Humans vs. AI in Detecting Vehicles and Humans in Driving Scenarios

Alice Yumeng Yang
u3555074@connect.hku.hk
Department of Psychology, University of Hong Kong

Guoyang Liu
gyangliu@hku.hk
Department of Psychology, University of Hong Kong

Yunke Chen
cyk1028@connect.hku.hk
Department of Psychology, University of Hong Kong

Ruoxi Qi
ruoxiqi@connect.hku.hk
Department of Psychology, University of Hong Kong

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

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

Abstract

To inform Explainable AI (XAI) design for updating users' beliefs about AI based on their mental models, we examined the similarities and differences between humans and AI in object detection in driving scenarios. In humans, individuals differed in adopting focused or explorative attention strategies, with better performance associated with the focused strategy. AI (Yolo-v5s) had higher similarity in attended features to the focused than the explorative strategy in humans, and achieved human-expert-level performance in vehicle detection even in difficult cases such as occlusion and degradation. In contrast, it performed much poorer than humans in detecting humans with low attended feature similarity due to humans' attention bias for stimuli with evolutionary significance. Also, higher similarity to humans' attended features was associated with better AI performance, suggesting that human attention may be used for guiding AI design. These findings have significant implications for both AI and XAI designs.

Keywords: eye movements; object detection; EMHMM; artificial intelligence; explainable AI

Introduction

Deep learning methods and availability of large datasets have revolutionized Artificial Intelligence (AI) research and at the same time demand more research on human-AI interaction (Lindebaum et al., 2020). Although many Explainable AI (XAI) methods have been proposed to help humans visualize AI's operations and decision-making processes, the understanding of how humans perceive and understand AI through these visualizations remains very limited (Hsiao, Ngai, et al., 2021; Mueller et al., 2019; Páez, 2019). In particular, in understanding others' behavior, humans often attribute mental states to others and to selves, an important ability in social interaction referred to as theory of mind (Frith & Frith, 2005). Thus, highlighting similarities and differences between AI and human decision-making processes may be essential for explanations that can truly enhance human understanding of AI. With the rapid advances, several current AI systems have been claimed to match or even outperform humans. For instance, PReLU-nets became the first model to surpass human-level performance (5.1% top-5 error) on the ImageNet dataset with a 4.94% top-5 error (He et al., 2015). In medicine, fine-grained methods could match expert physicians in performance in detecting skin cancer and diabetic retinopathy (Esteva et al., 2017; Gulshan et al., 2019). Nevertheless, it remains unclear whether these AI models employ similar strategies to humans in order to achieve or surpass human-level performance.

Early attempts to compare human and deep neural networks (DNNs) in image classification have been made. For example, Lake et al. (2015) found that human category typicality could be predicted by a DNN trained for classification on raw naturalistic images. Kheradpisheh et al. (2016) found that DNNs could use representations consistent with humans. These findings suggest that AI models may share similar strategies and internal representations with humans. However, some differences have been noted. For example, Geirhos et al. (2017) argued that in visual object recognition, the human visual system is more robust to various image conditions, including contrast reduction, noises, or distortions. Indeed, a fundamental difference between AI and humans is in their attention mechanisms: humans process bits of information at a time through a sequence of eye fixations, whereas AI processes all information simultaneously (Qi et al., 2023a). Nonetheless, humans are able to recognize the global scene gist across the entire image at a glance, and this information is sufficient to guide visual search and plan subsequent attention shifts (Navon, 1977; Oliva & Torralba, 2006; Wolfe et al., 2022). Top-down factors, including experience, task demands, and age, influence how people plan eye fixations to acquire information from an image (Betz et al., 2010; Hsiao & Chan, 2023; Hsiao, An, et al., 2021). In contrast, AI systems typically focus on bottom-up information and lack top-down attention to guide object detection and recognition (Oliva et al., 2003).

Although AI and humans may differ in the way they extract visual information, it remains possible that similar features, which are diagnostic to the task, are used by human experts and best-performing AI models, resulting in similar levels of performance. Here we aimed to examine this possibility by

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

comparing AI and human performance and attended features in object detection in driving scenarios due to its importance in self-driving applications where safety is a critical concern. In particular, we examined whether AI and humans differed in difficult object detection conditions including occlusion and degradation (i.e., blurry/low spatial frequency dominant cases). In these difficult cases, humans typically rely on context and prior experience to identify critical features for successful detection (Shulman & Wilson, 1987). It remains unclear whether AI systems can learn to perform the task in a similar way. Previous research has suggested that AI systems may not perform as well as humans in these difficult conditions (Zhang et al., 2018; see Gilroy et al., 2021 for a review). However, with recent advances in both the model design and training methods, it is possible that state-of-the-art AI systems have learned to identify critical features similar to humans even for difficult cases.

In addition, humans and AI may differ in detecting object categories with evolutionary significance to humans. More specifically, humans are shown to have category-specific biases in object detection, where detection of animals is more spontaneous and reliable than that of artificial objects. These biases can be attributed to ancestral influence (New et al., 2007; Öhman, 2007). In particular, humans show high sensitivity in detecting human bodies and faces (Downing et al., 2001; Hodzic et al., 2009). fMRI studies have revealed cortical regions that correspond uniquely to human faces and bodies, suggesting that visual perception of these categories is distinct from others (e.g., Aleong & Paus, 2010; Downing et al., 2001; Kanwisher et al., 1997). This category-specific bias cannot be explained by visual characteristics or objects' attractiveness alone, suggesting that human object detection systems prioritize stimuli with evolutionary significance. Thus, human observers may outperform AI models particularly in detecting humans as compared with artificial object categories such as vehicles in driving scenarios.

Accordingly, here we examined humans' and AI systems' performance and attended features in detecting vehicles and humans in driving scenarios, and how they were affected by occlusion and degradation conditions. Humans' attended features were measured using eye tracking, and AI's attended features were generated using a saliency-based XAI. Since humans may differ significantly from one another in eye movement patterns during object detection (Boot et al., 2009; Hsiao, Chan, et al. 2021), we used a data-driven 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), to discover representative participant groups where group members adopted similar eye movement patterns to one another across stimuli. We then examined whether a particular eye movement pattern group was associated with better object detection performance (i.e., experts), and compared it with a widely used object detection AI model, Yolo-v5s, which has great real-time performance and thus has been commonly used for driving scenarios (Redmon et al., 2016). We hypothesized that humans may outperform Yolo-v5s in detecting humans but not in detecting vehicles, and they may have higher similarity in detecting vehicles than detecting humans. Both humans and Yolo-v5s may be similarly affected by occlusion and degradation. Assuming that human experts attended to the most critical features for vehicle/human detection, higher similarity between Yolo-v5s' and humans' attended features may be associated with better Yolo-v5s' performance.

Study 1: Humans vs. AI in Detecting Vehicles

Methods

Participants We recruited 60 participants with normal or corrected-to-normal vision, aged 18 to 40 (M = 23.9; SD = 4.33; 48 females). To facilitate identification of experts, all participants had a driver's license.

Materials and Apparatus The Berkeley DeepDrive 100KImage (BDD100K) Dataset was used (Yu et al., 2020). It comprises 10 target categories. We chose the car, truck, and bus categories as the vehicle target in this task. Previous research has suggested a location memory limit of 3 to 5 items in young adults (Cowan, 2010). Accordingly, we selected all images with 1 to 4 vehicle targets to form our stimulus set (1366 images). Occlusion and degradation conditions in each image were rated according to the majority choice of three human raters (with good inter-rater reliability, Occlusion: $\alpha = .881$; Degradation: $\alpha = .877$; Cronbach, 1951). Images with any target occluded by any other object or the image boundary were rated as 'occluded.' Images containing any target whose feature identification was influenced by night vision, motion blur, uneven illumination, shadow, or light reflection were rated as 'degraded.' We assessed Yolo-v5s' performance on this stimulus set. We then randomly selected 160 images as the stimuli for the human study to compare AI and human performance and attention maps.

In the human study, the stimuli were displayed one at a time at the center of a 15.6-inch monitor (1920 x 1080 pixels), spanning 34.2 ° x 20.8° of visual angle at a viewing distance of 55 cm. Participants' eye movements were recorded using EyeLink 1000 Plus. A nine-point calibration procedure was performed before the experiment and occurred whenever drift check error exceeded 1° of visual angle.

Design EMHMM with co-clustering was used to discover representative eye movement pattern groups in the participants and quantify eye movement pattern similarities among the participants (See Eye Movement Analysis for details). We examined whether participants belonging to different representative pattern groups differed significantly in performance using an independent sample t-test and considered the better-performing group as the human experts to be compared with Yolo-v5s. We conducted by-items ANOVA to examine the effect of occlusion and degradation (as between-item variables) on Yolo-v5s' performance over the 1366 images. Then, we conducted ANOVA with humans vs. AI as an additional (within-item) independent variable to

compare the performance of Yolo-v5s with human experts on the 160 randomly selected images. Correlation analyses were used to measure the similarity between the attention strategies of the human experts and Yolo-v5s and to examine the relationship between their similarity and Yolo-v5s' performance.

Procedure Each trial started with a fixation cross at the screen center. The experimenter initiated the stimulus presentation when a stable fixation was observed at the fixation cross. Participants were asked to detect all vehicle targets in the image and press a key as soon as they thought they had detected all. Their eye movements during the visual search before the key press were used for data analysis. To assess detection performance, immediately after the key press, participants were asked to use a mouse click to place a marker at each detected target location on a blank screen. Then, they were asked to click again on the same objects they had clicked previously on the original image to confirm their selection (Figure 1). Here we reported the results based on the clicks on the blank screen (similar results were obtained using the clicks on the original image). Performance was assessed by hit rate: the number of correctly detected targets divided by the total number of targets (same for Yolo-v5s). Haladjian and Pylyshyn (2011) reported an average location error of 2.2° of visual angle in a spatial memory task with clicking responses. Accordingly, we used 64-pixel (2.2°) location error tolerance when calculating the hit rate.

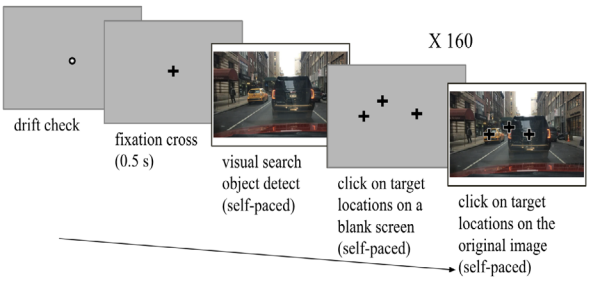


Figure 1: Procedure of the object detection task

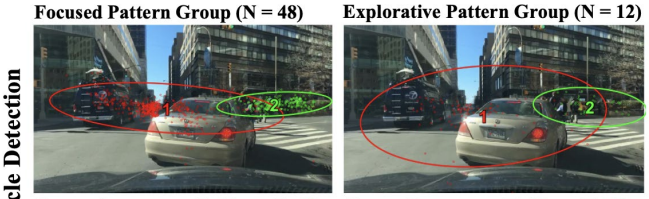
Eye Movement Analysis EMHMM with co-clustering was used. More specifically, each participant's eye movement data for each stimulus was summarized using a hidden Markov model (HMM) with personalized regions of interest (ROIs) and a transition matrix indicating the transition probabilities among the ROIs. The optimal number of ROIs for each HMM was determined from a range of 1 to 5 using a variational Bayesian approach. Each HMM was trained 300 times to select the model with the greatest log-likelihood. Participants were co-clustered into two pattern groups (A and B) where group members used similar eye movement patterns to one another across stimuli. A representative HMM with the number of ROIs set to the median number of the individual HMMs was generated for each group and each stimulus. We repeated the co-clustering procedure 300 times to select the model with the greatest log-likelihood.

Following previous studies (e.g., Hsiao, Lan, et al., 2021), we quantified each participant's attention strategy along the dimension contrasting the two group patterns using the AB scale: (A – B)/(A + B), where A and B are the log-likelihoods of a participant's data generated by Pattern Group A and B respectively. Larger/positive AB scales indicate higher similarity to Group A in contrast to Group B.

Human & AI Attention Maps To compare human and AI attention strategies, we generated human attention maps by applying a Gaussian smoothing kernel with a 30-pixel SD (equivalent to 1° of visual angle) to each fixation location over all (expert) participants. AI's attention (i.e., saliency) maps were generated by FullGrad-CAM++ algorithm, an XAI method designed for object detection model (Liu et al., submitted). Assume N_{obj} is the total number of detected objects, and with $m = 1, 2 \ldots, N_{\text{obj}}$ is the output classification probability of m-th detected object, the FullGrad-CAM++ can be defined as:

$$S_F^* = \sum_{m=1}^{N_{obj}} \mu\left(ReLU\left(\sum_{k=1}^{N_{ch}} ReLU\left(\frac{\partial y^m}{\partial A^k} \right) \odot A^k \right) \right), \quad (1)$$

where μ is the max-min normalization function that normalizes the data map to scale between 0 to 1, A^k is the activation map in the k-th layer, N_{ch} is the number of channels in A^k, $\odot$ is the Hadamard product, and $ReLu$ is the rectified linear unit function. This work computed the FullGrad-CAM++-based saliency maps using the last convolutional layer of the backbone in Yolo-v5s. By removing the global average pooling operation on gradient term, this method can better capture spatial information in object detection than the vanilla Grad-CAM and Grad-CAM++ methods, making its generated saliency maps more faithful. (Liu et al., 2023a, 2023b; Zhao & Chan, 2022). We used Pearson correlation coefficient (PCC) to assess the similarity between human and AI attention maps (Le Meur & Baccino, 2013).



Group A	To R	To G
Priors	.96	.04
From Red	.90	.10
From Green	.30	.70

Group B	To R	To G
Priors	.99	.01
From Red	.96	.04
From Green	.40	.60

Figure 2: The focused and explorative eye movement pattern group for vehicle detection. Ellipses show ROIs as 2-D Gaussian emissions. Priors show probabilities of the first fixation landing on each ROI and transition matrices show transition probabilities among the ROIs.

Results

Human Attention Strategy and Its Association with Performance EMHMM with co-clustering resulted in the focused and explorative pattern groups/attention strategies

(Figure 2). Participants using the focused strategy preferred to scan along the horizon, where vehicle targets usually occur. In contrast, those using the explorative strategy had larger/rounder ROIs, scanning across a broader area beyond the horizon. Accordingly, we referred to the AB scale as the FE (Focused-Explorative) scale. Pattern group by stimulus mixed ANOVA analysis on KL divergence estimation revealed that the two groups differed significantly, F =5.642, $p < .001$, $\eta^2_p = .089$, and this effect interacted with stimulus, $F(1, 159) = 2.16$, $p < .001$, $\eta^2_p = .036$, indicating that this group difference was larger for some stimuli than others.

Participants adopting the focused strategy outperformed those using the explorative strategy (Focused: $M = .756$, $SD = .07$; Explorative: $M = .713$, $SD = .04$), $t(58) = 2.61$, $p = .012$, $d = .842$. Consistent with this finding, participants' attention strategies as assessed by the FE scale showed a positive correlation with performance, $r(58) = .303$, $p = .019$.

Effect of Occlusion and Degradation in Humans vs. AI In AI's (Yolo-v5s) performance over the stimulus set (1366 images), the mean hit rate was .759 ($SD = .295$). There were main effects of occlusion, $F(1, 1362) = 4.63$, $p = .032$, $\eta^2_p = .003$, and degradation, $F(1, 1362) = 99.28$, $p < .001$, $\eta^2_p = .068$, with better performance in the non-occluded and the non-degraded conditions respectively. The interaction between occlusion and degradation was significant, $F(1, 1362) = 17.48$, $p < .001$, $\eta^2_p = .013$: The occlusion effect was significant in the non-degraded condition, $t = 3.56$, $p = .002$, $d = .474$, but not in the degraded condition, $t = 2.21$, $p = .120$, $d = .152$.

When we compared human experts (those using the focused strategy) and Yolo-v5s on their performance in the 160 randomly selected images (Figure 3), there was no main effect of humans vs. AI, $F(1, 156) = 0.35$, $p = .557$, $\eta^2_p = .002$, indicating that human experts ($M = .820$, $SE = .023$) and AI had comparable performance ($M = .809$, $SE = .029$; Figure 4). The main effect of occlusion was marginally significant, $F(1, 156) = 3.75$, $p = .055$, $\eta^2_p = .023$, and the main effect of degradation was significant, $F(1, 156) = 13.36$, $p < .001$, $\eta^2_p = .079$. However, they did not interact with humans vs. AI (Occlusion: $F(1, 156) = 0.02$, $p = .902$, $\eta^2_p = .000$; Degradation: $F(1, 156) = 1.45$, $p = .861$, $\eta^2_p = .009$), suggesting that occlusion and degradation influenced humans and Yolo-v5s similarly.

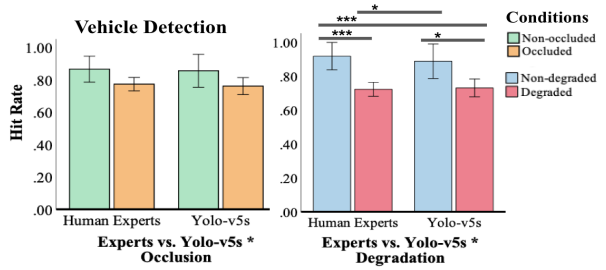


Figure 3: Difference in vehicle detection performance between humans and Yolo-v5s under occlusion and degradation conditions (error bars: 95% CI; $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$).

Do Yolo-v5s and Humans Attend to Similar Features when Detecting Vehicles? When we examined similarities between human and AI attention maps, we found that AI attention maps had a higher similarity to the focused group ($M = .659$, $SD = .205$) than the explorative group ($M = .644$, $SD = .204$) in humans, $t(159) = 2.55$, $p = .012$, $d = .202$. This result showed that AI's attention strategy was more similar to the participant group with better detection performance (See Figure 4A for an example). In addition, higher similarity of AI's attention strategies to human experts' was associated with better AI performance, $r(158) = .408$, $p < .001$. This correlation remained significant when we used only the trials where AI outperformed human experts, $r(156) = .394$, $p < .001$. This result suggested that features attended by human experts could be used as guidance for object detection AI to enhance their performance (Figure 4B).

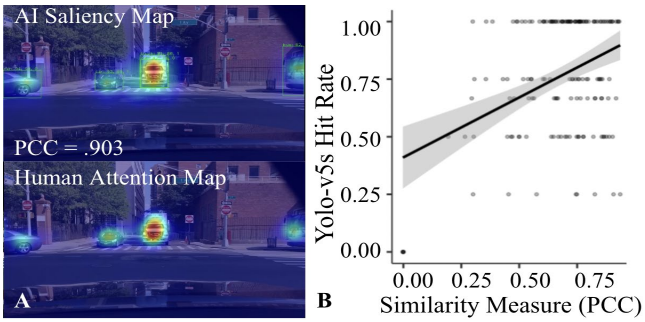


Figure 4: (A) Example showing high similarity between attended features of AI and human experts. (B) Positive correlation between Yolo-v5s' attended feature similarity with human experts and Yolo-v5s' performance (in selected trials where it outperformed humans).

Study 2: Humans vs. AI in Detecting Humans

Methods

Participants Following Study 1's criteria, we recruited 60 participants aged 18 to 40 ($M = 24.4$; $SD = 4.86$; 48 females).

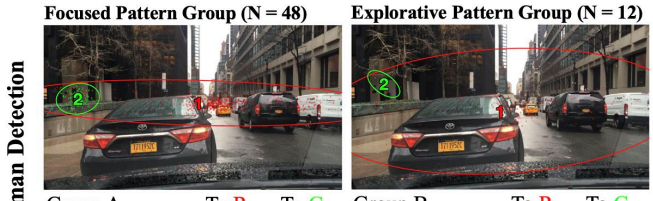
Materials and Apparatus The stimuli selection was similar to Study 1: 160 stimuli with human targets were randomly selected from all images with 1 to 4 human targets (2461 in total), including pedestrians and riders. The inter-rater reliability for occlusion ($\alpha = .886$) and degradation ($\alpha = .769$) ratings were relatively high.

Design and Procedure The design and procedure were identical to Study 1, and similar for eye movement analysis and the generation of human and AI attention maps.

Results

Human Attention Strategy and Its Association with Performance Consistent with Study 1, EMHMM with co-clustering resulted in the focused and explorative pattern

groups (Figure 5) that differed significantly: Pattern group by stimulus ANOVA on KL divergence estimation showed a significant main effect of pattern group: $F = 4.41$, $p < .001$, $\eta^2_p = .07$ (It interacted with stimulus, $F = 3.62$, $p < .001$, $\eta^2_p = .05$). Participants' eye movement patterns assessed by the FE scale revealed a positive correlation with performance, $r(58) = .318$, $p = .013$, although t-test on pattern group difference in performance did not reach statistical significance (Focused: $M = .688$, $SD = .08$; Explorative: $M = .647$, $SD = .09$), $t(58) = 1.55$, $p = .127$, $d = .499$.



Group A	To R	To G	Group B	To R	To G
Priors	1.0	.00	Priors	1.0	.00
From Red	.97	.03	From Red	.98	.02
From Green	.48	.52	From Green	.65	.35

Figure 5: The focused and explorative eye movement pattern group for human detection. Ellipses show ROIs as 2-D Gaussian emissions.

Effect of Occlusion and Degradation in Humans vs. AI In Yolo-v5s' performance over the stimulus set (2461 images), the mean hit rate was .473 ($SD = .435$). There were main effects of occlusion, $F(1, 2457) = 7.41$, $p = .007$, $\eta^2_p = .003$, and degradation, $F(1, 2457) = 8.89$, $p = .003$, $\eta^2_p = .004$: It performed in the non-occluded and in the non-degraded conditions respectively. The design and procedure were identical to Study 1, and similar for eye movement analysis and the generation of human and AI attention maps.

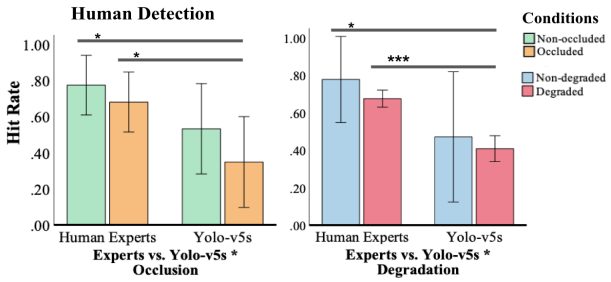


Figure 6: Difference in human detection performance between humans and Yolo-v5s under occlusion and degradation conditions (error bars: 95% CI; $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$).

Since the FE scale was positively correlated with performance, we used participants in the focused group as experts in the analysis. When we compared them with Yolo-v5s on their performance in the 160 randomly selected images, there was a main effect of humans vs. AI, $F(1, 156) = 14.90$, $p < .001$, $\eta^2_p = .087$, indicating that humans ($M = .728$, $SE = .059$) outperformed Yolo-v5s ($M = .441$, $SE = .090$). The main effects of occlusion, $F(1, 156) = 1.09$, $p = .297$, $\eta^2_p = .007$, and degradation were not significant, $F(1, 156) = 0.39$, $p = .534$, $\eta^2_p = .002$. In addition, occlusion and degradation did not interact with humans vs. AI (Occlusion: $F(1, 156) = 0.37$, $p = .902$, $\eta^2_p = .000$; Degradation: $F(1, 156) = 0.07$, $p = .790$, $\eta^2_p = .000$). Thus, Yolo-v5s performed poorly regardless of occlusion and degradation conditions (Figure 6).

Do Yolo-v5s and Humans Attend to Similar Features when Detecting Humans? The similarities between human and AI attention maps were low due to Yolo-v5s' poorer performance than humans (Figure 7A). Also, AI attention maps' similarities to the focused group ($M = .312$, $SD = .304$) and to the explorative group ($M = .308$, $SD = .311$) did not differ significantly, $t(159) = 0.24$, $p = .809$, $d = .019$. Since no saliency information could be generated for trials without any detected target in Yolo-v5s, we selected images (88 images) where both AI and humans detected at least one target (mean hit rate > 0) and found that AI attention maps had a higher similarity to the focused group ($M = .555$, $SD = .169$) than the explorative group in humans ($M = .492$, $SD = .245$), $t(87) = 2.74$, $p = .007$, $d = .292$. In addition, consistent with Study 1, higher similarity of AI's attention strategies with human experts was correlated with better AI performance, $r(158) = .823$, $p < .001$. This correlation remained significant when we only included trials where AI outperformed human experts, $r(146) = .850$, $p < .001$, suggesting that human attention strategies could be used to guide AI to enhance their performance (Figure 7B).

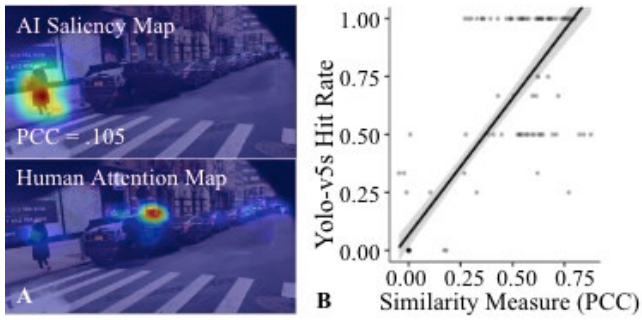


Figure 7: (A) Example showing low similarity between attended features of AI and human experts. (B) Positive correlation between Yolo-v5s' attended feature similarity to human experts' and Yolo-v5s' performance (in selected trials where it outperformed humans).

Discussion

Here we examined the similarities and differences between humans and AI (Yolo-v5s) in performance and attended features in detecting vehicles and humans in driving scenarios. Previous research has suggested that humans may outperform AI in object detection, particularly in difficult conditions such as occlusion and degradation. We speculated that with the recent advances in deep learning methods,

current AI models may have achieved human-expert-level performance with similar attended features to humans' even in difficult cases, with one exception: the detection of object categories with evolutionary significance such as animals and humans. Also, assuming that human experts have learned to attend to the most critical features for detection, AI models may have better performance when their attended features have higher similarity to humans'.

To test these hypotheses in driving scenarios, which have important implications for self-driving applications, we contrasted the detection of vehicles and humans. We first used a data-driven method, EMHMM with co-clustering, to identify representative attention strategies from humans. We discovered the focused and explorative strategies in both vehicle and human detection. Participants adopting the focused strategy preferred to scan the horizon where relevant targets typically occur, whereas those using the explorative strategy scanned a broader area. In both detection tasks, the focused strategy was associated with better performance, suggesting that using prior knowledge about the scene context (such as potential locations where targets may appear) facilitated detection. Interestingly, Yolo-v5s' attended features had higher similarity to the focused group than the explorative group in humans in both detection tasks. This result suggested that the current AI may have implicitly learned relevant contextual information similar to humans regardless of the lack of a top-down attention mechanism similar to humans' to guide the search due to their ability to process all information in parallel. Nevertheless, while Yolo-v5s showed comparable performance and high similarity in attended features to human experts in detecting vehicles even in difficult conditions, it performed significantly poorer than humans with low similarity in attended features in human detection. This result was consistent with our hypothesis: Humans have a category-specific attention bias for detecting animals due to their evolutionary significance (New et al., 2007). This higher vigilance, developed during evolution, for detecting animate objects than inanimate objects has been attributed to our ancestors' survival and adaptation in hunter-gatherer societies, where animals could be a meal or a dangerous threat; and humans could be friends or foes (Öhman, 2007). In contrast, vehicles are artificial objects created more recently. Although detecting vehicles has life-or-death consequences in the modern society, they did not exist in the ancestral environment (New et al., 2007). This difference may explain why humans outperformed Yolo-v5s in detecting humans but not in vehicles.

Consistent with its poorer performance in detecting humans, Yolo-v5s' attended features also showed low similarity to humans'. This finding suggested that despite the technological advances, current AI still attends to suboptimal features as compared with humans in human detection. Humans' superior ability in detecting animated objects may be related to the brain regions specialized for the recognition of object categories of evolutionary significance such as faces (e.g., Buiatti et al., 2019). In contrast, current AI systems are trained from a uniform architecture without specialized modules to learn to detect a particular category. To learn from humans' superiority resulting from a long history of evolutionary processes, future AI development may consider using human data as guidance to search for better solutions. Indeed, we found that in Yolo-v5s, higher similarity in attended features to humans' was associated with better performance, particularly in human detection, suggesting that human attention could be used to guide AI to learn better features for the task. These associations remained consistent for cases where humans performed worse than AI, suggesting that humans might be aware of where to attend to even when failed to detect the objects. Recently, human attention has been used to enhance AI' performance and explainability. For instance, Selvaraju et al. (2019) used human attention as guidance to improve model performance in visual question answering and image captioning tasks and outperformed other approaches with much less training data. Baek et al. (2021) trained a DNN model of human ventral visual stream to equip the model with face-selectivity in the absence of training. These findings along with our results suggest a promising direction in using human attention to enhance both AI and XAI designs (Liu et al., 2023a, 2023b).

Our findings about the differences between humans and AI also have important implications for XAI design. According to Yang, Folke et al. (2022), humans form beliefs about AI, assuming it would make similar decisions to their own, similar to how humans interact with each other (i.e., Theory of Mind ability). These beliefs can be updated through explanations generated by XAI methods, especially when the explanations highlight the discrepancy between what humans expected the AI to do and what the AI actually does. Therefore, the differences in performance and attended features across different detection tasks between humans and AI observed in this study could inform future XAI research to examine how to best use this information to update humans' beliefs about AI to enhance both user trust and understanding, and to truly satisfy stakeholders' desiderata (Hsiao, Ngai, et al., 2021; Qi et al., 2023a, 2023b).

In conclusion, here we showed that in humans, individuals differed in adopting more focused or explorative attention strategies in both vehicle and human detection in driving scenarios, with better performance associated with the focused strategies. Interestingly, current AI (Yolo-v5s) showed higher similarity in attended features to the focused than the explorative strategies, and achieved a similar performance level to human experts in vehicle detection even in difficult cases. Nevertheless, it performed much poorer than humans in detecting humans with low attended feature similarity due to humans' attention bias for detecting stimuli with evolutionary significance. Also, higher similarity to humans' attended features was associated with better AI performance, suggesting that human attention may be used to guide AI design. The observed differences between AI and humans could help XAI update users' beliefs about AI based on their mental models. Thus, our findings have significant implications for both AI and XAI designs.

Acknowledgments

We are grateful to Huawei and RGC of Hong Kong (Collaborative Research Fund No. C7129-20G to J. Hsiao). We thank Yi Yang, Caleb Cao, Yueyuan Zheng, Hilary Hei Ting Ngai, Hyunseo Cho for their help in data collection and helpful comments.

References

Aleong, R., & Paus, T. (2010). Neural correlates of human body perception. *Journal of Cognitive Neuroscience*, 22(3), 482-495. https://doi.org/10.1162/jocn.2009.21211

Boot, W. R., Becic, E., & Kramer, A. F. (2009). Stable individual differences in search strategy?: The effect of task demands and motivational factors on scanning strategy in visual search. *Journal of Vision*, 9(3), Article 7. https://doi.org/10.1167/9.3.7

Baek, S., Song, M., Jang, J., Kim, G., & Paik, S.-B. (2021). Face detection in untrained deep neural networks. *Nature Communications*, 12(1), Article 7328. https://doi.org/10.1038/s41467-021-27606-9

Betz, T., Kietzmann, T. C., Wilming, N., & Konig, P. (2010). Investigating task-dependent top-down effects on overt visual attention. *Journal of Vision*, *10*(3), Article 15. https://doi.org/10.1167/10.3.15

Buiatti, M., Di Giorgio, E., Piazza, M., Polloni, C., Menna, G., Taddei, F., Baldo, E., & Vallortigara, G. (2019). Cortical route for facelike pattern processing in human newborns. *Proceedings of the National Academy of Sciences*, 116(10), 4625-4630. https://doi.org/doi:10.1073/pnas.1812419116

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *Journal of Vision*, *14*(11), Article 8. https://doi.org/10.1167/14.11.8

Cowan, N. (2010). The Magical Mystery Four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51-57. https://doi.org/10.1177/0963721409359277

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. https://doi.org/10.1007/bf02310555

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470-2473. https://doi.org/10.1126/science.1063414

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*, 115-118. https://doi.org/10.1038/nature21056

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology*, 15(17), R644-R645. https://doi.org/10.1016/j.cub.2005.08.041

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). *Comparing deep neural networks against humans: object recognition when the signal gets weaker*. arXiv. https://arxiv.org/abs/1706.06969

Gilroy, S., Jones, E., & Glavin, M. (2019). Overcoming occlusion in the automotive environment—A review. *IEEE Transactions on Intelligent Transportation Systems*, 22(1), 23-35. https://doi.org/10.1109/TITS.2019.2956813

Gulshan, V., Rajan, R. P., Widner, K., Wu, D., Wubbels, P., Rhodes, T., Whitehouse, K., Coram, M., Corrado, G., Ramasamy, K., Raman, R., Peng, L., & Webster, D. R. (2019). Performance of a Deep-Learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmol*, *137*(9), 987-993. https://doi.org/10.1001/jamaophthalmol.2019.2004

Haladjian, H. H., & Pylyshyn, Z. W. (2011). Enumerating by pointing to locations: A new method for measuring the numerosity of visual object representations. *Attention, Perception, & Psychophysics,* 73(2), 303-308. https://doi.org/10.3758/s13414-010-0030-5

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.48550/arXiv.1502.01852

Hodzic, A., Kaas, A., Muckli, L., Stirn, A., & Singer, W. (2009). Distinct cortical networks for the detection and identification of human body. *Neuroimage*, 45(4), 1264-1271. https://doi.org/10.1016/j.neuroimage.2009.01.027

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. https://doi.org/10.1016/j.cognition.2021.104616

Hsiao, J. H., & Chan, A. B. (2023). Visual attention to own- vs. other-race faces: Perspectives from learning mechanisms and task demands. *British Journal of Psychology*. 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, 1933–1943. https://doi.org/10.3758/s13423-021-01944-7

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. https://doi.org/10.3758/s13428-021-01541-5

Hsiao, J. H. W., 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

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11), 4302-4311. https://doi.org/10.1523/jneurosci.17-11-04302.1997

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition.

Scientific Report, 6(1), 32672. https://doi.org/10.1038/srep32672

Lake, B.M., Zaremba, W., Fergus, R., & Gureckis, T.M. (2015). Deep neural networks predict category typicality ratings for images. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1), 251-266. https://doi-org.eproxy.lib.hku.hk/10.3758/s13428-012-0226-9

Lindebaum, D., Vesa, M., & Hond, F. d. (2020). Insights from "the Machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Academy of Management Review*, *45*(1), 247-263. https://doi.org/10.5465/amr.2018.0181

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*.

Liu, G., Zhang, J., Chan, A. B., & Hsiao, J. H. (2023b). *Human Attention-Guided Explainable Artificial Intelligence for Computer Vision Models*. arXiv. https://arxiv.org/pdf/2305.03601

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://doi.org/10.48550/arXiv.1902.01876

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353-383. https://doi.org/10.1016/0010-0285(77)90012-3

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, 104(42), 16598-16603. https://doi.org/10.1073/pnas.0703913104

Öhman, A. (2007). Has evolution primed humans to "beware the beast"? *Proceedings of the National Academy of Sciences*, 104(42), 16396-16397. https://doi.org/10.1073/pnas.0707885104

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429).* IEEE. https://ieeexplore.ieee.org/document/1246946

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, *155*, 23-36. https://doi.org/10.1016/S0079-6123(06)55002-2

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, 29, 441-459. https://doi.org/10.1007/s11023-019-09502-w

Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2023a). *Explanation strategies for image classification in humans vs. current explainable AI*. arXiv. https://arxiv.org/abs/2304.04448

Qi, R., Zheng, Y., Yang, Y., Zhang, J., & Hsiao, J. H. (2023b). Individual differences in explanation strategies for image classification and implications for explainable AI. *Proceedings of the Annual Conference of the Cognitive Science Society*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788). https://doi.org/10.1109/CVPR.2016.91

Selvaraju, R. R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., ... & Parikh, D. (2019). Taking a hint: Leveraging explanations to make vision and language models more grounded. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 2591-2600). https://doi.org/10.48550/arXiv.1902.03751

Shulman, G. L., & Wilson, J. (1987). Spatial frequency and selective attention to local and global information. *Perception*, 16(1), 89-101. https://doi.org/10.1068/p160089

Wolfe, B., Sawyer, B. D., & Rosenholtz, R. (2022). Toward a Theory of Visual Information Acquisition in Driving. *Human Factors*, 64(4), 694-713. https://doi.org/10.1177/0018720820939693

Yang, S. C.-H., Folke, N. E. T., & Shafto, P. (2022). A Psychological Theory of Explainability. *Proceedings of the 39th International Conference on Machine Learning* (pp. 25007–25021). https://doi.org/10.48550/arXiv.2205.08452

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.48550/arXiv.1805.04687

Zhao, C., & Chan, A. B. (2023). ODAM: Gradient-based Instance-Specific Visual Explanations for Object Detection. *International Conference on Learning Representations (ICLR).* https://doi.org/10.48550/arXiv.2304.06354

Zhang, Zhishuai, et al. (2018). "Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.48550/arXiv.1709.04577