Human Cognition for Mitigating the Paradox of AI Explainability: A Pilot Study on Human Gaze-based Text Highlighting

Changhyun Lee[a], Hun Yeong Kwon[b], and **Kyung Jin Cha[a]**

[a]Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea

[b]Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea

newdlckdgus@hanyang.ac.kr, khy0@korea.ac.kr, kjcha7@hanynag.ac.kr

Abstract

Artificial Intelligence (AI) explainability plays a crucial role in fostering robust Human-AI Interaction (HAI). However, circular reasoning compromises decision robustness due to limitations in existing AI explainability methods. To address this challenge, we propose leveraging human cognition to enhance explainability, aligning with analysis goals without relying on potentially biased labels. By developing text highlighting driven by human gaze patterns, our research demonstrates that human gaze-based text highlighting significantly reduces decision time for proficient readers, without significantly affecting accuracy or bias. This study concludes by emphasizing the value of human cognition-based explainability in advancing explainable AI (XAI) and HAI.

Introduction

AI explainability is vital for effective human-AI interaction, alongside achieving high performance (London, 2019). Biased data used to train AI models can lead to biased predictions, making explainability crucial for enabling human decision makers to rectify unfair reasoning and bolster confidence in their decisions (Arrieta et al., 2020). Recent research has highlighted situations where prioritizing explainable models with lower accuracy is warranted over black-box models with higher accuracy (Lee & Cha, 2023). However, current AI explainability face inherent limitations, leading to a circular reasoning challenge. While designed to empower human decision makers in rectifying AI bias stemming from biased training data, these methods are themselves based on such biased data, potentially resulting in biased highlighting. This biased form of explainability, with highlighted texts capturing human attention and potentially overshadowing non-highlighted cues, can lead humans to make decisions mirroring the inherent bias in the AI's decision-making process (Ponce and Mayer, 2014). This study underscores the critical need for AI explainability that can fulfill analytical purposes without solely relying on potentially biased labels. To address this, the study aims to determine whether human cognition can serve as an indicator for explainability through field experiment.

Literature Review

The Paradox of AI Explainability

Explainability plays a crucial role in enabling AI to effectively support HAI (Lee and Cha, 2023). While AI was traditionally valued for its performance, recent studies have emphasized that high performance does not inherently guarantee that an AI's decision can be attributed to HAI (London, 2019). Consider a non-explainable AI recruitment system that assigns a high score to a white male applicant. In such cases, it becomes challenging to discern whether the elevated score solely reflects the applicant's competence or if their race or gender influenced the outcome. Therefore, human decision-makers can confidently rely on AI decisions to foster robust HAI only when meticulously crafted explainability mechanisms ensure fairness, accountability, and transparency in the AI's decision-making process (Arrieta et al., 2020; Shin et al., 2020).

However, existing AI explainability face a structural paradox in their efforts to enhance the robustness of HAI. Errors can arise when humans base their final decisions on AI explainability (Wang et al., 2019). Feature importance track features considered crucial for predicting the assigned label, but these tracked features can become unreliable if the label itself is biased (Kusner et al., 2017). The pursuit of explainability to prevent biased decisions relies on labels that may themselves carry bias.

E-Z Reader Model

When individuals view visual content, their eye movements follow distinct patterns characterized by saccades and fixations (Henderson, 2003). Saccades are rapid, short eye movements used to search for important features, while fixations involve brief periods of stationary eyes to extract information from these features (Reichle et al., 2003). Scholars assume that individuals repeat saccades and fixations when perceiving visual information: searching for important parts through saccades, extracting information during fixations, and then performing another saccade to locate important parts (Pannasch et al., 2008).

Recent efforts have integrated of human cognition into explainability, offering several advantages that highlight its potential significance. This approach can enhance the performance of AI explanations (Karim et al., 2022; Yang et al., 2023). Combining the unique attributes of human and AI decisions can offer additional functions (Bertrand et al., 2022). Human cognition-based explainability is considered more plausible compared to traditional AI explainability (Yang et al., 2021; Liu et al., 2023).

Despite these benefits, the attempts to integrate human cognition and explainability are limited due to the concerns related to fairness and plausibility, especially when compared to research scenarios where AI models are designed to make subjective decisions (Meske, 2022). Nevertheless, it becomes increasingly imperative to address the challenge of AI explainability, which is particularly pertinent in cases involving subjective decision-making.

Method and Analysis Results

To construct a text highlighting model based on human gaze, 167 South Korean students participated in a reading task involving short documents. They responded to 10 quizzes to ascertain if a given proposition was addressed in the document. SeeSo, Visual Camp's eye tracking technology, was utilized to track the participants' gaze and record the sequential x and y coordinates of their gaze, along with the duration of fixations in milliseconds.

Based on the eye tracking, a model was developed to identify important text passages deserving of highlighting. This model incorporated hyperparameters such as window size (the number of adjacent words to the fixation), threshold (the number of passages eligible for highlighting), and time limit (the maximum duration considered).

After optimization, 60 Korean college students were assigned to read six documents. For each document, participants had to respond to a set of 10 true/false quizzes. They were randomly divided into control group and treatment group. Initially, both groups read three documents without any highlighting. Then, the former was instructed to read three additional non-highlighted documents, and the latter was directed to read the same documents, but with the optimized model applied.

Difference-in-differences (DID) analysis was performed to assess the effect of human gaze-based text highlighting on decision speed, accuracy, and bias. Decision speed was determined by the time taken by participants to solve the quizzes, decision accuracy was measured by their scores, and decision bias was assessed by their average Krippendorff's alpha (KA). The hypothesis would be supported if the DID coefficient for decision speed is significantly negative, and if decision accuracy and bias show significant positive results. Table 1 presents the analysis results.

Measurement	X_1	X_2	X_3	X_4	coef	t
Decision Speed	255.53	245.62	228.45	201.21	-17.33***	3.40
Decision Accuracy	21.65	23.03	21.35	22.13	-0.61	1.22
KA	0.503	0.575	0.562	0.628	-0.006	-

Table 1. The Results of the DID Analysis

Conclusion

Human gaze-based text highlighting is anticipated to improve decision speed by aligning with human information processing patterns. However, it is important to address certain limitations by integrating the human cognition-based approach with existing AI explainability methodologies within the specific context. By combining the strengths of both approaches, a more reliable AI explainability can be established, promoting effective collaboration between humans and AI systems in decision-making.

Explainability plays a crucial role in facilitating this interaction, especially in subjective decision-making scenarios (Lee and Cha, 2023). However, XAIs often rely on labels, and if these labels are unreliable, the trustworthiness of the XAI may be compromised (Ribeiro et al., 2016). Consequently, reliance on an unreliable XAI system can lead to biased decisions (Ponce and Mayer, 2014). Moreover, when biased HAI occurs, biases can be perpetuated and even amplified during the decision-making process, undermining the integrity of subjective decisions (Ahsen et al., 2019). Human cognition-based explainability offers a potential solution to break free from this circular reasoning dilemma by eliminating the dependence on unreliable labels in XAI. By aligning with human cognition, explainability can enhance the compatibility and effectiveness of HAI.

Acknowledgement

The authors gratefully acknowledge the support from the Visual Camp for this paper.

References

Ahsen, M. E., Ayvaci, M. U. S., & Raghunathan, S. (2019). When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Information Systems Research, 30*(1), 97-116.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58*, 82-115.

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022, July). How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 78-91.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in cognitive sciences, 7*(11), 498-504.

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

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems, 30*.

Lee, C., & Cha, K. (2023). FAT-CAT—Explainability and augmentation for an AI system: A case study on AI recruitment-system adoption. *International Journal of Human-Computer Studies, 171,* 102976.

Liu, G., Zhang, J., Chan, A. B., & Hsiao, J. (2023). Human attention-guided explainable AI for object detection. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. *45*(45). 2573-2580.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report, 49*(1), 15-21.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management, 39*(1), 53-63.

Pannasch, S., Helmert, J. R., Roth, K., Herbold, A. K., & Walter, H. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research, 2*(2), 1-19.

Ponce, H. R., & Mayer, R. E. (2014). An eye movement analysis of highlighting and graphic organizer study aids for learning from expository text. *Computers in Human Behavior, 41*, 21-32.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). Psychology of reading.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*(4), 445-476.

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

Shin, D., Zhong, B., & Biocca, F. A. (2020). Beyond user experience: What constitutes algorithmic experiences?. *International Journal of Information Management, 52*, 102061.

Simon, H.A., (1957). Models of Man - Social and Rational. John Wiley, New York.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1-15.

Yang, A., Liu, G., Chen, Y., Qi, R., Zhang, J., & Hsiao, J. (2023). Humans vs. AI in detecting vehicles and humans in driving scenarios. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. *45*(45). 1832-1839

Yang, Y., Zheng, Y., Deng, D., Zhang, J., Huang, Y., Yang, Y., ... & Cao, C. C. (2022). Hsi: Human saliency imitator for benchmarking saliency-based model explanations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. *10*(1). 231-242.