Ergonomics in Design, Vol. 47, 2022, 145–150
https://doi.org/10.54941/ahfe1001931

Toward Adaptive Trust Management for Human-Automation Teaming Using an Instance-Based Learning Cognitive Model

Wen-Li Dong, Wei-Ning Fang, Bei-Yuan Guo, Jian-Xin Wang, and Hai-Feng Bao

State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

ABSTRACT

Trust in automation is seen as a core factor affecting human-automation teaming. Inappropriate calibration of trust in automation can damage the performance and safety goals of the collaborative team. It is essential to develop automation that can correctly calibrate human trust in it. Herein, based on the view that trust comes from interaction, we use an instance-based learning cognitive model to obtain the cognitive process involved in the interaction between dispatchers and automated Decision Support Systems (DSSs) in the Fully Automatic Operation (FAO) circumstances, and obtain from the model an internal estimate of the calibration state of human trust. We consider integrating the model into automation so that it can judge the hidden calibration status of the human teammate's trust, and respond to the trust dynamics in an online and adaptive way. We discuss our results and the potential of the instance-based computational cognitive process model to improve human-automation teaming. Our model has great potential to avoid the sluggish effect caused by dispatchers failing to obtain effective decision support in time in the FAO circumstances, especially when dealing with emergencies under high time pressure.

Keywords: Trust in automation, Human-automation teaming, Cognitive model, FAO

INTRODUCTION

With the rapid development of rail transit, the FAO system has become a hot research topic in the current urban rail transit field. In the FAO circumstances, the train will be unmanned, the driver's responsibilities will be replaced by the system, and the emergency response level of the system will depend more on the remote operation of dispatchers (Commission, 2006). In some critical decision-making tasks, because raw data about the state of the system is not available to the dispatcher, or because the function of the system is opaque and unclear to the dispatcher, DSSs can be used by highlighting relevant areas, providing suggestions and the direction of action, and even in some cases executing it for the operator, helps supplement or clarify information already available to the dispatcher (Madhavan and Wiegmann, 2007), therefore, the FAO dispatcher work in teams with the DSSs to ensure the safe operation of the train.

Trust in automation has been identified as a key factor in mediating the relationship between human operator and automation (Lee and See, 2004; Hoff and Bashir, 2015) and the operator's trust in automated team members often has a significant impact on the decision-making process of the human-automation team (Drnec et al., 2016). Due to the safety-critical nature of the FAO system, the issue of trust in automation is likely to become a bottleneck that limits the performance of the system and impairs the security of the system. It is crucial to develop DSSs that can calibrate the dispatchers' trust in them appropriately.

For the development of trustworthy automation, traditional work focuses on studying the antecedents and consequences of trust in automation, improving automation design statically (Schaefer et al., 2016). However, trust in automation is not a static phenomenon—trust fluctuates dynamically as the interaction unfolds over time, and the design's efficacy is often dependent on the context and individual differences among human operators. To resolve this issue, recent work has tended to empower automation reasoning ability through methods such as probabilistic modeling (Xu et al., 2015; Akash et al., 2018; Akash et al., 2019a; Akash et al., 2019b), to develop strategies for automation to proactively acquire, calibrate, and maintain the human teammate's trust. However, a strong limitation of most of them is their inability to continuously and dynamically learn from experience and, thus, update the rules of learning.

A large number of related literatures show that existing conceptual theories (Lee and See, 2004) and computational modeling work (Akash et al., 2017; Hu et al., 2018) of trust in automation have reached a consensus: the operator's calibration of trust and reliance is the result of learning and memory from the experience of the automation technology or system. As a learning theory related to dynamic decision-making, Instance-Based Learning (IBL) can well explain the dynamic development mechanism of trust from the perspective of cognitive structure (Gonzalez et al., 2003). IBL argues that in dynamic decision-making, people learn through the accumulation and refinement of instances, which contain situations, actions, and utility of decisions. When decision makers interact with dynamic tasks, they learn based on the similarity of the situation to past instances to identify situations, adjust their judgment strategies from heuristic-based to instance-based, and refine accumulated knowledge based on feedback on the outcomes of their actions.

Therefore, this paper constructs the Dispatching Multitasking Trust Paradigm (DMTP), studies the dynamic change process of dispatcher's trust in automation in the FAO circumstances, builds a computational cognitive model of trust based on IBL theory, captures the dynamic changes of trust, and take the first step to create adaptive trust management to improve the efficiency and safety of FAO metros.

The Human-Automation Teaming Task

According to the investigation of Yanfang Line of Beijing Metro dispatching operation and related case study (Balfe et al., 2018), the dispatching work can be summarized into the following four categories: monitoring, intervention,

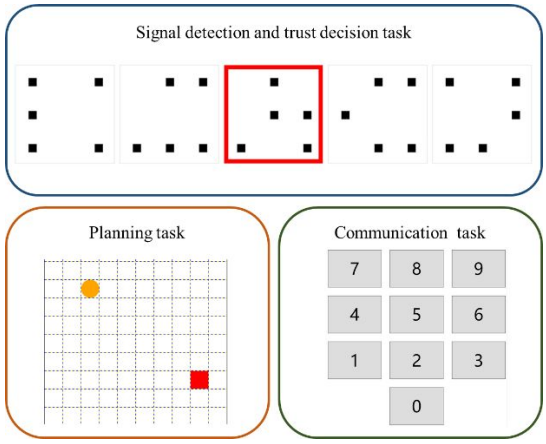

Figure 1: The main components of DMTP, where the planning task is a gridworld-based path planning task, and the communication task is an auditory 2-back task.

planning, and communication. On this basis, we developed DMTP, the basic components of which are shown in Figure 1.

In DMTP, we can examine the underlying cognitive processes underlying dispatcher trust decisions in the presence of DSS unreliability. In DMTP, subjects will play the role of dispatchers, performing monitoring tasks and handling dangerous situations. The subjects were asked to detect and judge whether the abnormal signal was safe or dangerous. To make this decision, they took advice from a DSS. The system will prompt the subjects that the abnormal signal may be a danger signal. However, the DSS recommendations are not necessarily correct, false positives (declaring a safety signal as dangerous) and false negatives (declaring a danger signal as safe) may occur. Therefore, subjects' trust in the DSS fluctuates with its performance, which allows us to explore the trust dynamics of subjects as they interact with the DSS.

The goal of the subjects is to detect abnormal signals in the shortest possible time, judge whether the abnormal signals are dangerous signals, and solve the dangerous events by processing the planning task, so as to optimize their task performance. Among them, the subjects interact with DSS with different reliability in each block. If the subjects can detect abnormal signals, make correct trust decisions, and successfully complete the planning task of solving dangerous events, they can obtain three scores corresponding to the reward and penalty values associated with each task in various tasks (signal detection task, trust decision task, and planning task), and lose points otherwise.

When making trust decisions, subjects can view information describing the reward and penalty values for each DSS (we assume that during the reconnaissance phase before anomalous signal processing, the dispatcher can observe and obtain as much information as possible about the reliability of the DSS).

At the end of each trial, subjects were asked to explicitly indicate their level of trust in the DSS using a 7-point Likert scale (1: distrust, 7: trust), with a score of 4 indicating moderate trust, or there is no inclination.

Cognitive Model of Trust Decision

In IBL, decisions are made by generalizing about past experiences or instances that are similar to the current situation. Typically, experiences are encoded as chunks in declarative memory, containing attributes that describe the context of each decision, the decision itself, and the outcome of the decision. In this model, the context attributes include whether the DSS cue is present (present or absent), the reward value (ranging from 0 to 10), the penalty value (ranging from 0 to 10), and the possible decision to trust or distrust. The result is the actual score based on the action. In a given situation, for each possible decision, the relevant utility (i.e., the expected outcome) is computed by blending: the average of past outcomes weights the probability of memory retrieval, which depends on contextual similarity to past instances. Make the decision with the highest expected outcome.

The cognitive model is implemented in the ACT-R cognitive architecture(Anderson et al., 2004) and follow the IBL approach for decision making(Gonzalez et al., 2003). According to the blending mechanism of ACT-R, the retrieval of past instances is based on the activation strength of relevant instances in memory and their similarity to the current context. The activation A_i of instance i is determined by:

$$A_i = \log\left(\sum_{j=1}^{n} t_j^{-d}\right) + \varepsilon_i \tag{1}$$

where, t_j is the time since the jth occurrence of instance i, d is the decay rate of each occurrence, set to the default ACT-R value of 0.5. ε_i is the transient noise, a random value from the logistic distribution, whose mean value is 0 and the variance parameter s is 0.25(common act-r value), which introduces randomness into the retrieval.

Relevant memories are retrieved in specific contexts by combining their activation and relevancy to calculate their match scores:

$$M_i = A_i + \sum_{j=1}^{l} \mathrm{MP} \times \mathrm{Sim}\left(d_j, v_{\mathrm{ij}}\right) \tag{2}$$

where, $\mathrm{Sim}(d_j, v_{\mathrm{ij}})$ is the similarity between the current context element (d_j) and the corresponding context element (v_{ij}) of the instances in memory, and similarities between numerical slot values are computed on a linear scale from 0.0 (exact match) to -1.0(largest difference), the symbolical value is either an exact match or the largest difference. MP is mismatch penalty (set to ACT-R default of 1.0).

The IBL model uses the blending mechanism of ACT-R to generate expected outcomes of possible actions based on similarity to past instances. The desired result is the value V that best satisfies all constraints that match instance i, weighted by the probability of retrieval, where satisfaction is defined as minimizing the dissimilarity between the consensus value V and the actual answer $V_i$contained in instance i:

$$V = \underset{V_j}{\operatorname{argmin}} \sum_{i=1}^{k} P_i \times \operatorname{Sim}\left(V_j, v_{ij}\right)^2 \tag{3}$$

where, V is the consensus value in the possible value set V_j, P_i is the probability weight of memory i, and its matching score M_i is reflected through the Boltzmann softmax distribution.

$$P_i = \frac{e^{\frac{A_i}{t}}}{\sum_j e^{\frac{A_j}{t}}} \tag{4}$$

The temperature parameter t can be used to scale the probability according to the activation, i.e. low temperature results in a larger proportion assigned to the highest activation instance, while high temperature results in a more randomly distributed proportion, regardless of the activation strength. The current model sets the temperature to 1.0, which results in retrieval probabilities that reflect the original probability distribution, not biased towards or against the most active instances.

In summary, results from past instances are weighted by their recency, frequency, and similarity to the current instance (i.e., the probability of memory retrieval) to produce an expected result by blending. The corresponding trust action is made according to the generated expected result.

CONCLUSION AND FUTURE WORK

The current method is an initial attempt to solve the problem of FAO dispatchers' trust in automation. Studying trust from the perspective of cognitive structure can not only integrate various empirical findings in the trust literature, but also understand the relationship between trust and other cognitive mechanisms and phenomenon. More importantly, cognitive models are generative, in the sense that they actually make decisions in a human-like manner, based on their own experience, rather than being data-driven and requiring a lot of training set. Therefore, after conducting experiment with professional FAO dispatchers and DMTP to verify the validity of the constructed model, the future work considers equipping the cognitive model in the automated DSS, to deliver the right information at the right time to dispatchers based on their trust needs, improving the accuracy of dispatcher trust calibration, and developing human-automation teaming that facilitates more effective automated DSSs. In addition, whether the model's ability to predict dispatcher trust calibration can be generalized to less constrained dispatching environments requires further exploration.

ACKNOWLEDGMENT

This work was supported by Beijing Natural Science Foundation (L191018).

REFERENCES

Akash K, Hu W-L, Reid T, et al. (2017) Dynamic modeling of trust in human-machine interactions. 2017 American Control Conference (ACC). IEEE, 1542–1548.

Akash K, Polson K, Reid T, et al. (2019a) Improving Human-Machine Collaboration Through Transparency-based Feedback–Part I: Human Trust and Workload Model. IFAC-PapersOnLine 51: 315–321.

Akash K, Reid T and Jain N. (2018) Adaptive Probabilistic Classification of Dynamic Processes: A Case Study on Human Trust in Automation. 2018 Annual American Control Conference (ACC). IEEE, 246–251.

Akash K, Reid T and Jain N. (2019b) Improving Human-Machine Collaboration Through Transparency-based Feedback–Part II: Control Design and Synthesis. IFAC-PapersOnLine 51: 322–328.

Anderson JR, Bothell D, Byrne MD, et al. (2004) An integrated theory of the mind. Psychological review 111: 1036.

Balfe N, Sharples S and Wilson JR. (2018) Understanding is key: an analysis of factors pertaining to trust in a real-world automation system. Human factors 60: 477–495.

Commission IIE. (2006) Railway applications–Urban guided transport management and command/control systems–Part 1: System principles and fundamental concepts. Genf.

Drnec K, Marathe AR, Lukos JR, et al. (2016) From trust in automation to decision neuroscience: applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. Frontiers in human neuroscience 10: 290.

Gonzalez C, Lerch JF and Lebiere C. (2003) Instance-based learning in dynamic decision making. Cognitive Science 27: 591–635.

Hoff KA and Bashir M. (2015) Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors 57: 407–434.

Hu W-L, Akash K, Reid T, et al. (2018) Computational modeling of the dynamics of human trust during human–machine interactions. IEEE Transactions on Human-Machine Systems 49: 485–497.

Lee JD and See KA. (2004) Trust in automation: Designing for appropriate reliance. Human factors 46: 50–80.

Madhavan P and Wiegmann DA. (2007) Similarities and differences between human–human and human–automation trust: an integrative review. Theoretical Issues in Ergonomics Science 8: 277–301.

Schaefer KE, Chen JY, Szalma JL, et al. (2016) A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human factors 58: 377–400.

Xu AQ, Dudek G and Acm. (2015) OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. Proceedings of the 2015 Acm/Ieee International Conference on Human-Robot Interaction. New York: Assoc Computing Machinery, 221–228.