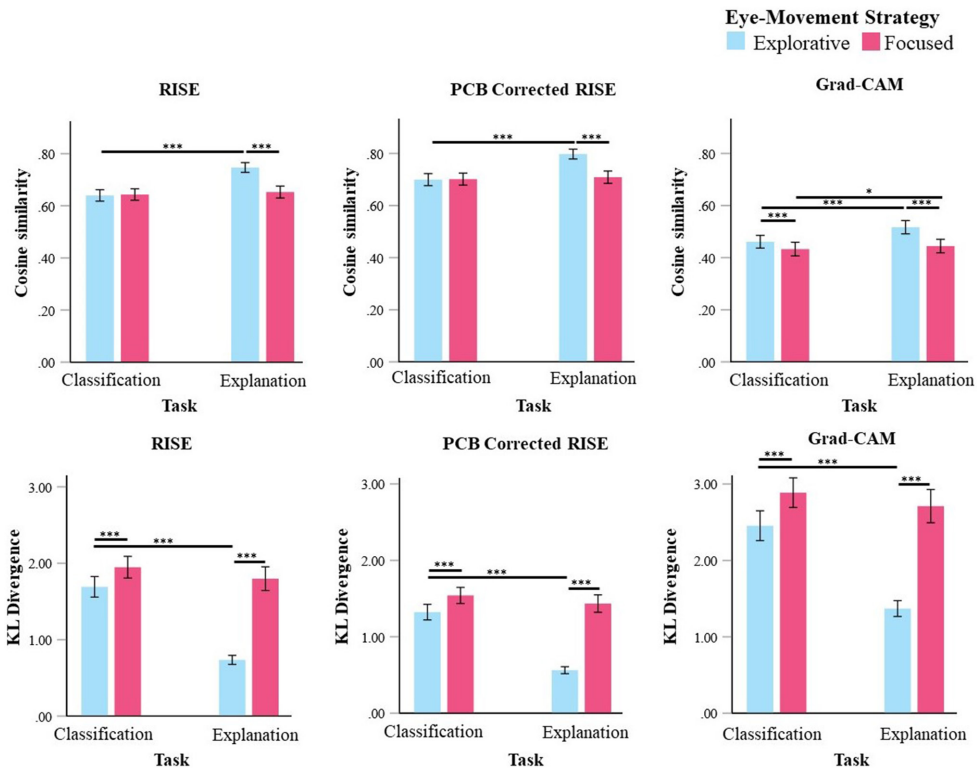


FIGURE 6 Difference in cosine similarity (row above) and KL divergence (row below) between the two strategies for the two tasks and each of the three XAI methods (error bars: 95% CI; *$p < .05$, **$p < .01$, ***$p < .001$). Note that greater similarity is indicated by higher cosine similarity and lower KL divergence.

DISCUSSION

Here, we examined human attention strategies for image classification and for explaining image classification and compared them with current saliency map-based XAI explanations. Using EMHMM, we discovered focused (focused visual scanning on the foreground object) and explorative (explorative scanning at a broader region) attention strategies in both classification and explanation tasks.

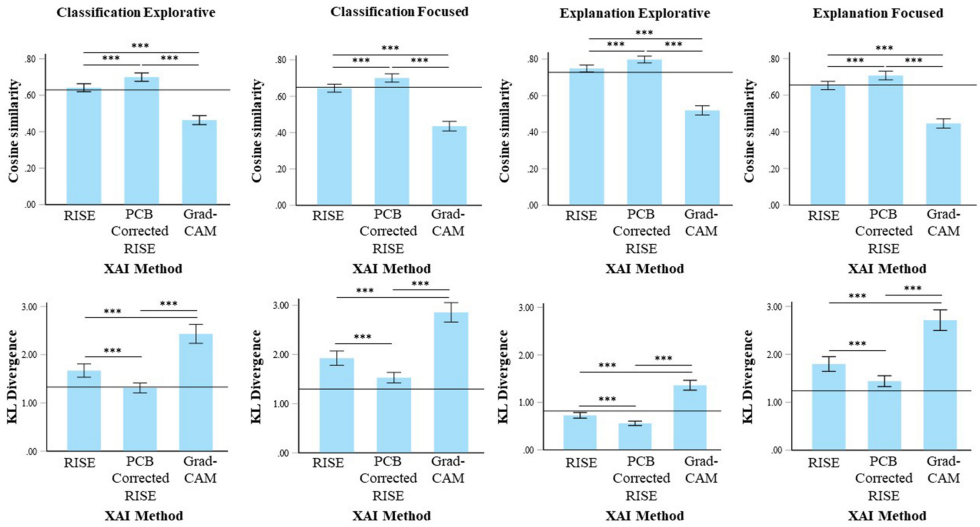


FIGURE 7 Difference among the three XAI methods (RISE, PCB Corrected RISE, Grad-CAM) in cosine similarity (row above) and KL divergence (row below) in each of the four task and strategy combinations (error bars: 95% CI; $*p < .05$, $**p < .01$, $***p < .001$). The black reference lines refer to the average similarity between human-segmented object masks and human attention maps for comparison purposes.

Participants did not consistently adopt the same strategies across the two tasks, and they adopted more explorative strategies for explanation than the classification task itself. This result suggested humans adjust their attention strategies according to the task demand (Chuk, Chan, & Hsiao, 2017; Hsiao, An, et al., 2021; Kanan et al., 2015). In addition, in image classification, focused strategies predicted faster responses. In contrast, in explanation, focused strategies were associated with explanations with higher diagnosticity, that is containing more specific information for inferring image classes, whereas explorative strategies were associated with higher frequency to attend to the image region and explanations rated higher for effectiveness for early category learning. Interestingly, current saliency-based XAI explanations were more similar to human attention strategies in the explanation task, especially the explorative strategies, rather than those during image classification. In particular, saliency maps generated by perturbation-based XAI methods including PCB corrected RISE and RISE, which highlight input features that lead to output class probability change when being perturbed, had higher similarity to human attention maps than the backpropagation-based XAI method Grad-CAM. This result was consistent with previous research suggesting causal reasoning based on observed regularities as an important feature in human explanations (Einhorn & Hogarth, 1986; Holzinger et al., 2019; Zemla et al., 2017).

Theoretical implications

The finding that participants used more focused attention strategies for image classification and more explorative strategies for explanation was consistent with our hypothesis that human attention strategies during explanation may cover more relevant features than those during image classification itself. More specifically, image classification may require attention to just sufficient information for making a classification decision (Hsiao & Cottrell, 2008; Smith & Ratcliff, 2004), whereas explanation may require attention to as much relevant information as possible to be comprehensive (Gelman et al., 1998). Consistent with this finding, here, we found that in image classification, a more focused attention strategy had faster response speed, suggesting that focusing on identifying critical features of the foreground object is beneficial for classification.

In explanation, a more explorative attention strategy was associated with higher ratings in effectiveness for novel category learning, higher frequency to attend to the image region than the textbox region, and more use of visual information in the explanation text, as compared with the focused strategy. This result suggested that explorative scanning for relevant visual features to the object class is beneficial for providing explanations for early category learning. In contrast, explanations with a more focused scanning on the foreground object were rated lower in effectiveness but were associated with higher diagnosticity for inferring the class label and more use of conceptual information in the explanation text. We also found that visual information enhanced effectiveness, consistent with the previous finding that visual information is more important for early category learning (Kloos & Sloutsky, 2008). In contrast, conceptual information made it easier for people who already knew the image classes to guess the class label, consistent with the finding that conceptual information is more dominant in category representations developed later (Fisher & Sloutsky, 2005). Thus, attention strategies revealed participants' preference in information use during explanation, which was in turn associated with explanations that served different purposes. EMHMM allowed us to discover representative attention strategies and quantify individual strategies, leading to these novel findings.

Although saliency map-based XAI methods are designed to highlight features used by AI for performing the task, we found that saliency maps generated by XAI for image classification had higher similarity to human attention strategies during explanation than during the image classification task itself. This finding suggests that the current XAI methods highlight all features that are relevant to AI's classification decision, similar to how humans explain image classification. While this finding may also suggest that AI uses more features for image classification than humans, it is important to note that AI's decision processes can be fundamentally different from humans'. More specifically, a fundamental difference between humans and AI is in their attention mechanisms: humans process bits of visual information at a time through a sequence of eye fixations, whereas AI models do not have this visual anatomy constraint and can process all information simultaneously (Hsiao et al., 2022). Thus, human decisions involve accumulation of evidences sequentially (Lee & Cummins, 2004), whereas in AI all relevant information can be processed in parallel (Raschka et al., 2020).

We also found that XAI saliency maps had higher similarities to the explorative than the focused attention strategy during human explanations, and in humans' explorative strategy was associated with higher reliance on visual information. Indeed, another difference between human and AI image classification is the type of information available for decision-making: the AI model under examination is designed to use visual information only; in contrast, human representations for object classes contain both visual and conceptual information (Martin et al., 2018), which can be flexibly and selectively attended to for decision-making. Also, in human category learning, both visual exemplars and verbal explanations play an important role: verbal explanations provide crude rules for the category structure, while visual exemplars can be used for finer adjustments based on these rules (Moskvichev et al., 2019).

Through comparing human attention maps with XAI saliency maps, we found that the XAI saliency maps generated by perturbation-based methods (RISE and PCB corrected RISE) consistently had higher similarity to human attention maps than those from the backpropagation-based method (Grad-CAM). This result suggested that human attention strategies during explanation may be more similar to the perturbation-based than the backpropagation-based XAI approach. Perturbation-based methods highlight input features that have causal influence on the classification output probability. In contrast, backpropagation-based methods highlight features according to the gradient output class score in a particular input layer. Our findings are consistent with the literature where human explanations are typically characterized by the emphasis on observable causality (Einhorn & Hogarth, 1986; Holzinger et al., 2019).

Practical implications

Our results showed that humans can use both visual and conceptual information for explaining image classification. This finding has important implications for developing human-accessible explanations

in XAI. For example, in addition to XAI saliency maps, some recent studies have developed concept-based approaches to provide explanations using human-friendly concepts (e.g. striped, curly, etc.) for image classification models (Kim et al., 2018). Our findings suggest that visual and conceptual explanations serve different purposes and thus both are important for providing human-accessible explanations.

We also found that human attention strategies during explanation had higher similarity to the perturbation-based than the backpropagation-based XAI approach. This finding suggested that XAI saliency maps generated by perturbation-based methods may better match human attention strategy for explanation, potentially more accessible to AI users. Indeed, in human category learning, contrastive explanations with exemplars highlighting features discriminative of different categories have been shown to facilitate learning performance as compared with non-contrastive explanations with within-category exemplars (Hammer et al., 2009; Kang & Pashler, 2012; Nosofsky & McDaniel, 2019). Thus, saliency maps from perturbation-based methods may better facilitate user understanding of AI than backpropagation-based methods. Future work may examine this possibility.

Another practical implication from our study is related to evaluating and benchmarking XAI saliency maps. Recent research has proposed to use human attention (Liu et al., 2023, 2024; Mohseni et al., 2021; Zhao et al., 2024) as a benchmark for evaluating plausibility of saliency map-based explanations. Our results suggested that we need to take individual differences into account when developing such benchmarks. More specifically, since individuals differ significantly in attention strategies in explanation, which were associated with different aspects of explanation quality, we may develop different benchmarks that better suit the explanation needs. Since collecting a large amount of human attention data for benchmarking purposes is often time-consuming, Yang et al. (2022) developed a Human Saliency Imitator model to automatically generate a human attention map given an input image using a deep learning model trained with human attention data with high accuracy (Pearson Correlation Coefficient = .88 on validation). These simulated data can also be used for developing other applications that require human attention data such as human-in-the-loop systems (Gil et al., 2019), demonstrating the importance of simulated data as a new trend in AI/cognitive science research (De Melo et al., 2021).

Our discovery of different explanation strategies from human explainers also suggested that explainees may differ in the type of explanations that is more accessible to them, and human explainers may adjust their strategy according to explainees' needs (Kaufman & Kirsh, 2022; Strauss & Ziv, 2012). Thus, future XAI development may consider learners' preferences for providing more accessible explanations. Indeed, most recent XAI research has started to consider the importance of inferring humans' mental state when providing explanations (Hsiao, 2024; Hsiao & Chan, 2023), as inspired by an important cognitive capacity in human social interaction, theory of mind (i.e. the capacity to understand others' behaviour by attributing mental states to them; e.g. Akula et al., 2022). For example, it may be beneficial to use more visual information when explaining novel categories or explaining to young children without much category knowledge. In addition, Hammer et al. (2009) discovered that in contrast to older children and adults, young children had difficulties with identifying between-category differences and thus learned better through comparing same-class exemplars. In this case, prototype-based XAI methods, which use the most representative objects of the category as explanation exemplars, may be more suitable. Indeed, humans are shown to prefer prototype-based approaches during early category learning (Minda & Smith, 2001; Smith & Minda, 1998).

Limitations and future work

Note that in the current study, we selected Grad-CAM as the representative backpropagation-based method and RISE and PCB-RISE as the representative perturbation-based methods to be compared with human explanation strategies. It remains unclear how other saliency-based XAI methods such as LIME (Ribeiro et al., 2016; it can be considered as a perturbation-based method, Das & Rad, 2020,

or a surrogate model method, Sokol et al., 2019) are compared with human attention strategies during explanation observed here. This can be examined in future work.

Our results, here, have demonstrated the similarities and differences between XAI and human explanation strategies. Humans and XAI may also work together to improve explanation quality. This concept of human-in-the-loop system has been adopted in AI design. For example, task-driven human attention can be integrated into AI systems to boost their performance, especially when humans provide better information extraction strategies (Lai et al., 2020; Rong et al., 2021). Future studies may explore how human attention can be incorporated into XAI methods to make XAI's explanations more accessible to humans. For instance, saliency-based XAI methods highlight important regions without providing a logical temporal sequence for users to understand the links among them (Kaufman & Kirsh, 2022). Integrating human attention, which contains temporal information, may help guide users to reach better comprehension. XAI with the ability to infer user strategy may compare it with AI's strategy and inform user when to trust or not to trust AI.

CONCLUSION

In conclusion, here, we showed that human explanation for image classification involves exploring more relevant features than the classification task itself, which only requires sufficient information for decision-making. Humans also differed in the use of explorative vs. focused attention strategies during explanation. These strategies were associated with differential reliance on visual and conceptual information in the explanation that served different purposes. The finding that features used by AI as revealed by current saliency-based XAI methods had the highest similarity to the explorative explanation strategy in humans demonstrated a fundamental difference between AI and human: AI could use all relevant information in parallel, whereas human attention involved sequential processing to accumulate evidence. Interestingly, XAI saliency maps that highlight discriminative features informing causality matched better with human attention strategies for explanation, suggesting that establishing causality characterizes human explanation and can potentially make explanations more accessible to AI users. These findings have important implications for developing user-centred XAI methods to enhance human-AI interaction.

AUTHOR CONTRIBUTIONS

Janet H. Hsiao: Conceptualization; methodology; data curation; supervision; project administration; writing – original draft; writing – review and editing; funding acquisition. **Ruoxi Qi:** Methodology; investigation; formal analysis; data curation; writing – original draft; visualization. **Yueyuan Zheng:** Investigation; writing – original draft; methodology; formal analysis; data curation; visualization. **Yi Yang:** Visualization; writing – review and editing. **Caleb Chen Cao:** Conceptualization.

ACKNOWLEDGEMENTS

This study was supported by Huawei. The eye tracker used was supported by RGC of Hong Kong (No. C7129-20G to Hsiao). We thank Van Kie Liew for data processing, Yundi Yang for data collection and preprocessing, Yunke Chen for data preprocessing, Yannie Lim for data collection, Luyu Qiu and Jindi Zhang for their advice.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial and/or non-financial interests.

DATA AVAILABILITY STATEMENT

Data and codes of this study that are sufficient to build on the results presented in this paper will be made available on a permanent third-party archive upon journal publication.

ORCID

Ruoxi Qi https://orcid.org/0009-0008-5820-7502
Yueyuan Zheng https://orcid.org/0000-0002-0913-9514
Yi Yang https://orcid.org/0009-0001-8224-1232
Janet H. Hsiao https://orcid.org/0000-0003-2271-8710

REFERENCES

Akata, Z. (2013). Label embedding for image classification. In *Proceedings of the 2013 IEEE conference on computer vision and pattern recognition* (pp. 819–826). IEEE.

Akula, A. R., Wang, K., Liu, C., Saba-Sadiya, S., Lu, H., Todorovic, S., Chai, J., & Zhu, S. C. (2022). CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *iScience*, *25*(1), 103581.

An, J., & Hsiao, J. H. (2021). Modulation of mood on eye movement pattern and performance in face recognition. *Emotion*, *21*(3), 617–630.

Balcikanli, C. (2011). Metacognitive awareness inventory for teachers (MAIT). *Electronic Journal of Research in Educational Psychology*, *9*(3), 1309–1332.

Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., & Intriligator, J. M. (2006). Information processing during face recognition: The effects of familiarity, inversion, and morphing on scanning fixations. *Perception*, *35*, 1089–1105.

Bender, A. (2020). What is causal cognition? *Frontiers in Psychology*, *11*, 3.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. O'Reilly Media.

Caldara, R., & Miellet, S. (2011). iMap: A novel method for statistical fixation mapping of eye movement data. *Behavior Research Methods*, *43*, 864–878.

Chan, C. Y. H., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2018). Eye movement patterns in face recognition are associated with cognitive decline in older adults. *Psychonomic Bulletin & Review*, *25*(6), 2200–2207.

Chan, F. H. F., Barry, T. J., Chan, A. B., & Hsiao, J. H. (2020). Understanding visual attention to face emotions in social anxiety using hidden Markov models. *Cognition and Emotion*, *34*(8), 1704–1710.

Chin-Parker, S., & Cantelon, J. (2017). Contrastive constraints guide explanation-based category learning. *Cognitive Science*, *41*(6), 1645–1655.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *Journal of Vision*, *14*(11), 8.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2017). Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling. *Vision Research*, *141*, 204–216.

Chuk, T., Crookes, K., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2017). Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures. *Cognition*, *169*, 102–117.

Cover, T. M., & Thomas, J. A. (2006). Entropy, relative entropy and mutual information. In *Elements of information theory* (pp. 12–49). John Wiley & Sons.

Coviello, E., Chan, A. B., & Lanckriet, G. R. G. (2014). Clustering hidden Markov models with variational HEM. *Journal of Machine Learning Research*, *15*(22), 697–747.

Das, A., & Rad, P. (2020). *Opportunities and challenges in explainable artificial intelligence* (*xai*): *A survey*. *arXiv preprint arXiv*:*2006.11371*.

De Melo, C. M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., & Hodgins, J. (2021). Next-generation deep learning based on simulators and synthetic data. *Trends in Cognitive Sciences*, *26*(2), 174–187.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3–19.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fisher, A. V., & Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. *Child Development*, *76*, 583–597.

Gelman, S. A., Coley, J. D., Rosengren, K. S., Hartman, E., & Pappas, A. (1998). Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child Development*, *63*(1), 1–148.

Gil, M., Albert, M., Fons, J., & Pelechano, V. (2019). Designing human-in-the-loop autonomous cyber-physical systems. *International Journal of Human-Computer Studies*, *130*, 21–39.

Goyal, Y., Mohapatra, A., Parikh, D., & Batra, D. (2016). *Towards transparent AI systems*: *Interpreting visual question answering models*. arXiv. https://arxiv.org/abs/1608.08974

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120.

Hammer, R., Diesendruck, G., Weinshall, D., & Hochstein, S. (2009). The development of category learning strategies: What makes the difference? *Cognition*, *112*(1), 105–119.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for explainable AI: Challenges and prospects*. arXiv. https://arxiv.org/abs/1812.04608

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python* [Python].

Hsiao, J. H. (2024). Understanding human cognition through computational modeling. *Topics in Cognitive Science*. Online ahead of print. https://doi.org/10.1111/tops.12737

Hsiao, J. H., An, J., Hui, V. K. S., Zheng, Y., & Chan, A. B. (2022). Understanding the role of eye movement consistency in face recognition and autism through integrating deep neural networks and hidden Markov models. *npj Science of Learning*, *7*(1), 28.

Hsiao, J. H., An, J., Zheng, Y., & Chan, A. B. (2021). Do portrait artists have enhanced face processing abilities? Evidence from hidden Markov modeling of eye movements. *Cognition*, *211*, 104616.

Hsiao, J. H., & Chan, A. B. (2023). Towards the next generation explainable AI that promotes AI-human mutual understanding. *NeurIPS XAIA 2023*. https://openreview.net/forum?id=d7FsEtYjvN

Hsiao, J. H., & Chan, A. B. (2023). Visual attention to own- versus other-race faces: Perspectives from learning mechanisms and task demands. *British Journal of Psychology*, *114*(S1), 17–20. https://doi.org/10.1111/bjop.12647

Hsiao, J. H., Chan, A. B., An, J., Yeh, S.-L., & Jingling, L. (2021). Understanding the collinear masking effect in visual search through eye tracking. *Psychonomic Bulletin & Review*, *28*(6), 1933–1943.

Hsiao, J. H., & Cottrell, G. W. (2008). Two fixations suffice in face recognition. *Psychological Science*, *9*(10), 998–1006.

Hsiao, J. H., Lan, H., Zheng, Y., & Chan, A. B. (2021). Eye movement analysis with hidden Markov models (EMHMM) with co-clustering. *Behavior Research Methods*, *53*, 2473–2486.

Hsiao, J. H., Ngai, H. H. T., Qiu, L., Yang, Y., & Cao, C. C. (2021). *Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI)*. arXiv. https://arxiv.org/abs/2108.01737

Hwu, T., Levy, M., Skorheim, S., & Huber, D. (2021). *Matching representations of explainable artificial intelligence and eye gaze for human-machine interaction*. arXiv. https://arxiv.org/abs/2102.00179

Jiang, Y., Ma, L., & Gao, L. (2016). Assessing teachers' metacognition in teaching: The teacher metacognition inventory. *Teaching and Teacher Education*, *59*, 403–413.

Kanan, C., Bseiso, D., Ray, N., Hsiao, J. H., & Cottrell, G. (2015). Humans have idiosyncratic and task-specific scanpaths for judging faces. *Vision Research*, *108*, 67–76.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*, 97–103.

Karim, M. M., Li, Y., & Qin, R. (2022). Toward explainable artificial intelligence for early anticipation of traffic accidents. *Transportation Research Record*, *2676*(6), 743–755.

Kaufman, R. A., & Kirsh, D. (2022). Cognitive differences in human and AI explanation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*, 2694–2700.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (pp. 2668–2677). PMLR.

Kindermans, P. J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). The (un) reliability of saliency methods. Explainable AI: Interpreting, explaining and visualizing deep learning (pp. 267–280).

Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, *137*, 52–72.

Lai, Q., Khan, S., Nie, Y., Sun, H., Shen, J., & Shao, L. (2020). Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, *23*, 2086–2099.

Lanfredi, R. B., Arora, A., Drew, T., Schroeder, J. D., & Tasdizen, T. (2021). *Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest x-rays*. arXiv. https://arxiv.org/abs/2112.11716

Lau, E. Y. Y., Eskes, G. A., Morrison, D. L., Rajda, M., & Spurr, K. F. (2010). Executive function in patients with obstructive sleep apnea treated with continuous positive airway pressure. *Journal of the International Neuropsychological Society*, *16*(6), 1077–1088.

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. *Psychonomic Bulletin and Review*, *11*, 343–352.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325–343.

Li, X.-H., Shi, Y., Li, H., Bai, W., Cao, C. C., & Chen, L. (2021). An experimental study of quantitative evaluations on saliency methods. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 3200–3208). Association for Computing Machinery.

Liao, W., Li, S. T. K., & Hsiao, J. H. W. (2022). Music reading experience modulates eye movement pattern in English reading but not in Chinese reading. *Scientific Reports*, *12*(1), 9144.

Liu, G., Zhang, J., Chan, A. B., & Hsiao, J. (2023). Human attention-guided explainable AI for object detection. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*.

Liu, G., Zhang, J., Chan, A. B., & Hsiao, J. H. (2024). Human attention guided explainable artificial intelligence for computer vision models. *Neural Networks*, *177*, 106392. https://doi.org/10.1016/j.neunet.2024.106392

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster sensorimotor norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291.

Malle, B. F., & Knobe, J. (1997). Which behaviors do people explain? A basic actor–observer asymmetry. *Journal of Personality and Social Psychology*, *72*, 288–304.

Markman, A. B., & Wisniewski, E. J. (1997). Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 54.

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *life*, *7*, e31873.

Maxwell, J. A. (2004). Using qualitative methods for causal explanation. *Field Methods*, *16*(3), 243–264.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Miller, T. (2021). Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, *36*, E14.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775.

Mohseni, S., Block, J. E., & Ragan, E. (2021). Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In *26th international conference on intelligent user interfaces* (pp. 22–31). Association for Computing Machinery.

Moskvichev, A., Tikhonov, R., & Steyvers, M. (2019). A picture is worth 7.17 words: Learning categories from examples and definitions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *41*, 2406–2412.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). *Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI*. arXiv. https://arxiv.org/abs/1902.01876

Nosofsky, R. M., & McDaniel, M. A. (2019). Recommendations from cognitive psychology for enhancing the teaching of natural-science categories. *Policy Insights From the Behavioral and Brain Sciences*, *6*(1), 21–28.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd international conference on neural information processing systems* (pp. 8026–8037). Neural Information Processing Systems Foundation, Inc.

Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological Science*, *24*(7), 1216–1225.

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the 2018 British machine vision conference* (p. 151). BMVA Press.

Phillips, L. H., Wynn, V. E., McPherson, S., & Gilhooly, K. J. (2001). Mental planning and the tower of London task. *The Quarterly Journal of Experimental Psychology*, *54*(2), 579–597.

Qi, R., Zheng, Y., Yang, Y., Zhang, J., & Hsiao, J. H. (2023). Individual differences in explanation strategies for image classification and implications for explainable AI. In *Proceedings of the 45th annual conference of the cognitive science society* (pp. 1644–1651). Cognitive Science Society.

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, *11*(4), 193.

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, *29*(9), 2352–2449.

Ridderinkhof, R. K., Band, G. P. H., & Logan, G. D. (1999). A study of adaptive behavior: Effects of age and irrelevant information on the ability to inhibit one's actions. *Acta Psychologica*, *101*(2), 315–337.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). Association for Computing Machinery.

Rong, Y., Xu, W., Akata, Z., & Kasneci, E. (2021). *Human attention in fine-grained classification*. arXiv. https://arxiv.org/abs/2111.01628

Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human metacognition across domains: Insights from individual differences and neuroimaging. *Personality Neuroscience*, *1*, E17.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(11), 2660–2673.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*(3), 161–168.

Sokol, K., Hepburn, A., Santos-Rodriguez, R., & Flach, P. (2019). *bLIMEy: surrogate prediction explanations beyond LIME*. ArXiv. https://arxiv.org/abs/1910.13016

Spalek, T. M., & Hammad, S. (2005). The left-to-right bias in inhibition of return is due to the direction of reading. *Psychological Science*, *16*(1), 15–18.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164–182.

Stoet, G., O'Connor, D. B., Conner, M., & Laws, K. R. (2013). Are women better than men at multi-tasking? *BMC Psychology*, *1*(1), 18.

Strauss, S., & Ziv, M. (2012). Teaching is a natural cognitive ability for humans. *Mind, Brain, and Education*, *6*(4), 186–196.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Wang, Z., Wang, H., Wen, J.-R., & Xiao, Y. (2015). An inference approach to basic level of categorization. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 653–662). Association for Computing Machinery.

Xie, W., Li, X.-H., Cao, C. C., & Zhang, L. (2022). *ViT-CX*: *Causal explanation of vision transformers*. arXiv. https://arxiv.org/abs/2211.03064

Yang, Y., Zheng, Y., Deng, D., Zhang, J., Huang, Y., Yang, Y., Hsiao, J. H., & Cao, C. C. (2022). HSI: Human saliency imitator for benchmarking saliency-based model explanations. In J. Hsu & M. Yin (Eds.), *Proceedings of the tenth AAAI conference on human computation and crowdsourcing* (Vol. *10*, pp. 231–242). The AAAI Press.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, *24*(5), 1488–1500.

Zhang, J., Chan, A. B., Lau, E. Y. Y., & Hsiao, J. H. (2019). Individuals with insomnia misrecognize angry faces as fearful faces while missing the eyes: An eye-tracking study. *Sleep*, *42*(2), zsy220.

Zhao, C., Hsiao, J. H., & Chan, A. B. (2024). Gradient-based instance-specific visual explanations for object specification and object discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18. https://doi.org/10.1109/tpami.2024.3380604

Zheng, Y., & Hsiao, J. H. (2023). Differential audiovisual information processing in emotion recognition: An eye-tracking study. *Emotion*, *23*(4), 1028–1039.

Zheng, Y., Ye, X., & Hsiao, J. H. (2022). Does adding video and subtitles to an audio lesson facilitate its comprehension? *Learning and Instruction*, *77*, 101542.

How to cite this article: Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2024). Explanation strategies in humans versus current explainable artificial intelligence: Insights from image classification. *British Journal of Psychology*, *00*, 1–24. https://doi.org/10.1111/bjop.12714