Do Saliency-Based Explainable AI Methods Help Us Understand AI's Decisions? The Case of Object Detection AI

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

Guoyang Liu (gyliu@sdu.edu.cn)
School of Integrated Circuits, Shandong University
Jinan, Shandong, China

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

Janet H. Hsiao (jhhsiao@ust.hk)
Division of Social Science, Hong Kong University of Science & Technology
Clear Water Bay, Hong Kong

Abstract

Saliency-based Explainable AI (XAI) methods have been commonly used for explaining computer vision models, but whether they could indeed enhance user understanding at different levels remains unclear. We showed that for object detection AI, presenting users with AI 's output for a given input was sufficient for improving feature-level and some instance-level user understanding, particularly for false alarms, and providing saliency-based explanations did not have additional benefit. This was in contrast to previous research on image classification models where such explanations enhanced understanding. Analyses with human attention maps suggested that humans already attended to features important for AI's output in object detection and thus could infer AI's decision-making processes without saliency-based explanations. However, it did not enhance users' ability to distinguish AI's misses and hits, or system-level understanding. Therefore, the effectiveness of saliency-based explanations is task-dependent, and alternative XAI methods are required for object detection models to better enhance understanding.

Keywords: explainable AI; user understanding; object detection; saliency map; eye movements

Introduction

Recently the performance of artificial intelligence (AI) models has improved greatly due to the advance of deep learning methods and increased availability of large datasets. However, they have also become black boxes whose inner workings cannot be easily understood by their users or even creators (Lillicrap & Kording, 2019). Various explainable AI (XAI) methods have been developed to address this issue. For computer vision models, a common approach is to generate saliency maps that highlight pixels important to the model's decision (Petsiuk et al., 2018; Selvaraju et al., 2017). Quality of saliency maps could be evaluated by their faithfulness to the model, i.e., whether the saliency map accurately highlights input features that can lead to changes in AI's output (Samek et al., 2016). Nevertheless, computational metrics such as faithfulness alone cannot ensure that an XAI method can help users understand the AI model. Recent research has pointed out the importance of considering users' cognitive states during the explanation process and proposed cognitive metrics (e.g., Hsiao, Ngai, et al., 2021).

One potential cognitive metric is to assess user understanding of AI by simulatability, or how well the user can predict AI's behavior on new inputs (Hase & Bansal, 2020). Simulatability can be measured by two types of tasks: forward and counterfactual simulations (Doshi-Velez & Kim, 2017). Forward simulation involves predicting the output given an input, while counterfactual simulation involves predicting the change to the output given a change to the input. Previous studies have used forward simulation tasks and found that saliency maps could improve user understanding for image classification models. For instance, Alqaraawi et al. (2020) found that saliency maps helped participants better predict the model's output category on a new image. Similarly, Yang et al. (2022) found that participants typically expected AI to make similar classifications as themselves, and saliency map explanations helped them update this belief and predict AI's output category more accurately. Previous studies have also revealed that explanations could affect different levels of understanding differently. Specifically, saliency map explanations helped users attend to specific features when asked to infer what the model was sensitive to, but this feature-level understanding had a limited facilitation effect on their performance in predicting the model's general behavior on novel images as assessed in forward simulation (Alqaraawi et al., 2020). In addition, Kenny et al. (2021) found that instance-level understanding as the result of explanations did not necessarily aggregate into overall improvements at the system level. These findings raised the possibility that explanations may affect lower-level understanding more than higher-level understanding.

Accordingly, here we aimed to investigate the effect of saliency-based explanations on the different levels of understanding, which we termed as feature-, instance-, and system-level understanding. We focused on object detection models because of their important role in critical AI systems such as autonomous driving and medical imaging. Specifically, we examined object detection tasks in driving scenarios using a widely used object detector with high real-time performance, Yolo-v5s (Jocher, 2020). We used both forward and counterfactual simulation tasks to assess participants' understanding of the object detector at the three levels before and after a training phase, where they saw its output on different images with or without the saliency map explanations. Forward simulation allowed us to examine instance- and system-level understanding separately, where

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

instance-level understanding could be measured by how well participants predicted AI model's hits, misses, and false alarms, while system-level understanding could be quantified by summing over participants' predicted performance of AI across instances. Counterfactual simulation assessed feature-level understanding by asking participants to judge whether covering certain features could lead to changes in AI's decisions. Previous studies found that explanations generally had greater impact on the understanding of AI's mistakes than correct instances (Kenny et al., 2021; Yang et al., 2022). In addition, current saliency-based explanations typically do not highlight features for objects that were not detected by AI, including misses. Therefore, saliency maps may help users better predict false alarms than hits and misses.

Previous research has reported an association between learners' understanding or knowledge level of visual stimuli and their eye movement behavior (Gegenfurtner et al., 2011; Kruger and Steyn, 2014). For instance, Zheng et al. (2022) found that during multimedia learning, learners who used a more centralized eye movement pattern (i.e., focusing on the screen center) had better comprehension of the lessons. These findings suggested that we may use eye movement behavior as an objective measure to monitor users' current understanding without interrupting their cognitive processes involved in the task by using subjective measures (Hsiao, Ngai, et al., 2021). Accordingly, here we performed eye tracking and used participants' eye movements during simulation as an alternative measure of understanding. To better capture the substantial individual differences in both the spatial (where participants look) and temporal dimension of eye movements (the order of where they look) observed in visual tasks (Peterson & Eckstein, 2013), we analyzed the eye movement data with a data-driven approach, Eye Movement analysis with Hidden Markov Models (EMHMM; Chuk et al., 2014) with co-clustering (Hsiao, Lan, et al., 2021). This approach allowed us to discover representative attention strategies among the participants and use them to quantify individual eye movement patterns. It also allowed us to quantify eye movement consistency using the entropy of the hidden Markov models (HMMs), with higher entropy indicating less consistent or more adaptive attention strategies. We hypothesized that saliency-based explanations would help enhance user understanding of an object detection model's decision beyond presenting the model's output alone, similar to what was found with image classification models. However, the effectiveness of such explanations might decrease at a higher level of understanding. At the instance level, saliency maps may be more helpful for predicting false alarms than hits and misses. Finally, participants' attention strategies during simulation may become more consistent with features used by AI as a result of training, especially after training with saliency maps.

Methods

Participants

We recruited 68 participants (43 females) aged between 18 to 30 years (M = 21.7, SD = 2.9) from a local university. The participants all had normal or corrected-to-normal vision. Power analysis indicated that assuming medium effect sizes, 34 participants were required to detect a within-between interaction for mixed ANOVA ($\alpha = .05$, $\beta = .08$, $f = .25$) and that 55 participants were required to conduct linear regression analyses with one predictor ($\alpha = .05$, $\beta = .08$, $f^2 = .15$).

Materials

Three sets of images, each with 32 images, were randomly selected from the CODA dataset, a novel dataset of object-level corner cases in real-world driving scenes (Li et al., 2022). Each image set was selected such that the number of hits was roughly equal to the number of misses and false alarms and the number of misses was roughly equal to the number of false alarms (the difference was less than 10% of the total number of objects).

Two image sets were used for the pre- and post-training simulation tasks, with the order counterbalanced across participants, and one set was used for training. For counterfactual simulation, we generated 20 perturbed images for each correctly detected target by adding blocks with randomly generated RGB values that cover different parts of the bounding box and randomly selected one perturbed image for each image. Specifically, we generated 20 distinct perturbation blocks for each bounding box, with central points evenly distributed across the bounding box in a 4 × 5 grid. The dimensions of each perturbation block were one-fourth of the bounding box's width and one-fifth of its height. Saliency maps were generated using FullGrad-CAM++ (Liu et al., 2023a), an effective gradient-based XAI method designed for object detection models. Compared with traditional gradient-based XAI methods such as Grad-CAM (Selvaraju et al., 2017), it can generate instance-specific saliency maps with higher faithfulness and plausibility. All images were resized to 1024 × 576 pixels and displayed on a 375mm × 300 mm monitor with a resolution of 1024 × 768 pixels. Each image spanned 29.99° × 18.26° of visual angle at a viewing distance of 70 cm.

Design

The design consisted of a within-participant variable, time point (pre- vs. post-training), and a between-participant variable, training condition (with vs. without saliency maps). The dependent variables were participants' performance and eye movement measures in the simulation tasks (see Data Analysis section). A 2 × 2 mixed ANOVA was used. We then examined whether training effects in eye movements could predict performance change through regression.

In a separate analysis, we examined how well participants' attention maps during the stimulation tasks reflected features important to AI's output by calculating the faithfulness measure of the attention maps given the AI model (see Data Analysis section) as an alternative measure of user understanding. To do this, for each image we generated four human attention maps by aggregating participants' eye movements for each time point and training condition

combination, and conducted a by-image 2 (pre-training vs. post-training) by 2 (with vs. without saliency maps) repeated-measures ANOVA on the faithfulness of these attention maps.

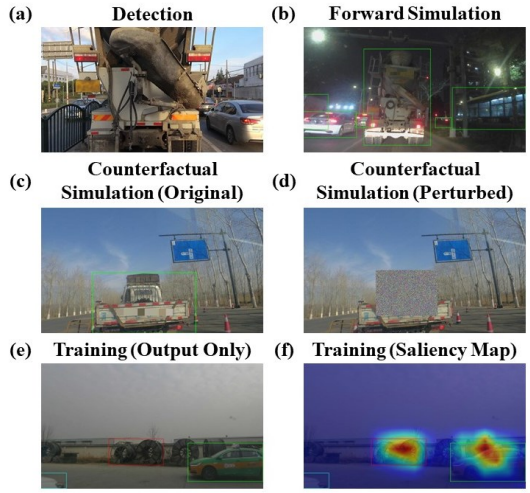


Figure 1: Example image stimuli.

Procedures

Participants first completed a vehicle detection task with stimuli from all image sets to measure baseline performance and eye movement behavior. They then completed the two pre-training simulation tasks, starting from forward simulation. Next, they went through the training phase with or without seeing saliency maps, depending on the condition to which they were randomly assigned. After training, they proceeded to the two post-training simulation tasks on a different set of images. Participants' eye movements were tracked with an EyeLink Portable Duo eye tracker (SR Research). A nine-point calibration and validation procedure was performed at the beginning of the experiment, and re-calibration took place whenever drift check error exceeded 1° of visual angle. Each trial started with a drift check at the center of the screen, and the stimulus appeared after a stable fixation at the center was observed.

Detection Participants were presented with images of driving scenes one at a time (Figure 1a) and were asked to detect all vehicle targets, including cars, buses, and trucks, by placing a marker on each target with a mouse click.

Forward Simulation Participants saw images with the ground truth bounding boxes marked (Figure 1b). They were asked to click on all the targets they thought AI model (Yolo-v5s) could detect (hits) and all the non-targets that AI model would falsely detect (false alarms), while not clicking on the targets that they thought AI model would not detect (misses).

Counterfactual Simulation Participants were presented with images where one correctly detected target was marked (Figure 1c). They then saw a perturbed image where part of the marked target was covered by noise (Figure 1d) and were asked to judge whether the output of AI would change. They could freely switch between the two images.

Training Participants saw images where AI model's hits, misses, and false alarms were marked with color coded bounding boxes in green, blue, and red respectively (Figure 1e). For participants in the "with saliency map" condition, they also saw the saliency maps (Figure 1f) and could freely switch between the two views.

Data Analysis

Behavioral Measures Detection performance was measured by recall (number of correctly detected targets divided by total number of targets), and precision (number of correctly detected targets divided by total number of objects clicked).

For forward simulation, we evaluated instance-level understanding using three measures: (1) Discrimination sensitivity of AI's hits and misses as measured in D', i.e., how well participants could judge whether a ground truth bounding box was a hit or a miss. It was calculated as z(Number of Correctly Predicted Hits/Total Number of Hits) – z(Number of Incorrectly Clicked Miss/Total Number of Misses), where z indicates z-transformation. (2) Number of AI's false alarms guessed. (3) Accuracy of predicting AI's false alarms, i.e., the number of correctly predicted false alarms divided by the total number of false alarms. In addition, we examined system-level understanding first by the similarity between participants' performance in the baseline object detection task (recall and precision) and their predicted performance of the AI model across images using cosine similarity[1] as a measure of the similarity between their own mental model of the task and their mental model of how AI performed the task. A decrease in this measure could indicate that participants updated their mental models about AI, instead of merely assuming that AI would behave like themselves. Another measure of system-level understanding was the similarity between participants' mental models of how AI performed the task and AI's actual performance by calculating the cosine similarity between their predicted recall/precision of AI and AI's actual recall/precision across images. An increase in this measure could indicate that participants' mental models of AI had become more accurate.

For counterfactual simulation, we measured performance by accuracy to assess feature-level understanding.

Eye Movement Measures EMHMM (Chuk et al., 2014) with co-clustering (Hsiao, Lan, et al., 2021) was used to analyze participants' eye movement patterns during the baseline detection task. Each participant's eye movements on each image were summarized with one HMM, which contained person-specific regions of interest (ROIs) and

[1] Cosine similarity is calculated as $(A \cdot B)/(\|A\| \times \|B\|)$, where A and B are vectors containing measures of recall or precision across images. $A \cdot B$ is the dot product and $\|A\|$ and $\|B\|$ are the Euclidean norms of the two vectors.

transition probabilities among these ROIs. The optimal number of ROIs for each individual HMM was determined using a variational Bayesian approach from a preset range of 1 to 10. Each individual HMM was trained 200 times, and the HMM with the highest log-likelihood was chosen. Participants were then clustered into two pattern groups using the co-clustering method, which put participants who shared similar eye movement patterns across the stimuli into the same group. A representative HMM was generated for each pattern group and each image, with the number of ROIs set to be the median number of the group members' individual HMMs. The co-clustering procedure was repeated for 200 times; the result with the highest log-likelihood was selected. Following previous studies (e.g., Hsiao, Chan, et al., 2021; Qi et al., 2023), each participant's eye movement patterns could be quantified using the A-B scale, which was defined as $(L_A - L_B)/(|L_A| + |L_B|)$. L_A and L_B were the log-likelihoods of a participant's eye movement data being assigned to Pattern Group A and B, respectively, and a more positive A-B scale indicated higher similarity to Pattern Group A. We also calculated each participant's eye movement entropy to measure eye movement consistency by summing the entropies of their individual HMMs across the stimuli.

For eye movement data during simulation and training, the same methods were used to generate the individual HMMs and to calculate the entropy. To make the eye movement pattern measure (A-B scale) comparable across time points, we used the two representative pattern groups discovered from the baseline detection task to quantify the A-B scale in both the stimulation tasks and training. Counterfactual simulation involved looking at only one target, so the eye movement patterns would not be comparable to those during detection. Thus, we focused our analysis on entropy.

Faithfulness We generated human attention maps by applying a Gaussian kernel with a standard deviation of 30 pixels (approximately 1° of visual angle) to smooth the eye fixations. We then computed faithfulness of the human attention maps and XAI saliency map for the AI model on each image using the deletion approach, which deleted salient areas step-by-step according to the saliency scores and filled the deleted regions with random noise. 100 steps were conducted with 1% of the total area deleted in each step and the confidence changes were recorded (Chattopadhay et al., 2018, Liu et al., 2023a). The area under the deletion curve was used as the faithfulness measure.

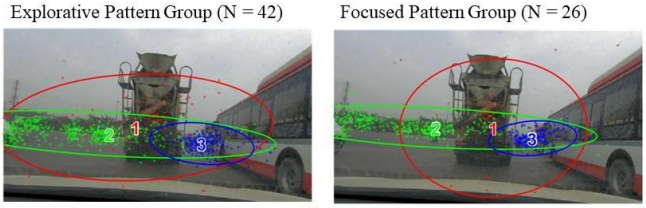
Results

Effect of Training on the Understanding of AI

Forward Simulation For instance-level understanding, after training participants guessed more AI's false alarms, $F(1, 66) = 29.35$, $p < .001$, $\eta^2_p = .308$, and had higher accuracy for predicting its false alarms, $F(1, 66) = 14.59$, $p < .001$, $\eta^2_p = .181$. However, they did not become better at distinguishing AI's hits and misses, $F(1, 66) = 1.38$, $p = .244$, $\eta^2_p = .020$. Meanwhile, none of training effects interacted with training condition (ps = .288, .845, .533, respectively), indicating that training had similar effects on instance-level understanding regardless of whether saliency maps were included.

For system-level understanding, participants' mental model of AI's performance became less similar to their own detection performance after training, both in terms of recall, $F(1, 66) = 6.87$, $p = .011$, $\eta^2_p = .094$, and precision, $F(1, 66) = 19.47$, $p < .001$, $\eta^2_p = .228$. These results indicated that participants no longer assumed that AI performed similarly to themselves after they updated their mental models of AI's performance as a result of training. However, their prediction of AI's performance did not become more similar to AI's actual performance, both for recall, $F(1, 66) = 0.61$, $p = .439$, $\eta^2_p = .009$, and for precision, $F(1, 66) = 0.86$, $p = .375$, $\eta^2_p = .013$, suggesting that their mental models of AI did not become more accurate. Similarly, none of the effects interacted with training condition (ps = .300, .851, .798, .158, respectively), suggesting that saliency maps did not enhance system-level understanding either.

Counterfactual Simulation After training, participants had improved counterfactual simulation accuracy, $F(1, 66) = 7.17$, $p = .009$, $\eta^2_p = .098$, indicating that participants had better understanding about which features could affect AI's output. However, similar to forward simulation, this effect did not interact with training condition, $F(1, 66) = 0.14$, $p = .713$, $\eta^2_p = .002$, indicating that saliency maps also could not enhance feature-level understanding.



Group A	To R	To G	To B
Priors	.58	.41	.02
From Red	.70	.24	.06
From Green	.01	.94	.05
From Blue	.01	.16	.83

Group B	To R	To G	To B
Priors	.56	.42	.02
From Red	.31	.48	.21
From Green	.03	.91	.06
From Blue	.05	.14	.81

Figure 2: Example representative HMMs for the explorative and focused pattern groups during detection. ROIs as 2-D Gaussian emissions are represented by the ellipses and raw fixations are represented by the dots. Priors reflect the probability for the first fixation to be on a certain ROI, while the transition matrices show the transition probabilities among the ROIs.

Eye Movement Pattern and Entropy During the Detection, Training, and Simulation

Detection Using EMHMM with co-clustering, we discovered the explorative and focused eye movement pattern groups, consistent with a previous study (Yang et al., 2023; Figure 2). Explorative participants scanned a wider region around the center of the image, while focused participants mainly scanned along the horizon, where

vehicles were most likely to appear. Therefore, the A-B scale was referred to as the Explorative-Focused (EF) scale. KL divergence estimation showed that the two groups differed significantly, $F(1, 66) = 170.01$, $p < .001$, $\eta^2_p = .720$.

Training Participants' eye movement patterns, as quantified by the EF scale, did not differ across the two training conditions, $t(66) = 0.10$, $p = .921$, $d = 0.02$ (Figure 3a). This finding indicated that participants attended to similar places during training regardless of whether they saw saliency maps or not. However, participants who saw saliency maps had marginally lower entropy, $t(66) = –1.97$, $p = .053$, $d = –0.48$ (Figure 3b), indicating that their eye movements during training were more consistent and predictable.

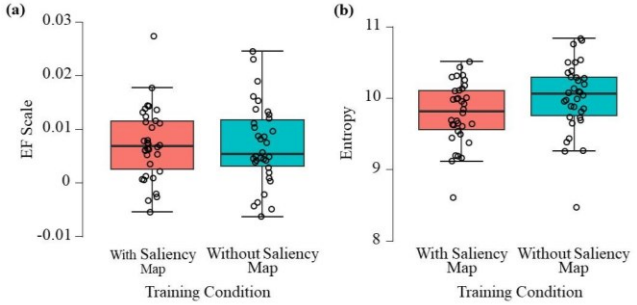


Figure 3: Difference in (a) EF scale and (b) entropy of participants' eye movements during training between the two training conditions.

Forward Simulation Participants' eye movement patterns, as quantified by the EF scale, did not change after training, $F(1, 66) = 0.01$, $p = .917$, $\eta^2_p < .001$, and there was no interaction effect between time and training condition, $F(1, 66) = 1.04$, $p = .312$, $\eta^2_p = .016$, indicating that participants attended to similar places when predicting AI's output before and after training for both training conditions. In contrast, participants' eye movement entropy significantly increased after training, $F(1, 66) = 9.70$, $p = .003$, $\eta^2_p = .128$. However, this effect again did not interact with training condition, $F(1, 66) = 1.20$, $p = .278$, $\eta^2_p = .018$, indicating that their eye movement strategies became less consistent, or more adaptive, after training, regardless of the training condition.

In addition, this change in entropy was associated with change in the performance measures related to false alarms. Specifically, participants who showed a greater increase in eye movement entropy after training had larger improvement in accuracy in predicting AI's false alarms, $R^2 = .07$, $F(1, 66) = 4.62$, $\beta = .26$, $p = .035$, and predicted a marginally greater number of false alarms, $R^2 = .05$, $F(1, 66) = 3.34$, $\beta = .22$, $p = .072$. They also updated their mental models about Yolo-v5s's precision more after training (i.e., becoming less similar to their own detection precision), $R^2 = .08$, $F(1, 66) = 6.05$, $\beta = –.29$, $p = .017$, but their beliefs also tended to deviate even more from Yolo-v5s's actual precision, $R^2 = .09$, $F(1, 66) = 6.83$, $\beta = –.31$, $p = .011$.

Counterfactual Simulation Participants' eye movement entropy during counterfactual simulation did not change after training, $F(1, 66) = 0.12$, $p = .735$, $\eta^2_p = .002$, and there was no interaction between time and training condition, $F(1, 66) = 1.83$, $p = .180$, $\eta^2_p = .027$. However, at individual level, participants who showed a greater increase in entropy had marginally greater improvement in accuracy, $R^2 = .04$, $F(1, 66) = 2.87$, $\beta = .20$, $p = .095$.

Faithfulness of Human Attention Maps

The faithfulness of participants' attention maps for the AI model did not differ between pre- and post-training in forward simulation, $F(1, 59) = 1.12$, $p = .294$, $\eta^2_p = .019$, or counterfactual simulation, $F(1, 59) = 0.66$, $p = .420$, $\eta^2_p = .011$. There was also no interaction between time point and training condition in either task (ps = .113, .238). These results indicated that when predicting AI's behavior, participants did not attend more to the features important to AI after training, either with or without saliency maps.

Discussion

Here we investigated the effect of providing saliency-based explanations on different aspects of user understanding of object detection AI models. We found that presenting users with AI's decisions given an input image, regardless of whether saliency maps were provided or not, enhanced feature-level and some instance-level user understanding, including better performance in predicting AI's decision change when certain features were perturbed (counterfactual simulation) and in predicting AI's false alarms in forward simulation. However, these learning effects were obtained regardless of the availability of saliency-based explanations, suggesting that such explanations did not provide additional benefits. This result was in contrast to previous research on image classification models, where enhancement at the feature and instance level due to saliency map explanations has been reported (Alqaraawi et al., 2021). Our analyses on participants' attention strategies during forward simulation showed that their eye movement pattern did not change after training. In contrast, training (in either condition) increased eye movement entropy, suggesting more adaptive attention strategies. The increase in eye movement entropy was associated with better performance in predicting AI's false alarms and a larger change in the mental model of AI's precision. These findings suggested that although training helped users engage in a more adaptive attention strategy, saliency maps provided limited benefits beyond presenting AI's decisions alone.

This limited facilitation effect from saliency-map explanations on user understanding of object detection models may be because participants already attended to the features used by AI and thus could not gain any additional information from the features highlighted by the saliency maps. This speculation was consistent with our finding that participants looked at similar places during training regardless of whether saliency maps were provided. Also, the faithfulness of participants' attention maps during simulation in explaining the AI model did not differ across different time points or training conditions. Thus, training (with or without

saliency maps) did not make them attend more or less to features important to AI's decisions. Indeed, Liu et al. (2023b) has recently suggested that current object detection AI models may attend to similar features as humans, but this phenomenon may not apply to image classification models. More specifically, they found that for explaining object detection AI models, increasing the similarity of XAI saliency maps to human attention maps when performing the same task increased their faithfulness, but the same procedure decreased the faithfulness of XAI saliency maps for image classification models. To examine whether our participants indeed attended to features important to AI's output, we performed exploratory analyses and found that the faithfulness of participants' overall attention maps did not differ significantly from the faithfulness of XAI saliency maps in either forward, $t(59) = 0.05$, $p = .960$, $d = 0.01$, or counterfactual simulation, $t(55) = 0.07$, $p = .942$, $d = 0.01$. In addition, this faithfulness was highly correlated with that of XAI saliency maps across images in both forward, $r(58) = .787$, $p < .001$ (Figure 4a), and counterfactual simulation, $r(54) = .531$, $p < .001$ (Figure 4b). These results indicated that participants' attention during simulation indeed matched well with features important for AI's output. Since humans already attend to similar features as AI models in object detection, human users could use their own mental model of the task to infer AI models' decision-making processes. As a result, giving an input image, presenting AI's decisions alone could already enhance user understanding and saliency-based explanations did not provide additional benefit, in contrast to the case of image classification.

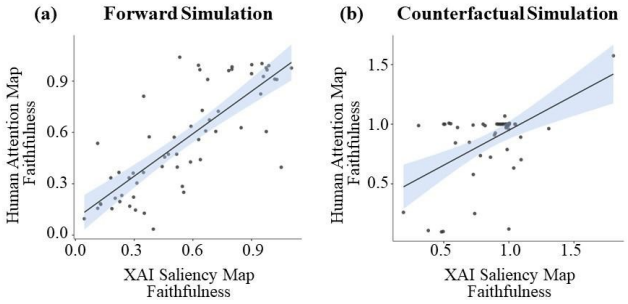


Figure 4: Correlations between the faithfulness of human attention maps and XAI saliency maps across images for (a) forward simulation and (b) counterfactual simulation.

Although presenting AI's decisions alone was sufficient to enhance user understanding for object detection, our current results suggested that the training effect was limited to some feature- and instance-level understanding. Specifically, training improved participants' knowledge about which features would affect AI's output. At the instance level, training helped participants better predict false alarms, but not hits and misses. At the system level, participants first assumed that AI would behave similarly as themselves and then updated their mental models of AI's performance with increased dissimilarity to themselves, consistent with previous findings (Yang et al., 2022). However, their mental models of AI's performance did not become more consistent with AI's actual performance. These results suggested an illusive sense of system-level understanding. A similar finding was observed in a previous study, which found that participants tended to overestimate the system-level understanding gained from instance-level explanations and explained this phenomenon with a cognitive bias known as the illusion of explanatory depth (Chromik et al., 2021).

Since presenting users with object detection AI's decisions given an input had limited effects on user understanding and saliency-based explanations could not provide additional benefit, better XAI methods are required to enhance user understanding for object detection models. It is particularly important for system-level understanding, which can be prone to the cognitive bias of illusion of explanatory depth, and the understanding of misses, which can lead to severe consequences in critical systems (e.g., missing a vehicle in autonomous driving or missing a tumor in medical diagnosis). One possible future direction is to consider human explanation processes where the explainer possesses the ability to infer the learner's mental states (i.e., the theory of mind ability) so that they can recognize and reduce the knowledge gap between the explainer and the learner (Olson & Bruner, 1996; Strauss & Shilony, 1994). Accordingly, we may equip XAI with such ability to ensure that the explanations are accessible to human users. Indeed, it has been suggested that XAI methods should monitor users' mental model of AI and adjust the explanations accordingly (Rutjes et al., 2019). Hsiao and Chan (2023) further proposed a theory-of-mind-based XAI framework, which posits that an effective XAI method should be able to infer users' strategy and performance of the given task as well as users' current understanding of AI's strategy and trust towards AI, so that it can provide user-centered explanations through comparing users' and AI's mental models of the task and estimating current user understanding, similar to real-life teachers.

In conclusion, here we showed that presenting AI's decisions for a given input alone is sufficient to enhance users' understanding of AI's decision processes for object detection, and providing saliency-based explanations to highlight important features used by AI did not provide additional benefit, in contrast to the case of image classification AI models where such explanations were reported to enhance user understanding. This result may be because humans and AI attended to similar features when performing object detection, and thus humans could infer AI's decision-making processes without saliency-based explanations. Nevertheless, the observed effect in user understanding was limited to feature-level and some instance-level understanding, particularly in predicting AI's false alarms, but did not enhance users' ability to distinguish AI's misses and hits, or the understanding at the system level. Our findings thus demonstrated that the effectiveness of saliency-based explanations is task-dependent and call for alternative XAI methods for object detection models to better enhance user understanding and induce an appropriate level of user trust, especially for those in critical systems such as autonomous driving and medical image analysis.

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 Yunke Chen for helping with data collection. We are also grateful for the helpful comments and suggestions from Yi Yang, Yueyuan Zheng, Alice Yang, and Caleb Cao.

References

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. *Proceedings of the International Conference on Intelligent User Interfaces*, *25*, 275–285.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE winter conference on applications of computer vision (WACV)*.

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think I get your point, AI! The illusion of explanatory depth in explainable AI. *26th International Conference on Intelligent User Interface*s, 307–317.

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. https://arxiv.org/abs/1702.08608

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: a meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, *23*(4), 523–552.

Hase, P., & Bansal, M. (2020). Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *58*, 5540–5552.

Hsiao, J. H., & Chan, A. B. (2023). Towards the next generation explainable AI that promotes AI-human mutual understanding. *XAI in Action: Past, Present, and Future Applications*.

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., 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.

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

Jocher, G. (2020). *YOLOv5 by Ultralytics* (7.0) [Python]. https://github.com/ultralytics/yolov5

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, *294*, Article 103459.

Kruger, J.-L., & Steyn, F. (2014). Subtitles and eye tracking: Reading and performance. *Reading Research Quarterly*, *49*(1), 105–120.

Li, K., Chen, K., Wang, H., Hong, L., Ye, C., Han, J., Chen, Y., Zhang, W., Xu, C., Yeung, D.-Y., Liang, X., Li, Z., & Xu, H. (2022). CODA: *A real-world road corner case dataset for object detection in autonomous driving*. arXiv. https://arxiv.org/abs/2203.07724

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. B., & Hsiao, J. (2023a). 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. (2023b). *Human attention-guided explainable artificial intelligence for computer vision model*s. arXiv. https://arxiv.org/abs/2305.03601

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. http://arxiv.org/abs/2304.04448

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Rutjes, H., Willemsen, M., & IJsselsteijn, W. (2019). Considerations on explainable AI and users' mental models. In *Where is the Human? Bridging the Gap Between AI and HCI: Workshop at CHI' 19*. Association for Computing Machinery, Inc.

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. (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.

Strauss, S., & Shilony, T. (1994). Teachers' models of children's minds and learning. In *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 455–473). Cambridge University Press.

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, S. C.-H., Folke, N. E. T., & Shafto, P. (2022). A psychological theory of explainability. *Proceedings of the International Conference on Machine Learning*, *39*, 25007–25021.

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