Individual differences in explanation strategies for image classification and implications for explainable AI

Ruoxi Qi (ruoxiqi@connect.hku.hk)
Department of Psychology, University of Hong Kong
Pokfulam Road, Hong Kong

Yueyuan Zheng (mercuryzheng@connect.hku.hk)
Department of Psychology, University of Hong Kong
Pokfulam Road, Hong Kong

Yi Yang (yang.yi4@huawei.com)
Hong Kong Research Center, Huawei
Shatin, Hong Kong

Jindi Zhang (zhangjindi2@huawei.com)
Hong Kong Research Center, Huawei
Shatin, Hong Kong

Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology, the State Key Laboratory of Brain and Cognitive Sciences, and the Institute of Data Science,
University of Hong Kong
Pokfulam Road, Hong Kong

Abstract

While saliency-based explainable AI (XAI) methods have been well developed for image classification models, they fall short in comparison with human explanations. Here we examined human explanation strategies for image classification and their relationship with explanation quality to inform better XAI designs. We found that individuals differed in attention strategies during explanation: Participants adopting more explorative strategies used more visual information in their explanations, whereas those adopting more focused strategies included more conceptual information. In addition, visual explanations were rated higher for effectiveness in teaching learners without prior category knowledge, whereas conceptual explanations were more diagnostic for observers with prior knowledge to infer the class label. Thus, individuals differ in the use of visual and conceptual information to explain image classification, which facilitate different aspects of explanation quality and suit learners with different experiences. These findings have important implications for adaptive use of visual and conceptual information in XAI development.

Keywords: explanation; explainable AI; image classification; text analysis; eye movements; EMHMM

Introduction

Due to the advance of deep learning methods and availability of large datasets, artificial intelligence (AI) systems have been greatly improved in performance and increasingly applied to a wide range of fields. However, their inner workings have become more difficult to be understood by users or creators due to the black-box nature of deep learning (Lillicrap & Kording, 2019). Image classification is one such area that has seen both significant progress and growing needs for explainability. Various explainable AI (XAI) methods have been developed. In particular, saliency-based methods, which highlight input pixels that contribute to AI systems' decisions, have emerged as a popular approach to explain image classification models (e.g., Petsiuk et al., 2018; Selvaraju et al., 2017; Yang et al., 2022; Liu et al., 2023a, 2023b).

Nevertheless, it remains unclear whether saliency-based XAI methods can indeed promote user understanding. As pointed out by Kaufman and Kirsh (2022), AI models are sensitive to pixel-level changes that may not correspond to features comprehensible to humans. In addition, these methods highlight image regions without providing any information on what to look at and in what order, which is essential for explainees to derive meanings in human-to-human explanations. Therefore, current XAI explanations still fall short when compared with human explanations.

Our current understanding about how humans provide explanations for performing image classification remains very limited. Previous studies examining human explanations typically focused on complex phenomena such as scientific questions or questions related to causality or functions of different phenomena (Zemla et al., 2017; Lombrozo & Carey, 2006). There have been limited studies on explanations for tasks that involve automatic and unconscious perceptual processes, such as object recognition/image classification, since humans are typically able to perform the task with good performance without any explanation. However, explanations on these tasks are now required to facilitate human understanding of AI systems. Here we aimed to fill this gap through examining human explanation strategies for image classification and their relationships with explanation quality in order to inform better XAI designs.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

When classifying images, humans use both perceptual information and existing object category knowledge to inform their judgments (de Lange, et al., 2018; Kersten et al., 2004). Their existing knowledge involves both visual and abstract conceptual features of object categories (Martin et al., 2018). During category learning, humans first learn perceptual regularities within and between categories, which then give rise to complex conceptual knowledge (French et al., 2004; Samuelson & Smith, 1999). Sloutsky (2010) proposed two systems of category learning: a compression-based system that filters out idiosyncratic features of category members while retaining common features, and a selection-based system that directs attention to dimensions beneficial for error reduction. The compression-based system allows for unsupervised learning of dense categories that have many common features, while the selection-based system works better for sparse categories that share few relevant features, especially when they are defined by explicit rules. Therefore, people form perceptually rich representations through the compression-based system when learning novel dense categories (Kloos & Sloutsky, 2008). Familiar dense categories (e.g., common basic-level categories such as dogs and cats) can also be represented abstractly by category inclusion rules or lexical entries (Fisher & Sloutsky, 2005), thus enabling the use of unobservable properties to define categories (e.g., dogs and cats are alive) and in turn allowing further generalizations to organize concepts hierarchically (e.g., dogs and cats are animals, which are alive; Rakison & Poulin-Dubois, 2001). Thus, humans likely use both visual and abstract conceptual information to explain image classification, in contrast to current XAI explanations that mainly rely on visual information.

The use of visual or conceptual information in explanations may serve different purposes. For instance, natural categories are typically based on perceptual similarities, whereas artifacts are typically categorized according to rule-based connections between features (Stibel, 2006). Thus, visual information may be more effective for explaining natural categories while conceptual information is better for artificial categories. In addition, providing explanations requires metacognitive abilities to be aware of one's own thought processes (Jiang et al., 2016), which vary across individuals (Rouault et al., 2018). Together these findings suggest that in explaining image classification, individuals may differ in how well they can use visual or conceptual information, and these differences may be reflected in where they attend to during explanation. Different attention strategies and information use may also be associated with differences in explanation quality.

To investigate these possibilities, here we examined individual differences in explanation strategies for image classification. We asked participants to provide text explanations for why a presented image should be assigned a certain class label with eye tracking. We then examined whether participants differed in attention strategies during the task, and whether these differences were associated with variations in reliance on visual and conceptual information in the explanation text and the explanation quality. To quantify individual differences in attention strategies, we used a data-driven machine-learning-based approach, Eye Movement analysis with Hidden Markov Models (EMHMM; Chuk et al., 2014), which takes both spatial and temporal information of eye movements into account. To quantitatively assess explanation text characteristics, we used the visual dimension of perceptual strength from Lancaster Sensorimotor Norms (Lynott et al., 2020) to measure visual information and WordNet path similarity (Bird et al., 2009; Miller, 1995) to the label to measure conceptual information. We also included the action strength measure from Lancaster Sensorimotor Norms, which might be related to some functional information. Explanation quality was assessed on two different aspects: effectiveness was measured using subjective ratings of how effective the explanation could help learners without prior category knowledge understand the classification, whereas diagnosticity was measured as how well naïve observers could infer the class label given the explanation. We hypothesized that individual difference in attention strategies during explanation would be associated with both text characteristics and explanation quality. Specifically, individuals who used more visual information in explanation may adopt more explorative attention strategies for viewing the image to extract more visual features. Since visual information is more important for early learning and conceptual information is typically derived from later abstraction using explicit rules, visual explanations may be rated higher for effectiveness, whereas conceptual explanations could be more diagnostic. Thus, explanation text characteristics may mediate the relationship between attention strategy and explanation quality: Individuals using different attention strategies included different types of information in their explanations, which in turn affected the explanation quality.

Method

Participants

Sixty-two participants (52 females) with age ranging from 18 to 37 years (M = 22.5, SD = 3.8) were recruited from a local university. The participants included 7 native speakers of English. For the non-native speakers, they started to learn English at a mean age of 5.2 (SD = 2.4). On average, the participants scored 71.20% (SD = 12.79%) on the Lexical Test for Advanced Learners of English (LexTALE; Lemhöfer & Broersma, 2012). All participants had normal or corrected-to-normal vision.

Materials

The participants were shown 160 images obtained from ImageNet (Deng et al., 2009) and PASCAL VOC (Everingham et al., 2010). The images were in 20 classes, with 8 images in each class. Among the 20 image classes, 9 were natural classes (ant, corn, horse, jellyfish, lemon, lion, mushroom, snail, and zebra) and 11 were artificial classes (broom, cellphone, fountain, harp, laptop, microphone, pizza, shovel, sofa, tennis ball, and umbrella). The image classes were selected such that each class was a basic level category for humans (Markman & Wisniewski, 1997) and a common output class for image classification AI models (Russakovsky et al., 2015). Since the images had different sizes and aspect ratios, all images were resized to fit into a 400 × 520 pixel frame on a blank canvas.

Design

To examine individual differences in attention strategies, we first used EMHMM with co-clustering (Hsiao, Lan et al., 2021) to discover representative participant groups where group members shared similar eye movement patterns to one another (Pattern Groups A and B). ANOVA was used to examine the effect of this individual difference on text characteristics and explanation quality. The design consisted of a between-participant variable eye movement pattern group (Group A vs. B), and a within-participant variable image type (natural vs. artificial). The dependent measures were text characteristics including visual strength, WordNet similarity, and action strength, and explanation quality included effectiveness and diagnosticity. Correlation and mediation analyses were used to examine the relationships among eye movement pattern, text characteristics, and explanation quality.

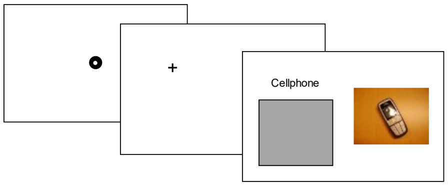


Figure 1: Trial procedure.

Procedures

In each trial, the participants were shown an image with a label and typed an explanation in a textbox about why the label should be assigned to the image (Figure 1). They were asked to imagine explaining to someone without any previous knowledge about perceptual categories, such as a very young child, and to provide sufficient information to help the person learn to assign correct labels to the images.

The participants' eye movements were recorded with an EyeLink Portable Duo eye tracker (SR Research) at a sampling rate of 1000 Hz, and E-Prime 3.0 with the extensions for EyeLink (Psychology Software Tools) was used to program the experiment. The stimuli were displayed on a 255 mm × 195 mm laptop with a resolution of 1024 × 768 pixels, and each image spanned 9.68° × 12.32° of visual angle at a viewing distance of 60 cm. A nine-point calibration and validation procedure was performed at the start of the experiment, and recalibration took place whenever drift correction error was over 1° of visual angle. Each trial began with a drift check at the center of the screen. A fixation cross was then displayed at the upper left corner of the screen, where the label would appear. The image, label, and textbox appeared on the screen once the participant fixated on the cross for more than 250 ms, so that the participant always saw the label first.

Data Analysis

Eye-Movement Patterns Participants' eye-movement data were analyzed using EMHMM (Chuk et al., 2014) with co-clustering (Hsiao, Lan, et al., 2021). Only fixations on the image area were included, and fixations that were more than three standard deviations from the mean fixation location for each image were removed as outliers.

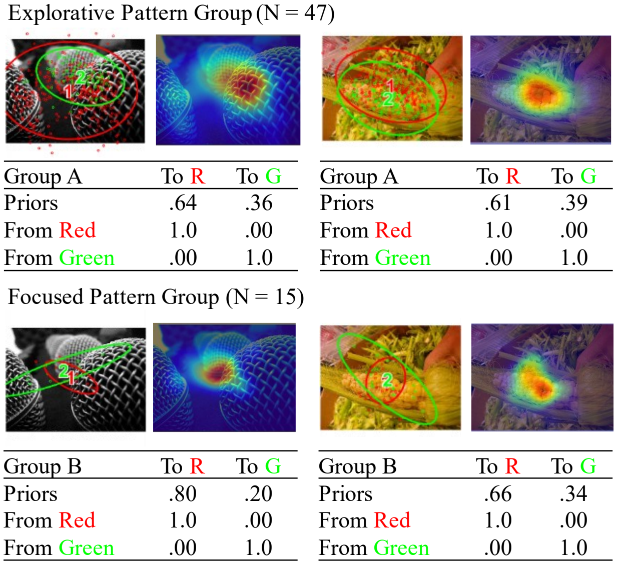
Each participant's eye movements when viewing each image were summarized using one hidden Markov model (HMM) with personalized regions of interest (ROIs) and transition probabilities among the ROIs. The optimal number of ROIs for each HMM was determined from a preset range of 1 to 10 using a variational Bayesian approach. Each HMM was trained for 200 times and the model with the greatest log-likelihood was selected. The participants were clustered into two groups, Pattern Group A and Pattern Group B, using the co-clustering algorithm, such that participants in the same group had similar eye-movement patterns to one another across the stimuli. A representative HMM was generated for each group and each stimulus, where the number of ROIs was set to be the median number of the individual HMMs. The co-clustering procedure was repeated for 200 times, and the result with the highest log-likelihood was selected.

Following previous studies (Chan et al., 2018; Hsiao, An, et al., 2021; Hsiao, Chan et al., 2021; Hsiao et al., 2022; Zheng et al., 2022), participants' attention strategies were quantified using A-B scale, which was calculated as $(L_A - L_B)/(|L_A| + |L_B|)$, where L_A and L_B represent the log-likelihoods of the participant's eye-movement patterns being classified as Pattern Group A and Pattern Group B respectively. A more positive A-B scale indicated greater similarity to Pattern Group A in contrast to Pattern Group B.

Explanation Text Characteristics Explanation texts were quantified using three measures: visual strength, WordNet similarity to the label, and action strength. Explanations were preprocessed by correcting typos, misused words, and major grammatical errors. The processed explanations were tokenized and lemmatized using spaCy (Honnibal et al., 2020) to obtain the three measures for each word, and the mean scores were calculated for each explanation.

The visual strength and action strength measures were retrieved from Lancaster Sensorimotor Norms (Lynott et al., 2020), which contain ratings for the extent to which words are experienced by using different perceptual senses and by performing actions with different parts of the body. Visual strength reflected the visual dimension of perceptual strength, while action strength was a composite measure of the dimensions for different parts of the body (e.g., "black white stripes" has high visual strength and "move slowly" has high action strength). WordNet similarity was computed using the NLTK interface (Bird et al., 2009) for WordNet (Miller, 1995). WordNet organizes words into sets of synonyms and connects the sets with semantic relations. Path similarity, based on the inverse of shortest path length in the hypernym/hyponym taxonomy, was used to measure similarity (e.g., “chair” has high similarity to “sofa”). Since WordNet only links words with the same part of speech, we calculated the similarity to the label for nouns. For words with multiple senses or meanings, we always chose the first one, which tended to be most commonly used.

Explanation Quality Assessment Explanation quality was evaluated by: (1) Effectiveness as measured by ratings from two computer vision experts on a scale from 1 to 7 while looking at the images. It indicated the effectiveness for teaching how to classify the image to someone without any prior knowledge of image classes. The two raters had good inter-rater reliability, Cronbach’s alpha = .858, and average rating was used for the analyses. (2) Diagnosticity as measured by how well naïve observers can infer the image class from the explanation text. It was obtained by presenting the explanations to 124 naïve observers (88 females) without the image or the label and asking them to infer the label. Their age ranged from 18 to 32 (M = 20.1, SD = 2.0) and their mean LexTALE score was 74.29% (SD = 14.29%). Each of them guessed the labels of a different set of 160 explanations with at most one explanation from each participant-class combination. Each explanation was evaluated by two observers. Diagnosticity for each participant was measured using percent accuracy of the observers across the explanations. Answers using the synonyms, hyponyms, or close hyponyms were counted as correct.



Explorative Pattern Group (N = 47)

Group A	To R	To G
Priors	.64	.36
From Red	1.0	.00
From Green	.00	1.0

Group A	To R	To G
Priors	.61	.39
From Red	1.0	.00
From Green	.00	1.0

Focused Pattern Group (N = 15)

Group B	To R	To G
Priors	.80	.20
From Red	1.0	.00
From Green	.00	1.0

Group B	To R	To G
Priors	.66	.34
From Red	1.0	.00
From Green	.00	1.0

Figure 2: Example representative HMMs of the explorative and focused patterns, where ellipses show ROIs as 2-D Gaussian emissions and dots represent the raw fixations, and heatmaps with a Gaussian distribution (SD = 0.5° of visual angle) applied to each fixation. Priors indicate the probabilities of a fixation sequence starting from each ROI, and the transition matrices indicate the transition probabilities among the ROIs.

Results

Eye-Movement Patterns during Explanation

EMHMM with co-clustering discovered two representative eye-movement Pattern Groups: Explorative (Group A) and Focused (Group B; Figure 2). Thus, the A-B scale was referred to as the Explorative-Focused (EF) scale. KL divergence estimation showed that the two groups differed significantly, F(1, 60) = 122.77, $p < .001$, $\eta^2_p = .67$. Participants in the Explorative Group had significantly more fixations per trial, t(57.78) = 6.57, $p < .001$, d = 1.53, longer average fixation duration, t(60) = 2.34, p = .022, d = 0.70, more fixations on the image region, t(60) = 4.02, $p < .001$, d = 1.19, and fewer fixations on the textbox region, t(60) = 3.38, p = .001, d = 1.00, than those in the Focused Group. Participants’ eye-movement patterns were more explorative for natural images than for artificial images, t(61) = 9.02, $p < .001$, d = 1.15.

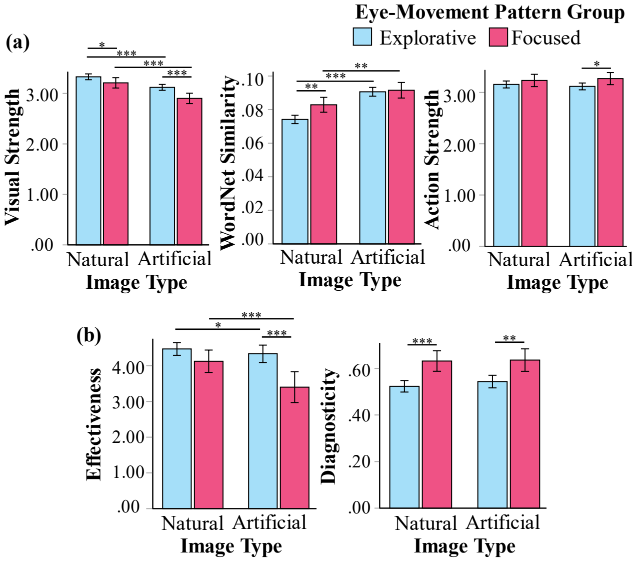


Figure 3: Difference in (a) explanation text characteristics and (b) explanation quality between the two pattern groups for natural and artificial images (error bars: 95% CI; $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$).

Pattern Group Differences in Explanation Text Characteristics and Quality

In explanation text characteristics, for visual strength, there were main effects of eye-movement pattern group, F(1, 60) = 9.33, p = .003, $\eta^2_p = .13$, where explorative participants used words with higher visual strength; and image type, F(1, 60) = 158.44, $p < .001$, $\eta^2_p = .73$, where participants used words with higher visual strength when explaining natural images. A significant interaction was also observed, F(1, 60) = 5.74, $\eta^2_p = .09$, with a greater difference between the two pattern groups for artificial images. For WordNet similarity to the label, we found main effects of eye-movement pattern group, F(1, 60) = 5.21, p = .026, $\eta^2_p = .08$, where focused participants used words more similar to the label; and image type, F(1, 60) = 69.20, p < .001, $\eta^2_p = .54$, where participants used words more similar to the label when explaining artificial images. There was also an interaction effect, F(1, 60) = 6.72, p = .012, $\eta^2_p = .10$, with the group difference only observed for natural images. For action strength, only a marginal effect of eye-movement pattern group was observed, F(1, 60) = 3.33, p = .073, where focused participants used words with higher action strength (Figure 3a). Consistent with these findings, EF scale was positively correlated with visual strength but negatively correlated with WordNet similarity and action strength (Table 1).

Table 1: Correlations between eye-movement patterns, explanation text characteristics, and explanation quality (Pearson's r with p-values in parentheses, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$).

	1	**2**	**3**	**4**	**5**
1. EF scale	–				
2. Visual strength	0.44 (< .001***)	–			
3. WordNet similarity	–0.29 (.023*)	–0.05 (.680)	–		
4. Action strength	–0.26 (.043*)	0.07 (.590)	0.19 (.130)	–	
5. Effectiveness	0.46 (< .001***)	0.50 (< .001***)	–0.42 (< .001***)	–0.54 (< .001***)	–
6. Diagnosticity	–0.42 (< .001***)	–0.36 (.004**)	0.31 (.013*)	0.29 (.021*)	–0.16 (.218)

In explanation quality, for effectiveness rating, we observed main effects of eye-movement pattern group, F(1, 60) = 9.28, p = .003, $\eta^2_p = .14$, with explorative participants being rated higher; and image type, F(1, 60) = 49.19, p < .001, $\eta^2_p = .45$, with explanations for natural images being rated higher. A significant interaction was also found, F(1, 60) = 23.06, p < .001, $\eta^2_p = .28$, where the difference between natural and artificial images was smaller for explorative participants. For diagnosticity, only a main effect of eye-movement pattern group was observed, F(1, 60) = 17.99, p < .001, $\eta^2_p = .23$, where explanations provided by focused participants had better diagnosticity (Figure 3b). Consistent with these findings, EF scale was positively correlated with effectiveness rating and negatively correlated with diagnosticity (Table 1).

The above results showed that a more explorative attention strategy during explanation was associated with explanation texts characterized by more visual information and less conceptual and action-related information, as well as higher effectiveness and lower diagnosticity. This result suggested that higher effectiveness in explanation quality may be associated with higher visual strength and lower WordNet similarity and action strength, whereas higher diagnosticity may be associated with the opposite characteristics. Correlation analyses confirmed these speculations (Table 1), suggesting that explanations relying more on visual information received better effectiveness ratings, whereas those with more conceptual and action-related information allowed naïve observers to infer the image class more easily. No significant correlation was found between the two quality measures, or among the three text characteristic measures.

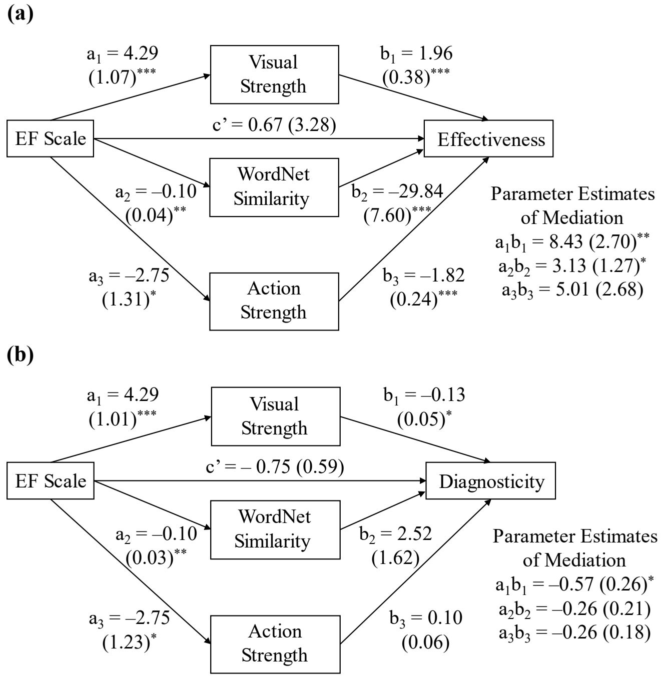

(a)
EF Scale
Visual Strength
WordNet Similarity
Action Strength
Effectiveness
$a_1 = 4.29$ $(1.07)^{***}$
$a_2 = –0.10$ $(0.04)^{**}$
$a_3 = –2.75$ $(1.31)^{*}$
$b_1 = 1.96$ $(0.38)^{***}$
$b_2 = –29.84$ $(7.60)^{***}$
$b_3 = –1.82$ $(0.24)^{***}$
c' = 0.67 (3.28)
Parameter Estimates of Mediation
$a_1b_1 = 8.43$ $(2.70)^{**}$
$a_2b_2 = 3.13$ $(1.27)^{*}$
$a_3b_3 = 5.01$ (2.68)
(b)
EF Scale
Visual Strength
WordNet Similarity
Action Strength
Diagnosticity
$a_1 = 4.29$ $(1.01)^{***}$
$a_2 = –0.10$ $(0.03)^{**}$
$a_3 = –2.75$ $(1.23)^{*}$
$b_1 = –0.13$ $(0.05)^{*}$
$b_2 = 2.52$ (1.62)
$b_3 = 0.10$ (0.06)
c' = – 0.75 (0.59)
Parameter Estimates of Mediation
$a_1b_1 = –0.57$ $(0.26)^{*}$
$a_2b_2 = –0.26$ (0.21)
$a_3b_3 = –0.26$ (0.18)

Figure 4: Indirect effect of EF scale on (a) effectiveness and (b) diagnosticity through the text characteristic measures. Path values represent the unstandardized coefficients with standard errors in parentheses (ab: indirect effect of EF scale on explanation quality through text characteristics, c': direct effect of EF scale on explanation quality, controlling for indirect effect; *p < .05, **p < .01, ***p < .001).

Did Attention Strategy Predict Explanation Quality through the Mediation of Its Association with Text Characteristics?

We examined whether participants' attention strategies predicted explanation text characteristics, which in turn led to differences in explanation quality. For explanation effectiveness, we found significant indirect effects of EF scale on effectiveness rating through visual strength, p = .002, and WordNet similarity, p = .014, and a marginal indirect effect through action strength, p = .061 (Figure 4a). These findings suggested that a more explorative attention strategy was associated with explanations that had higher visual strength, lower WordNet similarity, and lower action strength, which increased explanation effectiveness. The direct effect was not significant, p = .838. The total effect was significant, b = 17.24, SE = 3.28, p < .001. For diagnosticity, we observed a significant indirect effect of EF scale through visual strength, p = .030, but not through WordNet similarity, p = .208, or action strength, p = .146 (Figure 4b). These results indicated that a more explorative strategy predicted higher visual strength, which decreased diagnosticity. No significant direct effect was observed, p = .205, while the total effect was significant, b = −1.86, SE = 0.52, p < .001.

Discussion

Here we examined individual differences in attention strategies during explanation and whether such differences were associated with variations in reliance on visual or conceptual information in the explanation text and explanation quality. Through EMHMM, we discovered the explorative (exploring a wider image region) and focused (focusing on the foreground object) attention strategies. Explorative participants used more visual information, whereas focused participants included more conceptual information. In addition, visual explanations were rated as more effective for teaching image classification to people without prior knowledge, whereas conceptual explanations allowed people (with prior knowledge) to infer the label more easily. Finally, mediation analyses showed that participants using different attention strategies provided explanations with different characteristics, which were in turn associated with differences in explanation quality.

Participants adopting explorative and focused strategies had preferences for using more visual and conceptual information respectively. Since the explorative participants had more fixations on the image area, it is likely that they extracted more visual features from the images, while the focused participants relied more on their previous conceptual knowledge. In addition, explanations for natural categories were more visual and less conceptual than those for artificial categories. This result was consistent with previous findings that perceptual similarities are more important for categorizing natural objects, while artifact categorization relies more on rule-based connections between the features (Stibel, 2006). Interestingly, a smaller group difference in visual information use was observed for natural categories, and in conceptual information use for artificial categories, suggesting that participants had flexibility to include information more useful for explaining a specific image class regardless of their information use preference. Finally, action strength did not differ across natural and artificial categories. Previous research suggested that people use more functional information to explain artificial objects (Lombrozo & Carey, 2006). Although action strength may reflect some functional information, it also reflects movement or taste, which may apply to natural categories as well. Visual explanations had higher effectiveness, whereas conceptual explanations had better diagnosticity. Effectiveness was rated according to how effective it could teach someone with no prior knowledge to classify the image, whereas diagnosticity measured how easy observers with prior category knowledge could infer the label from the explanation text. Thus, our results were consistent with previous findings that in category learning, novel categories are represented based more on perceptual regularities, whereas familiar categories are represented more abstractly, often with non-perceptual features more diagnostic to the category (Fisher & Sloutsky, 2005; Kloos & Sloutsky, 2008). These findings further indicated that different explanation styles could suit different purposes. In addition, explorative strategies were associated with greater effectiveness in explanation quality while focused strategies were associated with higher diagnosticity. We found that these effects were mediated by the use of information in the explanation text. More specifically, the effect of attention strategy on effectiveness was mediated by the use of visual, conceptual, and action-related information, where participants using explorative strategies provided explanations with more visual information and less conceptual or action-related information, and thus were rated as more effective for teaching beginners. In contrast, the effect of attention strategy on diagnosticity was only mediated by the inclusion of visual information. This result suggested that it may be the prior conceptual knowledge about the image classes, instead of attention strategies when viewing the images, that affected diagnosticity through the use of conceptual and action-related information. The use of conceptual and action related information that are associated with high diagnosticity may not be well captured in eye movements during image viewing.

Thus, humans use both visual and conceptual information to explain image classification, in contrast to current saliency-based XAI methods (Qi et al., 2023). In addition, different information types facilitate different aspects of explanation quality, which may suit learners with different learning experience. Although image classification may seem to be a purely visual task, conceptual information plays an important role in human explanations, especially for artificial or familiar categories. Also, individuals have preferences over the use of visual or conceptual information for explanations. Together these findings suggest that when humans are learning from explanations, their performance is likely to be affected by how well the provided information matches their experiences and preferences. Therefore, it would be beneficial for AI and XAI researchers to consider using both visual and conceptual information adaptively to enhance classification and explanation performance. In addition, since humans also first learn categories through perceptual regularities and later develop abstract conceptual knowledge and rules, future work may examine whether existing deep learning models are able to learn abstract rules after being trained on large datasets of images using approaches from Artificial Cognition (e.g., Ritter et al., 2017; Taylor & Taylor, 2021; Yang et al., 2023).

In conclusion, we showed that individuals used different attention strategies when explaining image classification, which were associated with explanations that used different information and served different purposes. Specifically, participants using explorative strategies tended to provide visual explanations, which were considered more effective for beginning learners, while participants using focused strategies tended to give conceptual explanations, which were more diagnostic. These findings shed light on the inadequacies of current XAI methods for image classification and raised the possibility of adaptive use of visual and conceptual information in XAI.

Acknowledgments

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, and Caleb Cao and Luyu Qiu for their advice.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

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.

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.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE.

Everingham, M., Van Gool, L., Williams, C. K. I., 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.

French, R. M., Mareschal, D., Mermillod, M., & Quinn, P. C. (2004). The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: Simulations and data. *Journal of Experimental Psychology: General*, *133*(3), 382–397.

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

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), Article 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*, Article 104616.

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*, 1933–1943.

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*(6), 2473–2486.

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

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.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*(1), 271–304.

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.

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.

Lillicrap, T. P., & Kording, K. P. (2019). *What does it mean to understand a neural network*? arXiv. https://arxiv.org/abs/1907.06374

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

Liu, G., Zhang, J., Chan, A., & Hsiao, J. H. (2023b). *Human attention-guided explainable AI for computer vision models*. arXiv. https://arxiv.org/abs/2305.03601

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204.

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.

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*, 54–70.

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. *ELife*, *7*, Article e31873.

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

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 British Machine Vision Conference 2018* (p. 151). BMVA Press.

Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2023).

Explanation strategies for image classification in humans vs. current explainable AI. arXiv. https://arxiv.org/abs/2304.04448

Rakison, D. H., & Poulin-Dubois, D. (2001). Developmental origin of the animate-inanimate distinction. *Psychological Bulletin*, *127*(2), 209–228.

Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 2940–2949). JMLR.org.

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

Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, *73*(1), 1–33.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the 2017 IEEE International Conference on Computer Vision* (pp. 618–626). IEEE.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science*, *34*(7), 1244–1286.

Stibel, J. M. (2006). The role of explanation in categorization decisions. *International Journal of Psychology*, *41*(2), 132–144.

Taylor, J. E. T., & Taylor, G. W. (2021). Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, *28*(2), 454–475.

Yang, A., Liu, G., Chen, Y., Qi, R., Zhang, J., & Hsiao, J. H. (2023). Humans vs. AI in detecting vehicles and humans in driving scenarios. *Proceedings of the Annual Conference of the Cognitive Science Society*, *45*.

Yang, Y., Zheng, Y., Deng, D., Zhang, J., Huang, Y., Yang, Y., Hsiao, J., & 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* (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

Zheng, Y., Que, Y., Hu, X., & Hsiao, J. H. (2022). Predicting reading performance based on eye movement analysis with hidden Markov models. In *Proceedings of the 2022 International Conference on Advanced Learning Technologies (ICALT)* (pp. 172-176). IEEE.