"Loss in Value": What it revealed about WHO an explanation serves well and WHEN

Md Montaser Hamid[1], Jonathan Dodge[2], Andrew Anderson[1,3] and Margaret Burnett[1]

[1] *Oregon State University, Corvallis, Oregon, 97331, USA*
[2] *Pennsylvania State University, State College, Pennsylvania, 16802, USA*
[3] *IBM Research, San Jose, California, 95120, USA*

Abstract

Evaluating *who* an explanation serves well and *when* is sometimes done empirically using prediction tasks—but measuring a user's success at predicting an AI's behavior in traditional ways can produce misleading empirical results. Koujalgi et al. recently proposed measurements to solve this problem, but they have not yet been tried as a way of answering "who" and "when" questions. In this paper, we show how we used one of these measurements, "Loss in Value", in the context of an XAI study with 69 participants, to learn who our explanations were serving well, who they were not serving well, and when. Our results showed that Loss in Value uncovered "who" and "when" differences that traditional measurements were unable to reveal.

Keywords

Adaptive Explainable AI, Prediction Accuracy, Problem-Solving Style

1. Introduction

Does *this* explanation help *this* user form a reasonably accurate mental model of the AI? Measuring a user's mental model is particularly important for adaptive eXplainable AI (XAI) systems, because their reason for existence is to deliver the right explanation to *this* user in *this* situation [1, 23]. One way to evaluate the effectiveness of such adaptive XAI systems is to evaluate users' mental models—i.e., the user's understanding of how an AI system works [10].

One common approach to measure mental models empirically is through prediction tasks, in which users attempt to predict the AI's next move [2, 8, 18, 21]. However, the common practice of binary framing of prediction tasks—right or wrong—can be misleading because it fails to capture how wrong the user's prediction was.

Fortunately, Koujalgi et al. proposed solutions to this problem through four new prediction measurements that capture degrees of wrongness (i.e., proximity to the correct prediction) [14]. Their empirical results were somewhat encouraging, but were done using only statistical aggregates. That is, their analyses covered entire treatments, but for *adaptive* XAI, a consideration of individual users and individual predictions is needed—to determine *when*, *how*, and for *which* particular users the explanation is helpful, and who is left out.

To help address this gap, in this paper, we present our in-the-trenches use of one of Koujalgi's techniques called "Loss in Value", in an empirical study of two versions of an XAI system in which an AI agent explained itself. We used the technique in a 69-participant study to evaluate users' mental models in a fine-grained way. The study, while not on an adaptive XAI system per se, can be seen as a prerequisite to adaptive XAI systems: it investigates *who* is not helped by our explanations in *which* situations and *why*, producing formative data for the design of future explanations that better serve diverse users, such as by adaptive XAI approaches. Specifically, we measured not only in aggregate, but also prediction-by-prediction and problem-solving style by problem-solving style. This paper presents what doing so revealed, compared with using only the common binary technique.

[1]Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy
hamidmd@oregonstate.edu (M. M. Hamid); jxd6067@psu.edu (J. Dodge); anderan2@oregonstate.edu (A. Anderson); burnett@eecs.oregonstate.edu (M. Burnett)
0000-0002-5701-621X (M. M. Hamid); 0000-0003-4964-6059 (A. Anderson)

2. Background and Related Work

Several researchers have investigated ways to evaluate various kinds of XAI approaches (e.g., [4, 6, 7, 12, 13, 16, 17, 19, 20]), but perhaps the most widely cited is Hoffman et al.'s work [10, 11]. Hoffman et al. recommended the following tasks as useful ways to produce measurements of XAI effectiveness in terms of the accuracy of users' mental models: a prediction task, a retrospection task, a diagramming task, and a self-explanation task. This paper focuses on the first of these, how to measure the outcomes of a prediction task to evaluate users' mental models.

Many empirical XAI researchers have used the prediction task for this purpose, most commonly measuring the outcomes of users' predictions in a binary way—considering a prediction to be either correct or incorrect (e.g., [2, 5, 18, 21, 22]). However, Koujalgi et al. point out that a binary measure of prediction accuracy is very susceptible to floor and ceiling effects, in which almost everyone gets easy predictions correct, but nobody gets any of the others correct [14].

To address this problem, Koujalgi et al. proposed four strategies: Loss in Value, Loss in Rank, Discretized Loss in Rank, and Modified Rank-Biased Overlap between agent's preferences and participants' group-wise preferences [14]. The purpose of these measures is to capture the subtle differences in the degrees of "wrongness" among the incorrect predictions. As Koujalgi et al. point out, besides mitigating the frequent floor/ceiling effects, measuring the subtle differences has an additional advantage: higher resolution. As the output space of an AI grows, so does the ratio of incorrect predictions to correct predictions. At the extremes, the number of incorrect predictions of what the AI will do next (#incorrect) approaches infinity while the number of correct predictions (#correct) remains 1, so the probability of a user making an incorrect prediction (#incorrect / #correct) becomes very high. Without Koujalgi et al.'s measures, nearly 100% would be classified as simply incorrect.

Of these four strategies, "Loss in Value" offers the highest resolution, so we chose "Loss in Value" for our study. This paper offers an in-the-trenches report on how we used this "Loss in Value" technique and what, in combination with other fine-grained information we collected, the technique enabled us to see.

3. How we used the technique

The context of our interest was a comparative study of two versions of some explanations of a game-playing AI agent which explained itself [8]. We wanted to see which version enabled end users to form more accurate mental models of the AI agent. The explanations were set in an MNK game, which is an expanded version of the tic-tac-toe game. Tic-tac-toe has an X-player and an O-player, and in the game, the agent that gave explanations of its behavior was the X-player. Its opposing agent was the O-player.

For that study, we had recruited 69 end users in a between-subject study and had collected data on the participants' understanding of the X-player's behavior while using either the Original or the Post-GenderMag versions of the AI explanations. The Original version included the original design of the explanations by the AI team who created the game; the Post-GenderMag version included the AI team's GenderMag-inspired[2] changes to make the explanations more inclusive to users' potentially wide range of problem-solving styles. 34 participants used the Original version and 35 used the Post-GenderMag version.

Participants filled out questionnaires about their problem-solving styles as per the validated GenderMag survey [9] and additional background information. Participants' background information showed that no participants had any background in AI. While participating in the study, they also answered questions collecting their predictions and comments. Full details of the study design are documented in [8].

We measured users' understanding of the XAI system in two ways for triangulation purposes (using multiple measures on different data to see if they would lead to the same conclusions). One

[2] GenderMag [3] is an inclusive-design method that enables software professionals to improve gender-inclusiveness of their technology by making the technology inclusive across wide ranges of problem-solving style approaches [3, 8, 24].

was a self-explanation task measured via a rubric-based evaluation of participants' written comments about how the AI worked. As this was a measure of their conceptual understanding, we term this measure the "mental model concepts score". The other measure was their predictions of what the AI would do next, i.e., their prediction accuracy. In this paper, we focus on the prediction accuracy score.

To calculate participants' prediction accuracy, we used Koujalgi et al.'s method of calculating "Loss in Value" for each prediction [14]. This method rests on the idea that when an AI agent assigns similar values to two actions, the agent perceives the actions to have similar outcomes. Example: if the agent decides that two moves would be equally good, the agent has to pick one arbitrarily; if the participant picks the other, the participant would still be considered correct (zero loss in value):

$$Loss\ in\ Value\ =\ Value\ estimated\ by\ the\ agent\ for\ the\ agent's\ selected\ action\ -\ Value\ estimated\ by\ the\ agent\ for\ the\ participant's\ predicted\ action$$

We used the absolute value of Koujalgi et al.'s Loss in Value method to calculate our participants' prediction loss per prediction. We termed this loss as PredError which indicates how erroneous a prediction is:

$$PredError\ =\ |Score\ of\ the\ square\ selected\ by\ the\ agent\ (correct\ answer)\ -\ Score\ of\ the\ square\ selected\ by\ the\ participant\ |$$

Where *Score* is the difference between the AI agent's predicted Win% and Loss% for a move in the gameboard. We took the absolute value because PredError can be either positive or negative, but what mattered was how close or far away a participant's prediction was from the agent's selection in any direction.

Our fine-grained analyses of *who* was left out of the benefits of explanations, and *when/how* this occurred, came from these sources:

- Who information: came from each participant's problem-solving styles, as per the GenderMag facets survey [9].
- When information: came from each participant's PredError for each of the 17 prediction tasks across three games.
- Additional context: came from participants' free-text comments, in which they pointed out fine-grained nuances with the explanations' differences.

4. What the technique enabled us to see

4.1. "When" the explanations served well (or not)

In the setting of users making predictions, "when" means which prediction task. As Figure 1: Right shows, the binary measure suggested no differences between treatments at all, except for the very last prediction—but this was very misleading. As the Loss in Value revealed, six tasks actually had visibly noticeable differences between the treatments (Figure 1: Left, bright tasks).

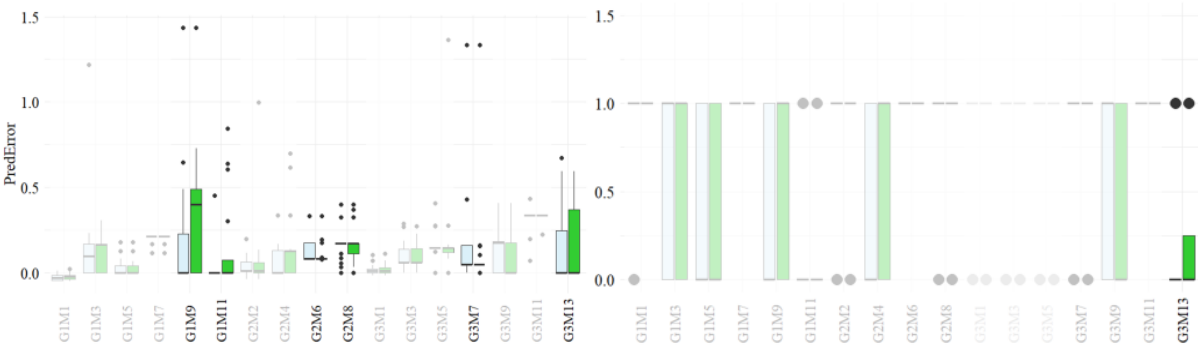

Figure 1: Prediction errors (PredError) for 17 prediction tasks across three games using Loss in Value (Left) and Binary measurements (Right). PredError revealed 6 game moves with visibly noticeable differences—G1M9, G1M11, G2M6, G2M8, G3M7, and G3M13—whereas Binary revealed only G3M13. Lower PredError is better (fewer prediction errors).

Legend:
G=Game, M=Move.

Graphs are boxplots, dots are outliers (paired dots mean multiple outliers), _ is median.
Faded: PredError very similar for Post-GenderMag and Original groups.
Bright: Differences revealed by the measure.

In-depth scrutiny of these six tasks revealed why. Participants' comment data showed that a new explanation present in only the Post-GenderMag treatment, "Top 5 Moves" (Figure 2), held the key. This explanation was popular—over a third of the Post-GenderMag participants explicitly commented on its usefulness. For three of the six tasks, the game remained on a similar trajectory as with the previous move, so predicting one of the previous top five moves was usually a good prediction. But overreliance became a problem if the game trajectory changed markedly with the most recent move, rendering the previous set of top-fives obsolete. This occurred in the other three tasks: Post-GenderMag participants incorrectly chose one of the (previously displayed) top-five choices significantly more often than Original participants, who did not see this explanation. Thus, using Loss in Value revealed a key issue with one of our explanations that was not visible with the Binary measure.

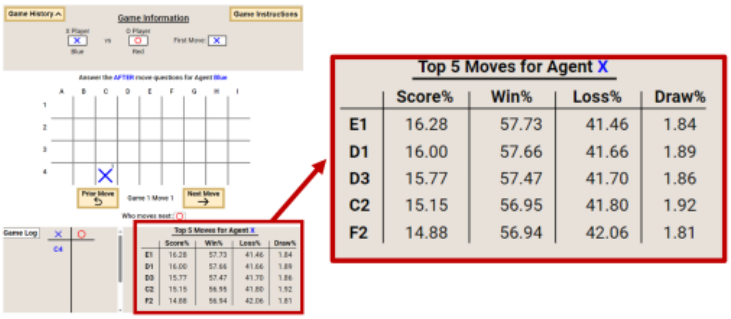

Figure 2: The Post-GenderMag treatment's explanations included the agent's (X-player) scoring details of its top 5 potential moves (red-bordered callout shows at readable size), to explain why the X-player made its most recent move.

4.2. "Who" was left out of the benefits of explanations

The "when" results raise "who" questions: who experienced problems in these 6 tasks? To answer this question, we used participants' problem-solving style data to investigate how participants with more Abi- vs. more Tim-like problem-solving styles[3] fared in each of these 6 tasks.

Here again, Loss in Value revealed the who's that Binary measures often obscured. Specifically, Loss in Value revealed that in all 6 of these 6 prediction tasks, Abi-like participants' PredErrors in the Post-GenderMag version were different than that in the Original version (Figure 3 (Top)-Left). However, for the same tasks, the Binary measure obscured 3 of these 6 (Figure 3 (Top)-Right). For Tim-like participants, Loss in Value revealed differences in 4 of these 6 (Figure 3 (Bottom)-Left), whereas Binary revealed only 2 (Figure 3 (Bottom)-Right).

An Abi-like vs. Tim-like comparison of the Loss in Value data reveals additional inclusivity and equity nuances that combine the who's with the when's. For example, Figure 3 (Top and Bottom Left) shows that Post-GenderMag explanations were harmful to both Abi-like and Tim-like participants in G1M9, with Tim-like participants being especially disadvantaged. An opposite example is G1M11, in which Tim-like participants fared well with both versions, but the Post-GenderMag Abi-like participants were disadvantaged. Contrasting with both of these examples, for G2M6, the Post-GenderMag version worked better for all and G2M8 was a win for Abi-like participants but nothing changed for Tim-like participants. G3M13 was particularly interesting, revealing opposite effects for Abi-like vs. Tim-like participants. As scrutinizing the figures shows, some of these results were not visible with the Binary measure (Figure 3-Top and Bottom Left).

[3] More Abi-like participants have at least 3 of the following 5 problem-solving styles, and Tim-like participants do not: risk-averse in tech settings, lower self-efficacy than peers, task-oriented motivations, comprehensive information processing style, and process-oriented learning styles [9].

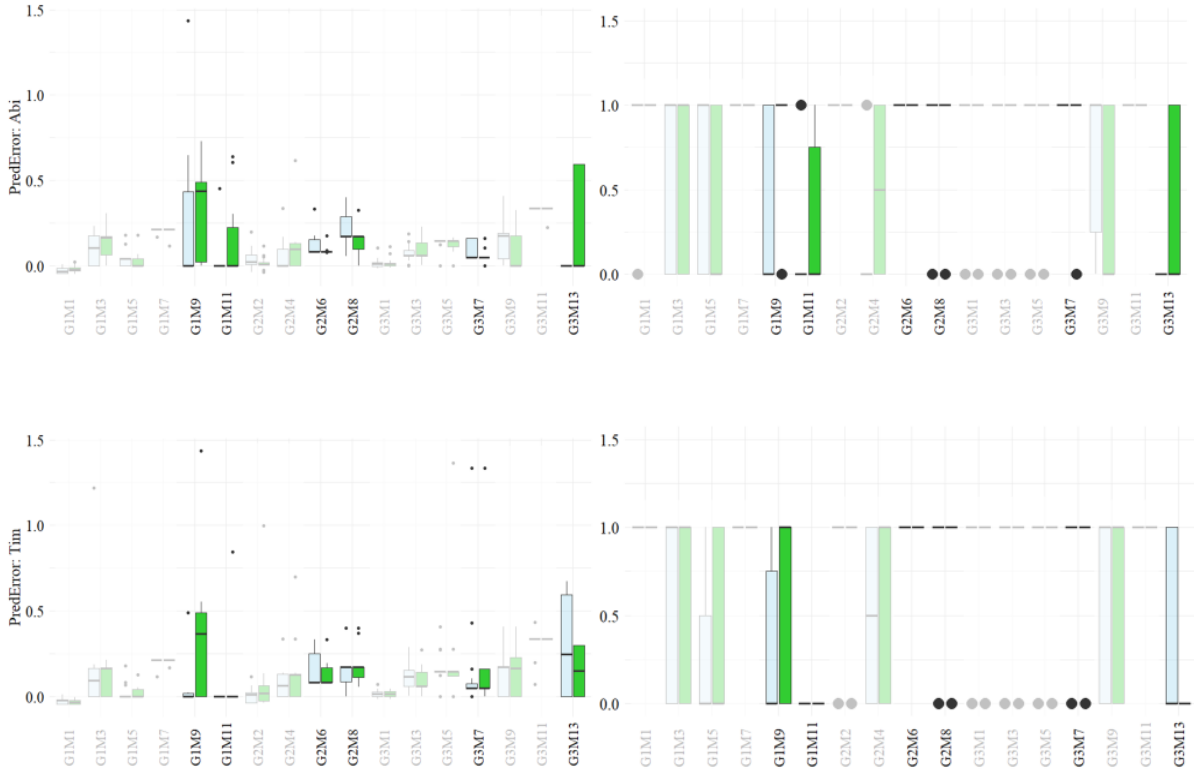

Figure 3: PredErrors for Abi-like and Tim-like participants (lower is better).

(Top): Loss in Value (left) revealed noticeable differences between Post-GenderMag Abi-like participants and Original Abi-like participants in the 6 "when" prediction tasks, but Binary measurement (right) obscured 3 of them.

(Bottom): With Loss in Value (left), Tim-like participants showed 4 noticeable differences in these 6 "when" tasks; Binary (right) obscured 2 of them.

Legend: see Figure 1.

Finally, Loss in Value revealed aggregate inclusivity results that the Binary measure did not. Comparing participants' average PredError by gender, Loss in Value revealed an equity gap: the men's prediction skills were not served as well as the women's in either version. Women had significantly ($p<.05$) better (i.e., fewer) PredErrors in both versions compared to men. These differences were not visible with the Binary measure.

5. Conclusion

In this paper, we presented our in-the-trenches experience with Koujalgi et al.'s Loss in Value method to evaluate users' mental models via their ability to predict an AI agent's next move. How to measure users' prediction successes matters, because prediction success is a common technique for evaluating users' mental models, which is a core reason for XAI.

In our experience, Loss in Value revealed numerous fine-grained results of the who, when, and how that binary measures obscured. Our study's setting was inclusive XAI rather than adaptive XAI, but shares with adaptive XAI a need to know which *particular* users (who) are well-served by an explanation vs. not, and when/how these phenomena occurred for whom. Given this commonality, our experiences provide encouraging evidence that the Loss in Value method is a more accurate and more useful measure than the traditional binary method of measuring the who's, when's, and how's of XAI work aiming to better serve the widely diverse users who deserve an explanation that is actually useful to them.

Acknowledgements

This work was supported in part by the USDA National Institute of Food and Agriculture #2021-67021-35344 and NSF #1901031 and #2042324

References

[1] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence," in Computer, vol. 53, no. 8, pp. 18-28, Aug. 2020, doi: 10.1109/MC.2020.2996587.

[2] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental models of mere mortals with explanations of reinforcement learning. ACM Transactions on Interactive Intelligent Systems (TiiS) 10, 2 (2020), 1–37.

[3] Burnett, Margaret, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. "GenderMag: A method for evaluating software's gender inclusiveness." *Interacting with Computers* 28(6). 2016: 760-787.

[4] Coroama, Loredana, and Adrian Groza. "Evaluation metrics in explainable artificial intelligence (XAI)." In International conference on advanced research in technologies, information, innovation and sustainability, pp. 401-413. Cham: Springer Nature Switzerland, 2022.

[5] Xinyue Dai, Mark T. Keane, Laurence Shalloo, Elodie Ruelle, and Ruth MJ Byrne. 2022. Counterfactual explanations for prediction and diagnosis in XAI. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. 215–226.

[6] Davis, Brittany, Maria Glenski, William Sealy, and Dustin Arendt. "Measure utility, gain trust: practical advice for XAI researchers." In 2020 IEEE workshop on trust and expertise in visual analytics (TREX), pp. 1-8. IEEE, 2020.

[7] Dodge, Jonathan, Andrew A. Anderson, Matthew Olson, Rupika Dikkala, and Margaret Burnett. "How do people rank multiple mutant agents?." In Proceedings of the 27th International Conference on Intelligent User Interfaces, pp. 191-211. 2022.

[8] Hamid, Md Montaser, Fatima Moussaoui, Jimena Noa Guevara, Andrew Anderson, and Margaret Burnett. "Inclusive Design of AI's Explanations: Just for Those Previously Left Out, or for Everyone?" *arXiv preprint arXiv:2404.13217* (2024).

[9] Hamid, Md Montaser, Amreeta Chatterjee, Mariam Guizani, Andrew Anderson, Fatima Moussaoui, Sarah Yang, Isaac Escobar, Anita Sarma, and Margaret Burnett. "How to measure diversity actionably in technology." In *Equity, Diversity, and Inclusion in Software Engineering: Best Practices and Insights*, pp. 469-485. Berkeley, CA: Apress, 2024.

[10] Hoffman, Robert R., Shane T. Mueller, Gary Klein, and Jordan Litman. "Metrics for explainable AI: Challenges and prospects." arXiv preprint arXiv:1812.04608 (2018).

[11] Hoffman, Robert R., Shane T. Mueller, Gary Klein, and Jordan Litman. "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance." Frontiers in Computer Science 5 (2023): 1096257.

[12] Kim, Jenia, Henry Maathuis, and Danielle Sent. "Human-centered evaluation of explainable AI applications: a systematic review." Frontiers in Artificial Intelligence 7 (2024): 1456486.

[13] Kong, Xiangwei, Shujie Liu, and Luhao Zhu. "Toward Human-centered XAI in Practice: A survey." Machine Intelligence Research (2024): 740-770.

[14] Sujay Koujalgi, Andrew Anderson, Iyadunni Adenuga, Shikha Soneji, Rupika Dikkala, Teresita Guzman Nader, Leo Soccio, Sourav Panda, Rupak Kumar Das, Margaret Burnett, and Jonathan Dodge. 2024. How to Measure Human-AI Prediction Accuracy in Explainable AI Systems. arXiv:2409.00069 [cs.HC] https://arxiv.org/abs/2409.00069

[15] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In 2013 IEEE Symposium on Visual Languages and Human Centric Computing. IEEE, 3–10.

[16] Lopes, Pedro, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. "XAI systems evaluation: A review of human and computer-centred methods." Applied Sciences 12, no. 19 (2022): 9423.
[17] Naveed, Sidra, Gunnar Stevens, and Dean Robin-Kern. "An Overview of the Empirical Evaluation of Explainable AI (XAI): A Comprehensive Guideline for User-Centered Evaluation in XAI." Applied Sciences 14, no. 23 (2024): 11288.
[18] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In 26th International Conference on Intelligent User Interfaces. 340–350.
[19] Rong, Yao, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. "Towards human-centered explainable AI: user studies for model explanations." Machine Intelligence vol. 46, (2022): 2104–2122.
[20] Schmidt, Philipp, and Felix Biessmann. "Quantifying interpretability and trust in machine learning systems." arXiv preprint arXiv:1901.08558 (2019).
[21] Schraagen, Jan Maarten, Pia Elsasser, Hanna Fricke, Marleen Hof, and Fabyen Ragalmuto. "Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 64, no. 1, pp. 339-343. Sage CA: Los Angeles, CA: SAGE Publications, 2020.
[22] Teerachart Soratana, X. Jessie Yang, and Yili Liu. 2021. Human Prediction of Robot's Intention in Object Handling Tasks.In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 65. SAGE Publications Sage CA: Los Angeles, CA, 1190–1194.
[23] Robert Thomson, Jordan Richard Schoenherr. (2020). Knowledge-to-Information Translation Training (KITT): An Adaptive Approach to Explainable Artificial Intelligence. In: Sottilare, R.A., Schwarz, J. (eds) Adaptive Instructional Systems. HCII 2020. Lecture Notes in Computer Science(), vol 12214. Springer, Cham. https://doi.org/10.1007/978-3-030-50788-6_14
[24] Vorvoreanu, Mihaela, Lingyi Zhang, Yun-Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. 2019. "From Gender Biases to Gender-Inclusive Design: An Empirical Investigation". In ACM Conference on Human Factors in Computing Systems (CHI '19). Glasgow, Scotland, UK, 1–14. https://doi.org/10.1145/3290605.3300283