Human Systems Engineering and Design (IHSED2025), Vol. 198, 2025, 197–206
https://doi.org/10.54941/ahfe1006776

A Predictive Model of Human Trust Evolution Over Time in AI-Based Recommendations

Vincent Fer[1,3], Daniel Lafond[2], Gilles Coppin[3], and Olivier Grisvard[1]

[1]Thales Defence Mission Systems, Brest, 29200, France
[2]Thales CortAIx Labs, Quebec, Canada
[3]IMT Atlantique, Lab-STICC UMR 6285, CNRS, France

ABSTRACT

Understanding the dynamics of human trust in AI-based recommendations is an important challenge for the design of decision support systems and human-machine teams. This study aims to advance quantitative modeling of reported trust levels evolving over successive trials across several weeks. Data from 53 participants was collected in a visual search experiment with AI system recommendations. The Feedback-based Dynamic Trust Model (F-DTM) proposed herein is based on ten predictors, focusing on variables linked to different types of delayed feedback. Six different types of machine learning regression models were compared, with the decision tree model demonstrating the best predictive performance ($R^2 = 64.9\%$, $RMSE = 0.72$) on held-out data. Some variables were then converted into cumulative sums to capture more effectively the sequential nature of the data with a memory of past outcomes. These modifications significantly improved the performance of the new 12-variable decision tree model ($R^2 = 69.92\%$, $RMSE = 0.66$). A subsequent analysis on this revised F-DTM model assessed the impact of eliminating variables one at a time, reaching an R^2 of 70.41% and an RMSE of 0.65. These findings help address the current lack of quantitative models of trust evolution in AI. However, the present cumulative sum memory approach of the F-DTM, employing supervised machine learning, may be improved on by using more complex models designed for time-series forecasting. Directions for future research include investigating temporal models, such as long short-term memory (LSTM), hidden Markov model (HMM), ARIMA or autoregressive models to predict trust evolution in AI-based recommendations.

Keywords: Trust, Artificial intelligence, Computational model, Human-autonomy teaming, Repeated measures, Regression model

INTRODUCTION

The topic of trust in artificial intelligence has been gaining interest for several decades, particularly in highly critical fields such as medicine, automotive, aviation, and military domains. In these contexts, we want to avoid individuals' over- or under-reliance in AI (Parasuraman, 1997), which could lead to serious incidents. This over- or under-reliance is also linked to good or poor use of the system. There can be many causes for these states, such as a lack of training, a high workload, or underlying miscalibrated trust in the system (Parasuraman, 1997). We therefore may talk to a similar extent about over- or under-trust (De Visser, 2018). In the case of over-trust, the individual has "too much" trust in the capabilities of AI ("too much" being considered in reference to the objective performance of the AI system observed on a reference set of situations), while under-trust is the opposite, where the individual does "not" have "enough" trust in the capabilities of AI. The main challenge of this approach is to determine an adequate level of trust that an individual should have in AI which is quite complex because of the complexity of concept of trust itself, as described below.

Trust in AI

Lee and See (2004) define trust as "the attitude that a trustor has toward a trustee who will help them achieve an objective in a risky and uncertain environment". According to this definition, trust is structured along a cycle of four stages: belief, attitude, intention, and behavior. Belief is the psychological construct that qualifies the expected assisting outcomes of the trustee's. Attitude is the inclination of the trustor to rely on the trustee in a certain context. Intention is defined as the willingness to act or behave in a certain manner, while taking into account all other factors influencing the decision (such as workload, stress, fatigue, perceived risk, etc. (Parasuraman, 1997). The behavior is the way the intention is made observable through an action which can be reliant (the user relies on the AI to conclude his/her choice) as well as meaningful parameters related to this action (decision time, etc.). To close the cycle, depending on the result, this action will ultimately have an effect on the trustor's previous beliefs about the trustee. To check the effect of results on the four states of the trust cycle, various types of measures exist (Kohn, 2021). Declarative measures that focus on beliefs and attitudes can be used at discrete times, before and after a set of interactions. Physiological measures (cardiac/respiratory/cerebral activity, skin conductance, etc.) focus on intention and can be used during interactions (de Visser, 2018), and finally behavioral measures (reliance, compliance, response time, etc.) provide real-time information on the trustor's relation to the trustee but also establish long-term trends.

Feedback

In most cases, and particularly in the military field, it is difficult to access to the "ground truth" and to the respective performance of user and AI that would be necessary to adequately calibrate the level of trust. As a practical answer to this issue, Miller (2018) suggests that users should be given the opportunity to consult explanations on AI decisions after a mission. He points out that this feedback process would fit perfectly into the work process of military operators, who debrief after each mission and learn continuously from their experience. In general terms, "feedback can be defined as any message generated in response to a learner's action" (Mason & Bruning, 2001). For sciences of education, one of the central objectives of giving a feedback is to fill the gap between the learner's current state of knowledge (impacting their current performance along missions) and an optimal state of knowledge (Lipnevich & Panadaro, 2021). To fulfill this objective, different types of feedback have been proposed, such as verification feedback and elaborated feedback (Bosc-Miné, 2014). "Verification feedback" can be similar to that of Dzindolet's (2003) work on trust who simply proposes to display a level of performance for the AI and the user for a given task. "Elaborated feedback" aims to provide explanations of how to achieve a satisfying result, generally under the form of decision rules. When applied to trust domain, verification feedback allows us to calibrate the beliefs related to the optimal performance of the system, while elaborated feedback modifies the beliefs the user can have upon the process of decision itself. The timing of feedback is also crucial: immediate feedback appears to be beneficial for immediate decision-making and tasks involving new information, while delayed feedback appears to be more effective for higher-level tasks and long-term memorization (Mason, 2001; Butler, 2007). The effectiveness of delayed feedback lies in the fact that it allows incorrect responses to dissipate and presents information in a spaced manner, which improves memorization (Butler, 2007).

Trust Estimation

Feedback appears to be an effective way of reducing or increasing the trust of a trustor or trustee over time (Fer et al., 2023). But reducing or increasing the level of trust obviously needs to be able to estimate its level and trend along time. It is thus necessary to detect at an early stage when a user is moving towards states of over- or under-trust and to apply methods for repairing or dampening trust (De Visser, 2018) through adequate feedback. To do so, different types of models exist (Rodriguez, 2023) with either behavioral, physiological, attitudinal or both types of input data. For example, Lee and Moray (1992, 1994) propose to use the VARMA models and stochastic difference equations to model the dynamics of trust based on system performance, faults, and history. One of the most common approaches is that of probabilistic models, with examples such as Bayesian models like OPTIMo (Xu, 2015). This approach models trust as a latent state updated online from performance, human interventions, and explicit feedback. There are also methods using Markov Models with the POMDP models (Williams, 2023; Akash, 2020), as well as Quantum Models (Roeder, 2023). Finally, statistical learning and machine learning (ML) models, such as multiple regression models or neural networks have been recently proposed (Akash, 2018). We see several limitations to these approaches: the first one is related to the estimation of trust after each interaction. Actually, trust evolves through experience and tends to stabilize over time if there is no trust violation (Beggiato, 2015; Zafari, 2024) or if external interventions do not come into play (De Visser, 2018). All methods mentioned here above do not take into account these aspects nor long term stabilization of trust, but only immediate and short-term variation of trust. The second limitation is the systematic use of AI performance levels in trust estimation models. As mentioned earlier, the performance of AI is not necessarily known to the trustor, or if it is known, it is known much later in some cases. It can also be misleading, as it may indicate that AI has a 90% success rate but has systematically erred in a particular class.

Study Objectives

The present study seeks to model trust evolution over the course of a visual search experiment with AI-based recommendations. The main objective is to evaluate the predictive accuracy of different supervised machine learning models for regression, and to examine outcomes as function of the feature set (predictors) utilized. We thus hypothesize that trust variance over successive trials can be explained by initial trust, the participant id, the type of feedback received (basic, performance or elaborated feedback), the moment of measures, different metrics representing the valence of those feedbacks, and metrics of past human-AI dissonance and compliance outcomes.

This paper is organized as follows. We present the data collection phase carried out during an experiment in Section 1. Section 2 explains the predictor variables used for modeling. In Section 3, we present our approach and the ML regression models used. Section 4 presents the results obtained by different models and finally iterates on the best model to refine results. We will conclude with a summary of the key findings and provide a discussion on current limitations and prospects for future work.

DATA COLLECTION

Task

We conducted a longitudinal study, as defined by Ployhart and Vandenberg (2010) (i.e., including at least 3 repeated measurements over time), experiment divided into five sessions with two days between each session. In each session, participants were asked to classify 30 images in which soldiers may or may not be present. For each image, participants had 30 seconds to decide whether a soldier was present or not, then an AI recommendation appeared for five seconds, and participants had 10 seconds to make their final decision. Before the first image, and then every 10 images, a question about trust in AI and self-confidence was asked on a slider from 1 to 7. The day after each session, feedback was provided, the type of which differed depending on the group assigned to the participant. The first group systematically received performance feedback including the participant's individual score, the AI score, and the final decision score. The second group systematically received elaborated feedback in which six cases encountered during the previous mission were accompanied by the AI's probabilistic explanation and the ground truth (soldier present or not). Then the third group was assigned to a baseline condition receiving only the ranking, which served as a motivator for the participants, indicating the participant's score, the first overall score, and the last overall score. This ranking was also provided to the other groups.

Participants

53 volunteers, comprising 17 women (M = 34.41, SD = 10.38) aged between 24 and 56, and 36 men (M = 28.63, SD = 8.2) aged between 20 and 51, took part in our experiment. We observed five levels of education: 1 person with a baccalaureate, 8 with two or three years of higher education, 3 with four years of higher education, 29 with five years of higher education, and 12 with eight years of higher education.

Data

The data collected was divided into three types: declarative data, collected through single questions about trust in AI and self-confidence; behavioral data such as compliance, initial decision time, or final decision time, which were directly integrated into the applications; and experience-related stimuli such as the number of dissonances (the AI disagrees with the initial decision) or information related to different types of feedback.

Table 1: Variables used for the first batch of training.

Variable	Short Description	Range
Group	Type of feedback received (1: performance, 2: elaborated, 3: baseline).	1 to 3 Integer
Id	Participant number	1 to 53 Integer
Measure	The time of measure (1 to 4: session 1; 5 to 8: session 2;...).	1 to 20 Integer
Initial Trust	Trust measured in measure 1.	1 to 7 Float
Last Delta	Difference in points mentioned in the feedback between the last participant and the actual participant (if the participant is the last participant, then the delta is 0).	0 to 391 Integer
First Delta	Similar to the above but in relation to the first participant.	0 to 391 Integer
Elaborated Delta	Difference in the number of errors made by AI compared to the participant in the six cases presented.	0 to 13 Integer
Performance Delta	Difference between the number of errors made by AI and the number of errors made by the participant on the 30 images.	0 to 31 Integer
Dissonance	Number of times the AI recommendation differs from the initial decision during a set.	0 to 10 Integer
Compliance	Number of times the participant's final decision corresponds to the AI recommendation during a set.	0 to 10 Integer

Variables

For the first round of training, input variables considered were: the group, the id, the time of measure, the initial level of trust, the last delta, the first delta, the elaborated delta, the performance delta, the number of dissonances and the number of compliance. These variable are defined in **Table 1.** In the second round of training, to align with the vision of feedback that modifies previous beliefs and the temporal aspect of the evolution of trust, a notion of memory was added to each of the delta, dissonance, and compliance variables. This memory is the cumulative sum of this variable since the first measure. In the final round, each variable was removed one by one to verify its impact on the model's performance.

The model output (dependent variable) corresponds to the declared trust in AI for each set ranging from 1 to 7 with .5 increment.

DATA PRE-PROCESSING

We observed significant variability in the reported trust during the first session compared to the other sessions, which was consistent with the theoretical framework of the trust-building phase (Söllner, 2016). We therefore decided to remove the data related to the first session from our model training to reduce noise in the dataset leaving us with 16 data lines per participant, or 848 lines in total. We then normalized all our variables on an individual basis from 0 to 1 prior to model training.

MODELS

We trained six supervised ML regression models: a multiple linear regression model, a decision tree regression model, a random forest regression model, a multi-layer perceptron model, a support vector regression model and a k-nearest neighbors model. We used Python and the Scikit-Learn library to train these models using the Cognitive Shadow toolkit (Lafond et al., 2020; MacLean et al., 2024).

RESULTS

First Round

The first round of training focused on the ten initial variables formulated for this study, without any cumulative memory. Model predictive accuracy was evaluated using the ten held-out folds in a 10-fold cross-validation. The RMSE corresponds to the average difference between the estimated value and the actual trust value (from 1 to 7).

Table 2: Training of the 6 models on the 10 basic variables without memory variables.

Model	R^2	RMSE
Linear Regression	52.27%	0.84
Decision Tree	**64.9%**	**0.72**
Random Forest	64.22%	0.73
K-Nearest neighbors	58.37%	0.78
Multi-Layer perceptron	59.46%	0.78
Support vector	58.74%	0.78

Results in the **Table 2** show that the Decision Tree and the Random Forest models were best at predicting trust ratings both in terms of accuracy (RMSE) and variance explained (R^2). Given the categorical nature of the id and the group, these two models naturally stand out from the rest, as they should allow similar profiles to be isolated by branch.

Given the focus of this model on the different feedback types received after each set, we call the resulting decision tree model the Feedback-based Dynamic Trust Model (F-DTM). Although our initial results are interesting, one limitation of the resulting model is the lack of memory of past feedbacks encountered in earlier sets.

Second Round

To compensate for the previous lack of consideration of temporality in the models, we decided to include a memory variable for six variables, namely the four deltas, dissonance, and compliance (effectively replacing them but also keeping the non-cumulative dissonance and compliance metrics for the current set). The group, id, trust initial and measurement variables did not lend themselves to a cumulative memory metric. The F-DTM was thus revised to go from 10 to 12 predictor variables with this cumulative memory approach.

Table 3: Training of the 6 models with the 16 input variables.

Model	R^2	RMSE
Linear regression	51.9%	0.85
Decision Tree	**68.57%**	**0.67**
Random Forest	64.17%	0.73
K-Nearest neighbors	45.83%	0.9
Multi-Layer Perceptron	60.33%	0.72
Support Vector	54.87%	0.81

Results, in **Table 3**, showed an improvement for the Decision Tree model which gained approximately 3.7% R^2 and improved RMSE by 0.05. With the addition of these sums, we believed that the unit variables had become redundant and could reduce the performance of the models. We first attempted to remove all unit variables whose cumulative values had been added (the four delta variables, compliance, and dissonance). Unexpectedly, we observed no deterioration or improvement in results (with the Decision Tree: $R^2 = 68.32\%$, RMSE = 0.68). To try to improve the results, we then re-added the compliance and dissonance variables, which improved the Decision Tree model results ($R^2 = 69.92\%$, RMSE = 0.66). By adding only the deltas (without compliance and dissonance), no improvement or deterioration was observed ($R^2 = 68.72$, RMSE = 0.67). In view of these results, for our next analysis, we decided to remove the four deltas but kept the associated sums, compliance, cumulative compliance, dissonance, and cumulative dissonances. We next sought to examine the impact of each variable on model performance, specifically for the Decision Tree model (the best model in Round 1 and Round 2).

Third Round

To verify the impact of each variable, we removed them one by one and retrained the Decision Tree model. We expected the performance of the best model, $R^2 = 69.92\%$ and RMSE = 0.66, to decrease for each variable removed. **Table 4** presents the resulting model statistics and the variables' removal impact in parentheses.

In **Table 4** we observe that the variables id and initial trust have a major impact on estimated trust. Given the structure of the decision tree model, the id allows for refinement by branch for each user, which makes it possible to tailor the prediction to the individual rather than the group. Next, through initial trust, we show the importance of anchorage for the model, which logically allows the model to establish a range of trust associated with the individual. We then observe a slight improvement for the removal of the Cumulated Delta Elab and Cumulated Delta Performance variables. This slight improvement suggests that those features are counterproductive, likely due to the fact that for those who do not receive them, it is a variable equivalent to 0 on all lines. In future work, we plan to train models separately by group and including only relevant features for each group (Liu, 2021) rather than globally as we did here.

For the remaining variables, we see a slight decrease in performance, showing that they are still important, although minimal for some, in estimating trust.

Table 4: Training with the decision tree model with 12 variables as input, i.e., the 12 variables minus the variable specified in the "Without" column.

Without	R^2	RMSE
Group	67.8% (–2.12%)	0.68 (+0.02)
Measure	69.16% (–0.76%)	0.66 (+–0)
Id	60.68% (–9.24%)	0.76 (+0.1)
Cumulated DLast	68.59% (–1.33%)	0.67 (+0.01)
Cumulated DFirst	67.9% (–2.02%)	0.68 (+0.02)
Cumulated DElab	70.51% (+0.59%)	0.65 (–0.01)
Cumulated DPerf	70.39% (+0.47%)	0.65 (–0.01)
Dissonance	68.25% (–1.47%)	0.68 (+0.02)
Cumulated Dissonance	69.18% (–0.74%)	0.67 (+0.01)
Compliance	67.72% (–2.2%)	0.68 (+0.02)
Cumulated Compliance	68.74% (–1.18%)	0.68 (+0.02)
Initial Trust	51.97% (–17.95%)	0.84 (+0.18)

SYNTHESIS OF RESULTS AND CONCLUSION

The present study investigated quantitative modeling of trust evolution in a visual search experiment where human participants were aided by AI-based recommendations. We trained six types of regression machine learning algorithms to estimate declared trust level (1 to 7) with ten key variables as inputs. Other variables were considered but did not show significant results with trust in AI, such as initial decision time and final decision time. Findings in a first round of training highlighted the decision tree and the random forest models, giving an R^2 of 64.9% and an RMSE of 0.72 for the decision Tree. However, given that the data points involved a history of successive trials, we attempted in a second round to add a cumulative memory variable for six of the key metrics (the four deltas, compliance, and dissonance) while removing the four non-cumulative deltas to avoid data redundancy. These modifications allowed us to achieve an R^2 of 69.92% and an RMSE of 0.66 with the decision tree while we observed a drop in the random forest model's performance.

Finally, this study shows that after removing two unnecessary variables, the Feedback-Based Dynamic Trust Model achieved a final predictive accuracy of 70.41% R2 and a RMSE of 0.65.

Despite interesting and encouraging results on these regression models, we are aware that our approach of using cumulative sums to act as memory could be replaced by more complex autoregressive models designed for time-series data. That is why, in our future work, we will use models with native temporality management such as LSTM, HMM, vector autoregression, or ARIMA. Additionally, models will be trained per group, or per user, as in Liu (2021), rather than globally. New dependent variables for trust modeling relative to AI trustworthiness will also be considered, such as under-trust, calibrated trust, and over-trust.

ACKNOWLEDGMENT

The authors would like to thank IMT Atlantique for its support in funding this collaboration with Thales Canada. They would also like to thank the ANRT for funding the Thesis No. 2022/1217.

REFERENCES

Akash, K., Hu, W.-L., Jain, N. & Reid, T. (2018). A Classification Model for Sensing Human Trust in Machines Using EEG and GSR. ACM Trans. Interact. Intell. Syst. (TiiS) **8**, 1–20.

Akash, K., McMahon, G., Reid, T. & Jain, (2020). N. Human Trust-Based Feedback Control. IEEE Control Syst. **40**, 98–116.

Beggiato, M., Pereira, M., Petzoldt, T. & Krems, (2015). J. Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. Transp. Res. Part F: Traffic Psychol. Behav. **35**, 75–84.

Butler, A. C., Karpicke, J. D. & Roediger, H. L. (2007). The Effect of Type and Timing of Feedback on Learning From Multiple-Choice Tests. J. Exp. Psychol.: Appl. **13**, 273–281).

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G. & Beck, H. P. (2003). The role of trust in automation reliance. Int. J. Hum.-Comput. Stud. **58**, 697–718.

Fer, V. et al. (2023). Trust in Automation: Analysis and Model of Operator Trust in Decision Aid AI Over Time. Conference on Artificial Intelligence for Defense, DGA Maîtrise de l'Information.

Kohn, S. C., Visser, E. J. de, Wiese, E., Lee, Y.-C. & Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. Front. Psychol. **12**, 604977.

Lafond, D., Labonté, K., Hunter, A., Neyedli, H. F., & Tremblay, S. (2020). Judgment analysis for real-time decision support using the Cognitive Shadow policy-capturing system. In Human Interaction and Emerging Technologies: Proceedings of the 1st International Conference on Human Interaction and Emerging Technologies (IHIET 2019), August 22–24, 2019, Nice, France (pp. 78–83). Springer International Publishing.

Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**, 1243–1270.

Lee, J. D. & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. Int. J. Hum.-Comput. Stud. **40**, 153–184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50–80.

Lipnevich, A. & Panadero, E. (2021). A Review of Feedback Models and Theories: Descriptions, Definitions, and Conclusions. Front. Educ. **6**, 720195.

Liu, J., Akash, K., Misu, T. & Wu, X. (2021). Clustering Human Trust Dynamics for Customized Real-Time Prediction. in 2021 IEEE International Intelligent Transportation Systems Conference (ITSC) 1705–1712.

MacLean, M. Lafond D., & Li, J. (2024). Capturing Expert Judgment Policies for Assessing Automated Sonar Range Prediction Credibility. 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS), Toronto, ON, Canada, 2024, pp. 1–5.

Mason, B. & Bruning, R. (2001). Providing Feedback in Computer-based Instruction: What the Research Tells Us. in vol. 15.

Mayer, R. C., Davis, J. H. & Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. Acad. Manag. Rev. **20**, 709–734.

Metcalfe, J., Kornell, N. & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. Mem. Cogn. **37**, 1077–1087.

Miller, C. A. (2018). Displaced interactions in human-automation relationships: Transparency over time. In International Conference on Engineering Psychology and Cognitive Ergonomics (pp. 191–203). Cham: Springer International Publishing.

Roeder, L. et al. (2023) A Quantum Model of Trust Calibration in Human–AI Interactions. Entropy **25**, 1362.

Smith, T. A. & Kimball, D. R. (2010). Learning From Feedback: Spacing and the Delay–Retention Effect. J. Exp. Psychol.: Learn., Mem., Cogn. **36**, 80–95.

Parasuraman, R. & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Hum. Factors: J. Hum. Factors Ergon. Soc. **39**, 230–253.

Parasuraman, R., Sheridan, T. B. & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Trans. Syst., Man, Cybern.- Part A: Syst. Hum. **30**, 286–297.

Ployhart, R. E. & Vandenberg, R. J. (2010). Longitudinal Research: The Theory, Design, and Analysis of Change. J. Manag. **36**, 94–120.

Söllner, M. & Pavlou, P. (2016). A Longitudinal Perspective on Trust in IT Artefacts. In: European Conference on Information Systems (ECIS), Istanbul, Turkey.

Rodriguez, L. R. et al. (2023). A review of mathematical models of human trust in automation. Front. Neuroergonomics **4**, 1171403.

Visser, E. J. de, Pak, R. & Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. Ergonomics **61**, 1409–1427.

Williams, K. J., Yuh, M. S. & Jain, N. (2023). A Computational Model of Coupled Human Trust and Self-confidence Dynamics. ACM Trans. Hum.-Robot Interact. **12**, 1–29.

Xu, A., & Dudek. (2015). Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, 221–228.

Zafari, S. et al. (2024). Trust Development and Explainability: A Longitudinal Study with a Personalized Assistive System. Multimodal Technol. Interact. **8**, 20.