Artificial Intelligence and Social Computing, Vol. 72, 2023, 121–130
https://doi.org/10.54941/ahfe1003280

Evaluating the Effect of Time on Trust Calibration of Explainable Artificial Intelligence

Ezekiel Bernardo and Rosemary Seva

Industrial and Systems Engineering Department, De La Salle University, Manila, Philippines

ABSTRACT

Explainable Artificial Intelligence (XAI) has played a significant role in human-computer interaction. The cognitive resources it carries allow humans to understand the complex algorithm powering Artificial Intelligence (AI), virtually resolving the acceptance and adoption barrier from the lack of transparency. This resulted in more systems leveraging XAI and triggering interest and efforts to develop newer and more capable techniques. However, though the research stream is expanding, little is known about the extent of its effectiveness on end-users. Current works have only measured XAI effects on either moment time effect or compared it cross-sectionally on various types of users. Filling this out can improve the understanding of existing studies and provide practical limitations on its use for trust calibration. To address this gap, a multi-time research experiment was conducted with 103 participants to use and evaluate XAI in an image classification application for three days. Measurement that was considered is on perceived usefulness for its cognitive contribution, integral emotions for affective change, trust, and reliance, and was analyzed via covariance-based structural equation modelling. Results showed that time only moderates the path from cognitive to trust and reliance as well as trust to reliance, with its effect dampening through time. On the other hand, affective change has remained consistent in all interactions. This shows that if an AI system uses XAI over a longer time frame, prioritization should be on its affective properties (i.e., things that will trigger emotional change) rather than purely on its cognitive purpose to maximize the positive effect of XAI.

Keywords: Explainable AI, XAI, Artificial intelligence, AI, Trust, Affect, Time, Moderation, Structural equation modelling, SEM

INTRODUCTION

Explainable AI (XAI) has become a potent tool in resolving the interpretability issue of AI. It allowed users to understand the complex inner workings of AI by providing human-level explanations (Barredo Arrieta et al., 2020; Das and Rad, 2020), which has been previously unachievable due to the convolution of machine learning and deep learning algorithms (Chowdhary, 2020). As such, it allowed users to calibrate their mental model and subsequently attune a proper stance for their trust and reliance on the system (Gunning et al., 2019).

Given the benefits it offers, many researchers have tried to improve the understanding of XAI. In the recent systematic review on the field, the progress had been on the development of newer techniques, research that focused on understanding how it calibrates trust, and newer perspectives that dive into the analysis of variables that can play in for XAI's utility (Adadi and Berrada, 2018; Forster et al., 2017; Haque et al., 2023; Rudin and Radin, 2019; Shin, 2021). This gave a massive push on the field, by expanding how it should be developed, integrated, and improved, which is heavily needed considering the trajectory of society's view on the role of AI (Lewis et al., 2021). However, there had been a key aspect of its use that has been left unanswered. Particularly, this is on the extent to which XAI is effective in resolving transparency.

The question of the extent of the effectiveness of XAI can be viewed from a multitude of perspectives. This can be for the inherent factors of XAI (e.g., design, components, features) or externally that can be induced in the human-XAI interaction (e.g., user's characteristics and disposition) (Ashoori and Weisz, 2019; Guerdan et al., 2021; Mohseni et al., 2021). However, one factor that directly pierces the user and XAI is the element of time. Operationally, this is on how long an XAI is effective and how time changes the user's stance on it.

Scholars have previously studied that time or experience may moderate trust. Kok & Soh (2020) identified that trust is not a static phenomenon and may dynamically change as the interaction unfolds. Another is the meta-analysis of Vanneste et al. (2014) where they stipulated the varying influential rate and the total effect of different mechanisms to build trust over time. Holliday et al. (2016) found out that there are different trust calibration journeys and explanations that alter such behaviors. Building on this, the time element for XAI can be explored to provide practical limitations on when and how should it be leveraged.

Considering the established arguments from the earlier discussion, this study was proposed to look at extending the XAI Affective Trust and Reliance (XATR) Calibration Model, with the consideration of time as a moderator. This is considered to viably know the limitations of the previous work of Bernardo & Seva (2023) as it is tested on a limited time frame and to create a suitable recommendation on a longer view of using XAI to calibrate trust for AI systems. With that, as presented in Figure 1, the following hypothesis was proposed:

***H1**: Time moderates the affective effect of XAI on (a) trust and (b) reliance*
***H2**: Time moderates the cognitive effect of XAI on (a) trust and (b) reliance*
***H3**: Time moderates the effect of trust on reliance.*

METHODOLOGY

To viably test the hypothesis proposed in the study, an asynchronous virtual experiment was conducted. The design was contextually lifted from the previous work of the research group (see Bernardo & Seva (2023)) with it using the same measurement tools to capture both the independent (i.e.,

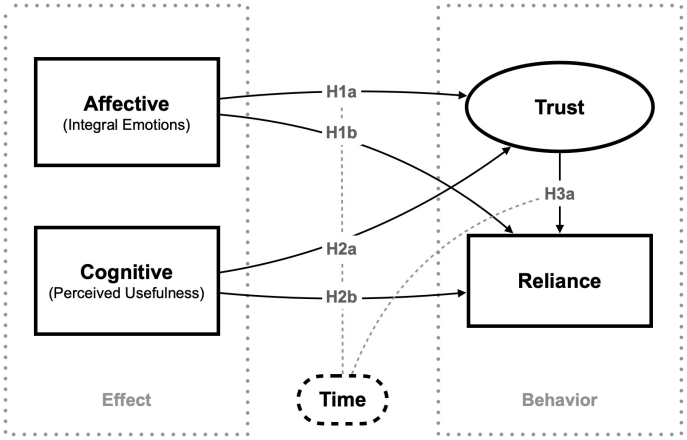

Figure 1: Time moderation on the XATR calibration model.

demographical data and disposition stance) and dependent variables (i.e., perceived usefulness for cognitive, integral emotion, trust, reliance).

Participants

Data was collected through convenience sampling method. The minimum target was set at 100 participants following the recommendation for detecting significant effects for a model structure by Westland (2010) and Cohen (1988) in their priori power computation. Preliminary leads were generated through direct invitations of the researcher's peers, which were supplemented by social networking ads. Qualifications were set as being at least 17 years old, being able to communicate in English, having used any AI-powered system in the recent year, and having normal to correct-to-normal vision. Considering the time element of the study, participants were also required to have at least three days to experiment with a minimum of 30 minutes of availability per day. As an incentive for participation, a reward of 150 PHP (∼3.00 USD) with a performance bonus of up to 100 PHP (∼2.00 USD) was guaranteed upon successful participation.

Tools

As mentioned earlier, the pre-experiment questionnaire and the XAI testbed of Bernardo & Seva (2023) were contextually used for the data gathering. This was decided considering the high-reliability score of the tools and to have a parallel view for comparative analysis. In terms of purpose, the former inquired about the consent, provided screening questions, and captured the demographical data. On the other hand, the latter focused on the presentation of XAI and capturing the main response variables resolved from the XATR calibration model in a seven-point unipolar slider (1 - strongly agree, 7 - strongly disagree). For reliability assurance, construct validity was retested using sample data from 20 participants, which includes AI programmers, app developers, user experience experts, and end-users.

Procedure

The experiment was divided into three phases: preliminaries, main experiment, and assessment. In terms of goal, the first phase aimed to onboard the users, disclosed data usage, and level the expectation with the experiment. For the second, the objective was to gather all measurements needed to test the hypothesis. Lastly, the third focused on the evaluation of experiment performance, post-interview, and token distribution.

As for the journey, after confirming to participate, subjects were required to attend a synchronous online meeting for the onboarding. Once done, confirmed subjects received the access link for the asynchronous experiment. Upon accessing, they were first directed to the pre-experiment questionnaire. The initial part was the consent clause confirmation and screening test. Only those who agreed and passed were allowed to continue to answer the demographic section. The priming condition then followed. The case of the study was that the subject was hired by an organization focused on creating a database of species seen all over the Philippines and subsequently providing a recognition AI to help common end-users if they wish to classify any species they encounter. To aid, the subjects were hired for three days to classify species sent to their system. The AI provides its recognition but the participants can input their own if deemed necessary. Once the participants accepted that they understand the case, they were then allowed to open the XAI testbed.

The general instructions and the three practice recognition were initially prompted to the participants. Each participant was required to experiment for three consecutive days, preferably at the same time of the day. Per day, they need to recognize 25 random photos provided in the application. On each trial, they were asked to provide their evaluation of the usefulness, emotions felt, and trust. To be able to record the progress of each participant, a unique reference code was generated by the testbed, which was required to be sent to the researchers after each experiment day. At the end of the third-day experiment, they were asked if they are willing to join the post-experiment interview and instructed on how the incentives were to be distributed. Those who agreed were provided with available schedules for the interview.

Data Recording and Analysis

The main difference between the tools from the work of Bernardo & Seva (2023) was that four emotions of XES were aggregated into one to reflect integral emotions, observing the natural valence of the emotions. Next, to observe the time element featured in the study, rating-based data was recorded based on its aggregate average value from the initial time to the momentary point of consideration. This follows the area under the curve (AUTC) recommendation of Yang et al. (2017) in their time analysis of trust. Lastly, for reliance, being the only dichotomous data, point observation (i.e., Yes or No) was recorded.

In terms of the analysis, the methodology was anchored on covariance-based structural equation modelling (CB-SEM). This was used considering its ability to do simultaneous analysis for multiple relationships, can incorporate time series analysis, insensitivity to hard parametric conditions, and

applicability for comparative analysis (Astrachan et al., 2014; Dash and Paul, 2021), which are all the major requirements to test the objectives. The 23rd version of the IBM statistical package for Social Science (SPSS) and Analysis of a Moment Structure (AMOS) was used for the computation and theory testing, with 0.05 alpha level being the significance limit. For brevity, the acceptance threshold for the other preliminary statistical test (e.g., fit, validity, reliability) will be given in its corresponding discussion.

RESULTS & ANALYSIS

The experiment ran successfully without any concerns raised or operating issues. It took 15 days for the entire data to be gathered, at an average usage time of 18 minutes per day, with access mostly happening from 5:00 PM to 12:00 PM. There was a decrease in usage time (SD = 7 minutes) from the first and third days. As for the post-interview, 31% of the subjects participated, with 10 minutes being its mean duration.

Data Screening

103 out of 123 subjects was the usable data sourced from the experiment. Abandonment (e.g., only participated for 2 days) was the main reason for removal, with some due to failing the screening requirements. Structurally, the majority of the participants were male (60.19%), belonging to the millennial age group (68.93%), and college graduates (83.50%). For AI-related demographics, most are (74.76%) innovators or have at least 5 years of experience, have used an image recognition AI before (60.19%), and have interacted with an XAI (71.84%).

Latent Variables Assessment

The factor analysis revealed that the data was valid and reliable for the measurements needed. This was proven by the Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity ($p < 0.001$), which were meritorious (0.833) and significant ($p < 0.001$). Also, all of the extractions were above the 0.700 thresholds showing that the contributed communalities were all relevant. As for the groupings, the three distinct latent variables proposed (i.e., cognitive, affective, and trust) were captured at a 72.883% variance explained, with a minimum factor loading of 0.621, and a Cronbach alpha of 0.753, 0.882, and 0.781 respectively. Confirmatory factor analysis via common latent bias also agrees with the findings. At an excellent fit ($\chi 2$/df - 1.421, RMSEA - 0.070, CFI - 0.951), all composite reliability scores were above 0.752, and all positive square root average variance extracted (AVE) was higher than the correlation amongst other latent variables.

Unconstrained Path and Relationship Analysis

A significant recursive structural model was determined from the 1000 bootstrapped iterations run. Notably, all solutions were resolved, estimates passed the three types of model fit (see Table 1), and no modification indices surfaced. As for the relationships tested, all of the proposed paths were statistically supported for the unconstrained model as shown in table 2.

Table 1. Model fit measures and threshold.

Type	Indices	Estimate	Threshold	Reference
Absolute Fit	RMSEA	0.075	< 0.08	Westland (Astrachan et al., 2014; Dash and Paul, 2021)
	SRMR	0.039	< 0.08	Hu & Bentler (1999)
Incremental Fit	CFI	0.984	> 0.95	Schreiber et al. (2006)
	NFI	0.965	> 0.95	Hu & Bentler (1999)
Parsimonious Fit	$\chi 2$/df	1.802	1 to 3	Hu & Bentler (1999)

Note: RMSEA - Root Mean Square Error of Approximation; SRMR - Standardized Root Mean Square Residual; CFI - Comparative Fit Index; NFI - Normed Fit Index; $\chi 2$/df - Chi-squared per Degrees of Freedom

Table 2. Direct path analysis results.

From		To	Std. Est. (ß)	P	Remarks
Integral Emotion	→	Trust	0.511	***	Supported
Perceived Usefulness			0.358	0.032	Supported
Integral Emotion	→	Reliance	0.385	0.021	Supported
Perceived Usefulness			0.379	***	Supported
Trust			0.520	***	Supported

*Notes: Std.Est. - Standard Estimates; P - significance; *** p-value < 0.01*

Table 3. Global multi-group moderation results and fit.

Type	CMIN	P	NFI	IFI	RFI	TLI	Remarks[a]
Time	39.758	<0.000	0.932	0.898	0.934	0.911	Moderated

Note: CMIN - Chi-square statistics; P - Significance; NFI - Normal Fit Index; IFI - Incremental Fit Index; RFI - Relative Fit Index; TLI - Tucker-Lewis coefficient; [a] Evaluated at p-value < 0.05

Moderating Analysis

The global multi-group moderation identified that time significantly moderates the proposed model. As shown in table 3, a highly significant interaction was determined at an excellent model fit all approaching 1.0. Looking at the local moderation difference in table 4, only the path for perceived usefulness to trust (H2a), reliance (H2b), and trust to reliance (H3) were statistically supported. These paths, based on estimates, weaken through time. On the other hand, all paths from integral emotion do not have significantly different results for the first and third days results. The consistency means that subjects reported the same degree and valance for all testing days.

DISCUSSION

Overall, results have highlighted that time was a moderator in the effectiveness of XAI. However, based on the specific local runs, it was determined that moderation was only partial, with only the cognitive route being affected. Specifically, both trust (decrease of 1.748) and reliance (decrease of 0.698) dampens through time. As for the affective route, the effect remained

Table 4. Local moderation difference.

From		To	*First-Day*		*Third-Day*		z-score	P
			Est.	P	Est.	P		
Integ. Emo	→	Trust	1.657	0.000	1.806	0.000	0.447	0.327
Perc. Use			2.429	***	0.681	0.000	4.541	***
Integ. Emo	→	Reliance	0.742	0.455	0.595	0.201	-0.134	0.447
Perc. Use			2.270	0.000	1.572	0.000	2.123	0.017
Trust			0.575	0.233	0.451	0.000	1.678	0.046[b]

*Notes: Est. - Unstandardized estimates, z -z-score; P - Significance; *** p-value < 0.01; b partial difference due to insignificance of first trial*

statistically consistent for all interactions, although there were slight changes experienced (i.e., increase of 0.149 for trust and decrease of 0.147 for reliance). To illustrate, Figure 1 shows the downward and consistent trend via fitted regression lines for cognitive and affective respectively.

The trend of its trust effect has resonated from the interviews conducted. As expressed by the subjects, a different stance was observed for the three days of the experiment. For the first day (trials 1-25), XAI primarily functions as a cognitive resource wherein subjects scrutinize the XAI to understand the limitations of the system. Subjects view the design less and focus more on the information it provides. On the second day (trials 26-50), familiarity played and subjects started to appreciate the design more as compared to the cognitive component of XAI. With the understanding of how the explanations were laid out, they value the presentation more which affects the integral emotions of the subjects. Lastly, on the third day (trials 51-75), the cognitive contribution plays lesser, with most of the subjects viewing the XAI as more of a cue rather than an information resource. This is evident by the faster experiment time compared to the first day (4 minutes faster). More so, subjects also mentioned that because they already have a bar on performance limitations of the AI, the explanations were evaluated on the confidence they felt upon viewing it.

Two theories can be said that heavily explained and supplements the findings. First is the elaboration likelihood model (ELM) of persuasion by Petty & Cacioppo (1986). In this, two types of processing can be done on a stimulus. This can follow a central route where processing is from scrutinizing the information presented or a peripheral route that values previous interaction and understanding to create a rapid evaluation. In the context of the study, the central route has been evident on the first days and further transitioned to the peripheral once the necessary information has been reserved and understood by the subjects. Adding to this is the affect infusion theory of Forgas (1995) where emotions alter the appraisal of stimuli and their behavioral effect. When the participants view XAI as a cue, valued emotion for their evaluation.

Another finding is that trust's effect on reliance significantly decreases over time. Based on the interview, this can be interpreted that other factors such as utility, performance, and hedonic were the factors they value as more

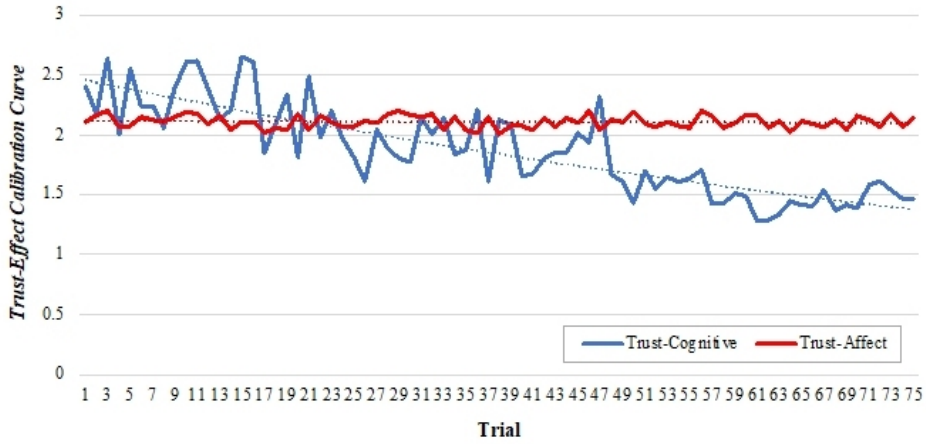

Figure 2: Slope trend of the trust-effect calibration curve.

interaction happened. Adding to the initial discussion, it can be seen that current cognitive prioritization will have a compounding effect on reliance, which is not ideal for propagating the use of AI. Thus, it is recommended to focus more on an affective route with a longer tenure of use.

CONCLUSION

Overall, the study was successful in uncovering the effects of time on XAI effectiveness. Being the first to test this, the study has highlighted three key findings. First, time functions as a moderator but is only relevant for cognitive route. Second, the effect of trust on reliance weakens through time. Lastly, it is recommended that the affective route should be prioritized to lessen the compounding dampening effect of trust to reliance. These results supplement current research on XAI trust calibration, highlighting that the current idea on the effectiveness of XAI should be retested and extended for its applicability in a longer time frame. More so, this opens new ideas how to strategize the use and improvement of XAI.

ACKNOWLEDGMENT

The Department of Science and Technology funded this research under the Engineering Research and Development for Technology grant.

REFERENCES

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ashoori, M., Weisz, J. D., 2019. In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. arXiv:1912.02675 [cs].

Astrachan, C. B., Patel, V. K., Wanzenried, G., 2014. A comparative study of CB-SEM and PLS-SEM for theory development in family firm research. Journal of Family Business Strategy 5, 116–128. https://doi.org/10.1016/j.jfbs.2013.12.002

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bernardo, E., Seva, R., 2023. Affective Analysis of Explainable Artificial Intelligence in the Development of Trust in AI Systems. Manuscript submitted for publication.

Chowdhary, K. R., 2020. Fundamentals of artificial intelligence. Springer, New Delhi.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences.

Das, A., Rad, P., 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. arXiv:2006.11371 [cs].

Dash, G., Paul, J., 2021. CB-SEM vs PLS-SEM methods for research in social sciences and technology forecasting. Technological Forecasting and Social Change 173, 121092. https://doi.org/10.1016/j.techfore.2021.121092

Forgas, J. P., 1995. Mood and judgment: The affect infusion model (AIM). Psychological Bulletin 117, 39–66. https://doi.org/10.1037/0033-2909.117.1.39

Forster, Y., Naujoks, F., Neukum, A., 2017. Increasing anthropomorphism and trust in automated driving functions by adding speech output, in: 2017 IEEE Intelligent Vehicles Symposium (IV). Presented at the 2017 IEEE Intelligent Vehicles Symposium (IV), IEEE, Los Angeles, CA, USA, pp. 365–372. https://doi.org/10.1109/IVS.2017.7995746

Guerdan, L., Raymond, A., Gunes, H., 2021. Toward Affective XAI: Facial Affect Analysis for Understanding Explainable Human-AI Interactions 10.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z., 2019. XAI—Explainable artificial intelligence. Sci. Robot. 4, eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Haque, A. B., Islam, A. K. M. N., Mikalef, P., 2023. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technological Forecasting and Social Change 186, 122120. https://doi.org/10.1016/j.techfore.2022.122120

Holliday, D., Wilson, S., Stumpf, S., 2016. User Trust in Intelligent Systems: A Journey Over Time, in: Proceedings of the 21st International Conference on Intelligent User Interfaces. Presented at the IUI'16: 21st International Conference on Intelligent User Interfaces, ACM, Sonoma California USA, pp. 164–168. https://doi.org/10.1145/2856767.2856811

Hu, L., Bentler, P. M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal 6, 1–55. https://doi.org/10.1080/10705519909540118

Kok, B. C., Soh, H., 2020. Trust in Robots: Challenges and Opportunities. Curr Robot Rep 1, 297–309. https://doi.org/10.1007/s43154-020-00029-y

Lewis, M., Li, H., Sycara, K., 2021. Deep learning, transparency, and trust in human robot teamwork, in: Trust in Human-Robot Interaction. Elsevier, pp. 321–352. https://doi.org/10.1016/B978-0-12-819472-0.00014-9

Mohseni, S., Zarei, N., Ragan, E. D., 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans. Interact. Intell. Syst. 11, 1–45. https://doi.org/10.1145/3387166

Petty, R. E., Cacioppo, J. T., 1986. The Elaboration Likelihood Model of Persuasion, in: Advances in Experimental Social Psychology. Elsevier, pp. 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

Rudin, C., Radin, J., 2019. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review 1. https://doi.org/10.1162/99608f92.5a8a3a3d

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., King, J., 2006. Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. The Journal of Educational Research 99, 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Shin, D., 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Vanneste, B. S., Puranam, P., Kretschmer, T., 2014. Trust over time in exchange relationships: Meta-analysis and theory: Research Notes and Commentaries. Strat. Mgmt. J. 35, 1891–1902. https://doi.org/10.1002/smj.2198

Westland, C., 2010. Lower bounds on sample size in structural equation modeling. Electronic Commerce Research and Applications 9, 476–487. https://doi.org/10.1016/j.elerap.2010.07.003

Yang, X. J., Unhelkar, V. V., Li, K., Shah, J. A., 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation, in: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. Presented at the HRI'17: ACM/IEEE International Conference on Human-Robot Interaction, ACM, Vienna Austria, pp. 408–416. https://doi.org/10.1145/2909824.3020230