Towards the next generation explainable AI that promotes AI-human mutual understanding

Janet H. Hsiao
Department of Psychology
The University of Hong Kong
jhsiao@hku.hk

Antoni B. Chan
Department of Computer Science
City University of Hong Kong
abchan@cityu.edu.hk

Abstract

Recent advances in deep learning-based AI has necessitated better explanations on AI's operations to enhance transparency of AI's decisions, especially in critical systems such as self-driving car or medical diagnosis applications, to ensure safety, user trust and user satisfaction. However, current Explainable AI (XAI) solutions focus on using more AI to explain AI, without considering users' mental processes. Here we use cognitive science theories and methodologies to develop a next-generation XAI framework that promotes human-AI mutual understanding, using computer vision AI models as examples due to its importance in critical systems. Specifically, we propose to equip XAI with an important cognitive capacity in human social interaction: theory of mind (ToM), i.e., the capacity to understand others' behaviour by attributing mental states to them. We focus on two ToM abilities: (1) Inferring human strategy and performance (i.e., Machine's ToM), and (2) Inferring human understanding of AI strategy and trust towards AI (i.e., to infer Human's ToM). Computational modeling of human cognition and experimental psychology methods play an important role for XAI to develop these two ToM abilities to provide user-centered explanations through comparing users' strategy with AI's strategy and estimating user's current understanding of AI's strategy, similar to real-life teachers. Enhanced human-AI mutual understanding can in turn lead to better adoption and trust of AI systems. This framework thus highlights the importance of cognitive science approaches to XAI.

1 Introduction

To ensure good use of AI to humans, researchers have long recognized the importance of explanation to enhance human-AI interaction. According to a review on Explainable AI (XAI) literature [46], the first-generation XAI was the expert systems developed in late 1970s to 1980s [47], aiming to aid decision-making processes. Nevertheless, the explanations were often considered inadequate in justifying rule-based inferences [63]. This issue was addressed in the second-generation XAI in 1990s and early 2000s, which focused on knowledge-based tutors that could infer user mental models from their behaviour and make context-sensitive explanations [5]. In addition, some have proposed to provide explanations according to user knowledge and capacities to provide better cognitive support [18].

Around mid-2010s, a third generation of XAI has emerged due to the advance of deep learning methods, whose decision-making processes are often obscured to both users and developers. Similar to the first generation, the third generation has focused on revealing the operation of AI, with better visualization techniques to make deep-learning classifiers explainable [22]. Others have proposed techniques to directly make classifiers more explainable [2]. Although some

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

have attempted to provide explanations with adequate justifications [49] or emphasize on user interactions [38] to address challenges faced in first two generations, related work remains limited.

Throughout the literature, the key issues have been how to help users develop a robust mental model of AI, instead of just presenting the rules used by AI. However, most current XAI solutions remain focusing on using more AI to explain AI without considering users' mental processes, such as using proxy models [72, 65], correlative techniques with saliency map visualization [55], or extracting concepts relevant to the decisions [16, 37] . Most recently, some have proposed that a good XAI should consider users' cognitive states during explanation processes [28, 34, 62], such as considering explanation as a conversational process that involves both social and cognitive processes across different levels of explanation [17], or as a model reconciliation problem, where the explanation is the minimal set of changes to the user's mental model so that the AI's decision is optimal with respect to the user's updated model, i.e., the task-important differences between the AI and the user's models. [10, 61, 11]. Cognitive metrics have been used for comparing XAI methods in addition to computational metrics [55]. A new concept of XAI, Artificial Cognition, has been proposed to learn from cognitive psychology methodologies on revealing the mechanisms underlying the black-box human brain through experimental approaches to study machine behaviour [64]. Together these suggest that cognitive science is going to play an essential role in the next generation of XAI.

At the same time in the history of Cognitive Science, there have been a gradual paradigm shift from box-and-arrow models with encapsulated cognitive components to the consideration of broader, more interactive, and more dynamic contextual effects and processes [60], including the interactions with the body and the environment (embodied cognition [7]), and with other individuals or cognitive systems (joint cognition [58]). Accordingly, the explanation process involved in XAI can be better conceptualized as dynamic interactions between the user and the XAI system, with an aim to provide explanations that can facilitate the user's construction of a robust cognitive model of the AI system. To make the explanations accessible to human users, XAI systems may learn from how humans explain and learn from explanations through explainer-explainee interactions [44, 54, 53]. In particular, an effective XAI system should consider an important cognitive capacity in human social interaction: *theory of mind (ToM)*, i.e., the capacity to understand others' behaviour by attributing mental states (knowledge, beliefs, etc.) to them [6, 3]. Indeed, it has been recently shown that humans tend to initially assume that AI performs a task in a similar way to them, and update this belief after receiving explanations [69, 68]. Thus, information about the difference between users' cognitive model of the task and the AI model of the task, and users' updated cognitive model of the AI system during the explanation process, are both essential for providing human-accessible explanations.

Accordingly, we propose that an effective XAI system should possess the ToM ability to evaluate how humans solve the same problem as the AI system (i.e., ToM about humans' cognitive models of the task), as well as the ToM ability to infer users' understanding of AI's operation (i.e., to infer human's ToM about AI's cognitive model). Using these ToM abilities, the XAI system can present the most informative explanations to update the human's current understanding of AI, as it relates to how the human solves the task. In doing so, the XAI can establish mutual understanding between the AI and the user. In this work, we outline our proposed framework of a ToM-based XAI system, and present our recent studies in experimental psychology and computational modeling on the path to achieve this XAI.

2 Theory-of-mind-based XAI

Our novel framework of ToM-based XAI is illustrated in Figure 1, where an effective XAI uses (1) information about performance and strategy differences between humans and AI models when performing the task, (2) information from models of human strategies of the task, and (3) information from models of users' understanding of AI's strategy and trust on AI, to provide explanations that are effective and accessible to humans. Built on the research reviewed above, below we discuss how we can use cognitive science methods, in particularly computational modelling and experimental psychology methods, to help obtain the required information for an effective ToM-based XAI that can truly facilitate human-AI interaction and mutual understanding and human trust towards AI.

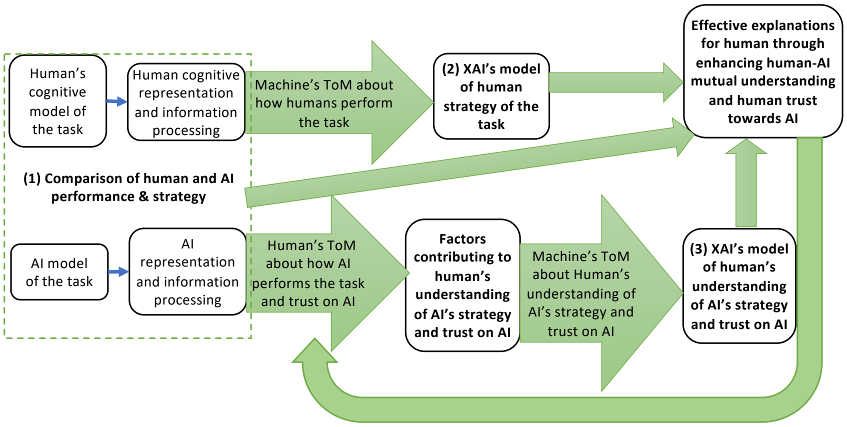

Human's cognitive model of the task
Human cognitive representation and information processing
Machine's ToM about how humans perform the task
(2) XAI's model of human strategy of the task
Effective explanations for human through enhancing human-AI mutual understanding and human trust towards AI
(1) Comparison of human and AI performance & strategy
AI model of the task
AI representation and information processing
Human's ToM about how AI performs the task and trust on AI
Factors contributing to human's understanding of AI's strategy and trust on AI
Machine's ToM about Human's understanding of AI's strategy and trust on AI
(3) XAI's model of human's understanding of AI's strategy and trust on AI

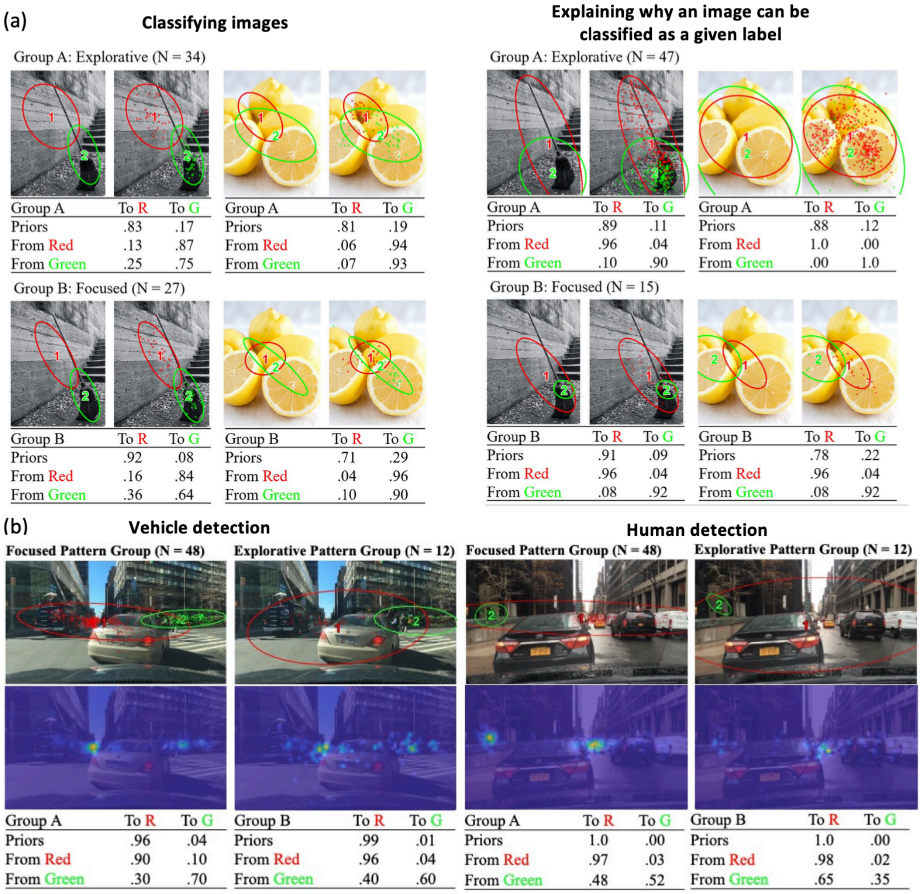
Figure 1: Proposed XAI framework that conceptualizes the explanation process as dynamic interactions between the user and the XAI system and considers the Theory of Mind (ToM) capacity in human social interaction.

2.1 Comparison of human and AI performance and strategy

Comparison studies between humans and AI models provide information about their difference in performance and strategy, which is essential for human-accessible explanations [68]. In humans, eye movement during a cognitive task has been commonly used as a direct measure of human attention strategy, and can be conceptualized as human predictions about the locations of diagnostic information for the task [25]. It is task-driven and can differ significantly when the task demand changes [36, 31]. Previous work has consistently reported substantial individual differences in eye movements during cognitive tasks [50], which are typically associated with differences in task performance and cognitive abilities (e.g., [15, 30, 14]). Understanding these individual differences in attention strategy and performance and their comparability with AI' attention strategies and performances will not only enhance our understanding of cognitive processes in humans, but also inform ways to enhance AI's performance and XAI's ability to provide user-accessible explanations.

Image classification When performing an image classification task, we have recently shown through Eye Movement analysis with Hidden Markov Models (EMHMM [13]) with co-clustering [32] that individuals differed in adopting more focused or more explorative attention strategies (Figure 2a), and those adopting focused strategies had faster response times [54, 53]. Also, humans adopted more explorative attention strategies when they explain how they classify an image as compared with when they simply classify the image. Interestingly, current salience-based explanations from XAI (e.g., GradCAM [59]; RISE [51]) had higher similarity to the explorative than the focused attention strategy during explanation. These results may be because during image classification, humans only need to attend to sufficient information for a decision [33], whereas during explanation, they may attend to all relevant information to provide a comprehensive explanation [21]. In addition, we found that saliency-based XAI explanations that highlight discriminative features from invoking observable causality through perturbation (e.g., RISE) had higher similarity to human strategies than those highlighting internal features associated with higher class score (e.g., gradient-based method such as GradCAM). This result was consistent with recent findings that human explanations typically focus on causal reasoning based on observed regularities in the world [29, 8].

Object detection In another study, we compared human participants and current deep learning object detection AI models, including one-stage detector Yolo-v5s [35], two-stage detector Faster RCNN [56], and transformer-based object detector DETR [9], in their performance and attention strategies in detecting vehicles and in detecting humans under different driving scenarios (i.e., occlusion and degradation [67]). We showed that DETR outperformed humans in vehicle detection particularly in degraded conditions but did not outperform humans in human detection, possibly related to humans' enhance sensitivity in detecting stimuli with high evolutionary significance [48, 27]. Also, both Yolo-v5s and Faster RCNN were more affected by occlusion than humans, whereas DETR was not. This phenomenon may be related to DETR's attention mechanism for



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

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

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

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

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

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

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

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

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

Group A	To R	To G
Priors	1.0	.00
From Red	.97	.03
From Green	.48	.52

Group B	To R	To G
Priors	1.0	.00
From Red	.98	.02
From Green	.65	.35

Figure 2: (a) Explorative and focused attention strategies discovered in human participants when they classified images and when they explained why an image can be classified as a given label. (b) Focused and explorative strategies discovered in human participants when they detect vehicles and when they detect humans in driving scenarios.

image context during encoding and the learnable "anchors" to facilitate learning where or how to search from experience. Using EMHMM with co-clustering, we discovered two representative human attention strategies: a focused strategy to scan mainly along the horizon where vehicle targets usually occur, and an explorative strategy with larger and rounder ROIs, scanning across a broader area beyond the horizon (Figure 2b), and the focused attention strategy was associated with better object detection performance. Interestingly, although DETR has outperformed the other models and achieved a similar performance level to human experts who adopted the focused attention strategy, its attended features were more similar to the human explorative strategy than focused strategy, suggesting difference in information use. Human experts who adopted the focused strategy may be using their existing knowledge about scene semantics in addition to visual information, and thus could better focus their search on areas with higher target appearance probability. We also found that in AI models, higher similarity to humans' attended features was associated with better AI model performance, suggesting that human attention may be used for guiding AI design.

These studies demonstrate how we can conduct experiments using methods from cognitive psychology to examine scenarios where humans and AI models differ, and to infer the information processing mechanisms associated with these differences. In addition to providing more user-centred explanations based on these differences, comparison studies between humans and AI can help us identify potential difficult/outlier instances, adversarial attacks (subtle modification of an image to make it misjudged by AI), or novel solutions from AI. The findings will inform ways to adopt and collaborate with AI safely.

2.2 XAI's model of human strategy of the task

Collecting behavioural and cognitive data from humans when they perform the same task as AI can help XAI develop models of human strategy of the task and generate predictions about possible human strategies in a novel scenario of the task. Once this information is available, it can be compared with AI models' strategies, and information about the difference between humans' and AI models' strategies can be used to update human users' beliefs about AI models' decision-making processes (Figure 1). The predicted human attention may also help XAI detect potential adversarial attack to the AI, i.e., subtle modification of an image/input to make it misclassified by AI: an AI error associated with a large deviation from human attention behaviour may indicate an adversarial attack.

Image classification We have recently developed a Human Saliency Imitator to predict human attention maps during image classification using a deep learning model trained with human attention data for benchmarking salience-based explanations [70].

Object detection In another study, we have developed a Human Attention-Guided XAI (HAG-XAI), in which we used trainable activation functions and smoothing kernels to maximize XAI saliency map's similarity to human attention maps when performing object detection tasks [41, 42]. We found that, for object detection AI models, enhancing the XAI salience maps' similarity to human attention maps also increased their faithfulness (i.e., how well the highlighted regions of a saliency map reflect features diagnostic to AI's decisions [57]) of the XAI saliency maps. This method could also be used as a human attention imitator for object detection tasks.

2.3 XAI's model of human's understanding of AI's strategy

To develop a model of human users' understanding of AI's strategy, we may collect human data when they are interacting with AI or learning about AI's operations from the explanations generated by XAI (such as AI's attention strategies), and then examine how human behaviour is associated with learning performance and improvement in understanding. For example, user understanding may be assessed using simulatability, i.e., how well the user can predict AI's behaviour [24], including both forward and counterfactual simulations [19]. *Forward simulation* involves predicting the output given an input, and thus evaluates user's general understanding. *Counterfactual simulation* involves predicting the output for a perturbed input given the original input and output, and thus assesses understanding about specific features used by the AI. In research on learning analytics, multimodal data, including eye movement data to assessing attention strategies, have been commonly used to monitor learners' behaviour, mental states, and cognitive processes in order to optimize their learning [45]. These multimodal data can be used by XAI to develop predictive models of user understanding of AI's strategy. The predictive model can not only enable XAI to predict users' current understanding of AI in a novel scenario, but also inform us about good learning strategies that can be used by XAI to guide users' attention and facilitate their learning from XAI explanations.

Another potential measure of human's understanding of AI is to examine the match between human attention map during the *simulation* tasks and important features used by AI in the task, e.g., as demonstrated in saliency-map based XAI approaches for computer vision models. This "human-AI understanding" can be quantified using similarity measures between the human's attention map and the XAI saliency map, e.g., normalized scanpath saliency, KL divergence, or Pearson's correlation coefficient [67, 4]. Alternatively, faithfulness metrics, which evaluate whether saliency values truly reflect feature importance to the AI, can be calculated between the human's attention map and the AI model, e.g., using insertion/deletion [12, 52, 66], image perturbation [4], or completeness/soundness [23], thus measuring how well highlighted regions of a human attention map reflects the importance for predicting AI's output (i.e., Faithfulness of human attention map in predicting AI's output). For example, we have recently used this faithfulness measure to examine how well human attention strategies during object detection match the diagnostic features used by current object detection models [41, 42] including Yolo-v5s [35] and Faster RCNN [56].

2.4 Towards effective XAI by enhancing human-AI mutual understanding and trust

With the considerations of information processing mechanisms involved in an explanation process, we posit that an effective XAI system should have the capacity to infer humans' mental model of the task (i.e., XAI's ToM about how humans perform the task) and to infer Human's ToM about how AI performs the task (i.e., XAI's ToM about humans' current understanding of the AI's operations; see Figure 1) to achieve a high-level ToM capacity and provide explanations accordingly. The construction of such system will involve the understanding of the information processing mechanisms underlying both humans and AI systems for the given task, the scenarios where they show similarities and differences, and how they infer each other's mental models for mutual understanding. In the previous sections we have presented some example findings, which could inform better AI and XAI designs to achieve better user trust and satisfaction and better adoption of AI systems in the society.

Trust is also an important factor for guiding users' reliance on complex AI systems where a full understanding is impractical [40], and an appropriate level of trust is essential for successful human-AI interaction [20, 28]. In Psychology, interpersonal trust has been shown to depend on both individual propensity to trust others and perceived trustworthiness of others, which could be further broken down into three attributes: ability, benevolence, and integrity [43]. Lee and See [40] adapted this model to trust in automation and proposed to change the three attributes to performance, process, and purpose. Accordingly, Körber [39] developed a questionnaire to measure trust in automation with subscales accounting for propensity to trust, three attributes of perceived trustworthiness (competence/reliability, understandability/predictability, and intention of developers), familiarity, and a direct measure of trust in automation itself. While trust has been conceptualized as an attitude, which affects behavior probabilistically rather than deterministically [1, 39] and thus is typically measured through subjective self-report methods, the intrusive nature of self-report measures makes it impractical for continuous monitoring purposes [71]. Thus, behavioral markers that are highly associated with subjective measures of trust can be used as an alternative for monitoring purposes without interrupting natural human-AI interaction. For example, eye gaze behavior has been used to infer user trust during autonomous driving [26], where more frequent attention to the system is associated with lower trust. In addition, accuracy of AI and user understanding of AI may be associated with user trust [71], as AI's competence, understandability, and purpose have been identified as the three attributes related to perceived trustworthiness. Our future work will consider how eye gaze behavior, AI accuracy, and user-understanding of AI can predict user trust levels.

3 Conclusion

Since the interdisciplinary endeavour on the scientific study of the human mind later known as Cognitive Science started to take place in 1950s, researchers across multidisciplinary fields have been working together to unravel the information processing mechanisms underlying the black box of human cognition at different levels of analysis and organization. In particular, the development of computational models and machine learning methods in AI continuously brings novel insights into the research on human cognition, especially in the understanding of the representational and computational capacities of the human mind. At the same time, the development of AI research is constantly influenced by the development of theories and methodologies in the study of the human mind. With AI models reaching or surpassing human-level performance in some cognitive tasks, the use of cognitive science theories and methods to understand black-box AI models has never been more important than they are now.

The understanding of the human mind through computational modelling and experimental methods in cognitive psychology facilitates studies on the comparability between human cognition and artificial cognition in AI models. Such comparisons not only can enhance our understanding of both human cognition and AI models, but also play an important role in developing effective XAI methods with the ToM capacities. More specifically, with the ToM capacities, humans tend to assume that AI performs a task in a similar way to them, and thus can naturally update this belief when given information about the differences between humans and AI. In addition, an effective XAI system requires the abilities to infer humans' cognitive model of the task and humans' cognitive model of current understanding of and trust on the AI system in order to provide more human-accessible explanations under new scenarios of the task, which in turn lead to better adoption AI systems in the society. Together these demonstrate well the importance of cognitive science approaches towards XAI, providing a path forward for interdisplinary research.

References

[1] Icek Ajzen. Understanding attitudes and predictiing social behavior. *Englewood cliffs*, 1980.

[2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015.

[3] Arjun R Akula, Keze Wang, Changsong Liu, Sari Saba-Sadiya, Hongjing Lu, Sinisa Todorovic, Joyce Chai, and Song-Chun Zhu. Cx-tom: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *Iscience*, 25(1), 2022.

[4] José P Amorim, Pedro H Abreu, João Santos, Marc Cortes, and Victor Vila. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Information Processing & Management*, 60(2):103225, 2023.

[5] John R Anderson, C Franklin Boyle, Albert T Corbett, and Matthew W Lewis. Cognitive modeling and intelligent tutoring. *Artificial intelligence*, 42(1):7–49, 1990.

[6] Ian A Apperly and Stephen A Butterfill. Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4):953, 2009.

[7] Dana H Ballard, Mary M Hayhoe, Polly K Pook, and Rajesh PN Rao. Deictic codes for the embodiment of cognition. *Behavioral and brain sciences*, 20(4):723–742, 1997.

[8] Andrea Bender. What is causal cognition? *Frontiers in psychology*, 11:3, 2020.

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[10] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. Plan explanations as model reconciliation–an empirical study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 258–266. Ieee, 2019.

[11] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317*, 2017.

[12] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[13] Tim Chuk, Antoni B Chan, and Janet H Hsiao. Understanding eye movements in face recognition using hidden markov models. *Journal of vision*, 14(11):8–8, 2014.

[14] Tim Chuk, Antoni B Chan, Shinsuke Shimojo, and Janet H Hsiao. Eye movement analysis with switching hidden markov models. *Behavior research methods*, 52:1026–1043, 2020.

[15] Tim Chuk, Kate Crookes, William G Hayward, Antoni B Chan, and Janet H Hsiao. Hidden markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures. *Cognition*, 169:102–117, 2017.

[16] Devleena Das, Sonia Chernova, and Been Kim. State2explanation: Concept-based explanations to benefit agent learning and user understanding. *arXiv preprint arXiv:2309.12482*, 2023.

[17] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299:103525, 2021.

[18] HP De Greef and Mark A Neerincx. Cognitive support: Designing aiding to supplement human knowledge. *International Journal of Human-Computer Studies*, 42(5):531–571, 1995.

[19] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[20] Fredrick Ekman, Mikael Johansson, and Jana Sochor. Creating appropriate trust in automated vehicle systems: A framework for hmi design. *IEEE Transactions on Human-Machine Systems*, 48(1):95–101, 2017.

[21] Susan A Gelman, John D Coley, Karl S Rosengren, Erin Hartman, Athina Pappas, and Frank C Keil. Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child development*, pages i–157, 1998.

[22] Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Towards transparent ai systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974*, 2016.

[23] Arushi Gupta, Nikunj Saunshi, Dingli Yu, Kaifeng Lyu, and Sanjeev Arora. New definitions and evaluations for saliency methods: Staying intrinsic, complete and sound. *Advances in Neural Information Processing Systems*, 35:33120–33133, 2022.

[24] Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*, 2020.

[25] John M Henderson. Gaze control as prediction. *Trends in cognitive sciences*, 21(1):15–23, 2017.

[26] Sebastian Hergeth, Lutz Lorenz, Roman Vilimek, and Josef F Krems. Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human factors*, 58(3):509–519, 2016.

[27] Amra Hodzic, Amanda Kaas, Lars Muckli, Aglaja Stirn, and Wolf Singer. Distinct cortical networks for the detection and identification of human body. *Neuroimage*, 45(4):1264–1271, 2009.

[28] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[29] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[30] Janet H Hsiao, Jeehye An, Veronica Kit Sum Hui, Yueyuan Zheng, and Antoni B Chan. Understanding the role of eye movement consistency in face recognition and autism through integrating deep neural networks and hidden markov models. *npj Science of Learning*, 7(1):28, 2022.

[31] Janet H Hsiao, Jeehye An, Yueyuan Zheng, and Antoni B Chan. Do portrait artists have enhanced face processing abilities? evidence from hidden markov modeling of eye movements. *Cognition*, 211:104616, 2021.

[32] Janet H Hsiao, Hui Lan, Yueyuan Zheng, and Antoni B Chan. Eye movement analysis with hidden markov models (emhmm) with co-clustering. *Behavior Research Methods*, 53(6):2473–2486, 2021.

[33] Janet Hui-wen Hsiao and Garrison Cottrell. Two fixations suffice in face recognition. *Psychological science*, 19(10):998–1006, 2008.

[34] Janet Hui-wen Hsiao, Hilary Hei Ting Ngai, Luyu Qiu, Yi Yang, and Caleb Chen Cao. Roadmap of designing cognitive metrics for explainable artificial intelligence (xai). *arXiv preprint arXiv:2108.01737*, 2021.

[35] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. A review of yolo algorithm developments. *Procedia Computer Science*, 199:1066–1073, 2022.

[36] Christopher Kanan, Dina NF Bseiso, Nicholas A Ray, Janet H Hsiao, and Garrison W Cottrell. Humans have idiosyncratic and task-specific scanpaths for judging faces. *Vision research*, 108:67–76, 2015.

[37] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[38] Tae Wan Kim. Explainable artificial intelligence (xai), the goodness criteria and the graspability test. *arXiv preprint arXiv:1810.09598*, 2018.

[39] Moritz Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*, pages 13–30. Springer, 2019.

[40] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[41] Guoyang Liu, Jindi Zhang, Antoni B Chan, and Janet Hsiao. Human attention-guided explainable ai for object detection. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

[42] Guoyang Liu, Jindi Zhang, Antoni B. Chan, and Janet H. Hsiao. Human attention-guided explainable artificial intelligence for computer vision models, 2023.

[43] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

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

[45] Su Mu, Meng Cui, and Xiaodi Huang. Multimodal data fusion in learning analytics: A systematic review. *Sensors*, 20(23):6856, 2020.

[46] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *arXiv preprint arXiv:1902.01876*, 2019.

[47] E Shortcliffe MYCIN. Computer-based medical consultations, 1976.

[48] Joshua New, Leda Cosmides, and John Tooby. Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, 104(42):16598–16603, 2007.

[49] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757*, 2016.

[50] Matthew F Peterson and Miguel P Eckstein. Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological science*, 24(7):1216–1225, 2013.

[51] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[52] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11443–11452, 2021.

[53] Ruoxi Qi, Yueyuan Zheng, Yi Yang, Caleb Chen Cao, and Janet H Hsiao. Explanation strategies for image classification in humans vs. current explainable ai. *arXiv preprint arXiv:2304.04448*, 2023.

[54] Ruoxi Qi, Yueyuan Zheng, Yi Yang, Jindi Zhang, and Janet Hsiao. Individual differences in explanation strategies for image classification and implications for explainable ai. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

[55] Luyu Qiu, Yi Yang, Caleb Chen Cao, Yueyuan Zheng, Hilary Ngai, Janet Hsiao, and Lei Chen. Generating perturbation-based explanations with robustness to out-of-distribution data. In *Proceedings of the ACM Web Conference 2022*, pages 3594–3605, 2022.

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[57] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[58] Natalie Sebanz, Harold Bekkering, and Günther Knoblich. Joint action: bodies and minds moving together. *Trends in cognitive sciences*, 10(2):70–76, 2006.

[59] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[60] Michael J Spivey. Cognitive science progresses toward interactive frameworks. *Topics in Cognitive Science*, 15(2):219–254, 2023.

[61] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301:103558, 2021.

[62] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. Hierarchical expertise level modeling for user specific contrastive explanations. In *IJCAI*, pages 4829–4836, 2018.

[63] William R Swartout and Johanna D Moore. Explanation in second generation expert systems. In *Second generation expert systems*, pages 543–585. Springer, 1993.

[64] J Eric T Taylor and Graham W Taylor. Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, 28(2):454–475, 2021.

[65] Pulkit Verma, Rushang Karia, and Siddharth Srivastava. Autonomous capability assessment of black-box sequential decision-making systems. *arXiv preprint arXiv:2306.04806*, 2023.

[66] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[67] Alice Yang, Guoyang Liu, Yunke Chen, Ruoxi Qi, Jindi Zhang, and Janet Hsiao. Humans vs. ai in detecting vehicles and humans in driving scenarios. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

[68] Scott Cheng-Hsin Yang, Nils Erik Tomas Folke, and Patrick Shafto. A psychological theory of explainability. In *International Conference on Machine Learning*, pages 25007–25021. PMLR, 2022.

[69] Scott Cheng-Hsin Yang, Wai Keen Vong, Ravi B Sojitra, Tomas Folke, and Patrick Shafto. Mitigating belief projection in explainable artificial intelligence via bayesian teaching. *Scientific reports*, 11(1):9863, 2021.

[70] Yi Yang, Yueyuan Zheng, Didan Deng, Jindi Zhang, Yongxiang Huang, Yumeng Yang, Janet H Hsiao, and Caleb Chen Cao. Hsi: Human saliency imitator for benchmarking saliency-based model explanations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 231–242, 2022.

[71] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces*, pages 307–317, 2017.

[72] Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen. Deepred–rule extraction from deep neural networks. In *Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19*, pages 457–473. Springer, 2016.