Calibrating Trust in Autonomous Systems in a Dynamic Environment

Kazuo Okamura (ok@nii.ac.jp)
The Graduate University for Advanced Studies, SOKENDAI, Tokyo, 101-8430, Japan

Seiji Yamada (seiji@nii.ac.jp)
National Institute of Informatics and SOKENDAI, Tokyo, 101-8430, Japan

Abstract

Appropriately calibrating trust in autonomous systems is essential for successful collaboration between humans and the systems. Over-trust and under-trust often happen in dynamically changing environments, and they can be major causes of serious issues with safety and efficiency. Many studies have examined the role of continuous system transparency in keeping proper trust calibration; however, not many studies have focused on how to find poor trust calibration nor how to mitigate it. In our proposed method of trust calibration, a behavior-based approach is used to detect improper trust calibration, and cognitive cues called "trust calibration cues" are presented to users as triggers for trust calibration. We conducted an online experiment with a drone simulator. Seventy participants performed pothole inspection tasks manually or relied on the drone's automatic inspection. The results demonstrated that adaptively presenting a simple cue could significantly promote trust calibration in both over-trust and under-trust cases.

Keywords: Trust Management,Trust Calibration

Introduction

Rapid advances in autonomous technologies are changing all aspects of our daily life. One of the early works (Chambers & Nagel, 1985) on human factors in flight automation already investigated a wide range of design considerations to realize a safe and efficient relationship between the pilot and the system.

Trust is known as one of the critical concepts in collaboration between human users and autonomous systems. Successful collaboration requires the users to appropriately adjust their level of trust to the actual reliability of systems. This cognitive process is called trust calibration (Muir, 1994; Lee & See, 2004). Users often fail to calibrate their trust in a system and end up in a state called over-trust or under-trust when the system's reliability changes for various reasons in an environment. Over-trust is poorly calibrated trust in which the user overestimates the reliability of the system, Under-trust is poorly calibrated trust in which the user underestimates the system's capability. Poor trust calibration often causes not only the performance of the collaboration to degrade but also serious safety issues (Parasuraman & Dietrich H. Manzey, 2010; NHTSA, 2017).

In keeping appropriate trust, it is necessary to be able to measure trust and to influence trust if necessary. However, these two elements are still challenging issues.

Measuring trust is difficult, as trust is a latent construct. Self-reported trust measures used by most of the trust research are too intrusive to use them during task executions. Trust questionnaires conducted at the end of an experiment sometimes do not correctly reflect real-time trust during the experiment (Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013). Some studies examined the effectiveness of physiological and neural measures such as gaze (Hergeth, Lorenz, Vilimek, & Krems, 2016), heart-rates, and EEG. Although these are promising approaches, further research would be necessary to clarify the correlation between trust and these metrics.

Many studies (Rempel, Holmes, & Zanna, 1985; Muir, 1987; Hoff & Bashir, 2015; Schaefer, Chen, Szalma, & Hancock, 2016) investigated the factors influencing trust. They suggested that it would be complex and difficult to calibrate trust by manipulating those factors, since there are many interactions and dependency among them.

Most of the existing research on trust calibration, such as (McGuirl, Sarter, John M. McGuirl, & Nadine B. Sarter, 2006; de Visser, Cohen, Freedy, & Parasuraman, 2014; Helldin, 2014; Haeuslschmid, Buelow, Pfleging, & Butz, 2017), emphasized the importance of system transparency to maintain appropriate trust. Studies on trust in autonomous driving such as Helldin, Falkman, Riveiro, and Davidsson (2013); Haeuslschmid et al. (2017) also demonstrated that providing good transparency by constantly presenting the system information helps maintain the proper trust in the vehicles. They claimed that appropriate trust could be developed if an AI system provides enough information for a human user to obtain a good understanding of the system. Their primary goal is how to avoid trust miscalibration. However, once human users fall into the categories of over-trust or under-trust, it might not be easy for them to escape from the miscalibration status with the system transparency information.

Although recent works such as (de Visser et al., 2019; Tolmeijer et al., 2020) proposed trust calibration models for human-robot teams, not many studies have focused on how to detect improper trust calibration nor how to mitigate it.

To address the research challenges, we propose a framework to define the status of improper trust calibration with a behavior based-measurement of trust. We also examine cognitive cues to notify the users of miscalibration status. A method of adaptive calibration is proposed with the framework and the cognitive cues. We conducted an online experiment using a drone simulator with the ABA/BAB scenarios of under-trust(A) and over-trust(B) by manipulating the weather conditions. The results demonstrate that adaptively presenting a simple cue could significantly promote trust calibration in both over-trust and under-trust cases. As the proposed method is simple but task-independent, we believe it could be a good design baseline for better human-autonomous system collaborations.

Current study

Detection Framework

Suppose a user and an autonomous system are jointly working on a set of tasks. The user should decide whether to rely on the system or do each task manually. In our framework, we focus on performance related factors for trust. Three probabilites, P_{auto}, P_{trust}, and P_{man}, are defined as follows.

- P_{auto}: Probability that a task done by a system will be successful. This is called the "reliability of the system."
- P_{trust}: User's estimation of P_{auto}. This is the user's trust in the system.
- P_{man}: Probability that a task done manually by a user will be successful. This is called the "capability of the user." Note that "*man*" means "*manual*."

P_{auto} changes depending on the conditions of the system. P_{trust} also changes accordingly and becomes equal to P_{auto} if trust is appropriately calibrated. Over-trust occurs if $P_{trust} > P_{auto}$, and under-trust occurs if $P_{trust} < P_{auto}$. Since directly measuring P_{trust} is quite difficult, we modified the definitions of over-trust and under-trust by introducing a third probability P_{man} in addition to P_{trust} and P_{auto} as follows:

- Over-trust: the user estimates that the system is better at a task than the user even though the actual reliability of the system is lower than the user's capability.

$$(P_{trust} > P_{man}) \wedge (P_{man} > P_{auto}) \tag{1}$$

- Under-trust: the user estimates that they are better at a task than the system even though the actual reliability of the system is higher than the user's capability.

$$(P_{trust} < P_{man}) \wedge (P_{man} < P_{auto}) \tag{2}$$

The reliance behaviors of a user can be explained by the user's perception of the reliability of a system and the user's own capability (Gac & Lee, 2006). When a user decides to rely on a system, it is reasonable to say that this behavior indicates $P_{trust} > P_{man}$. If the user decides to do a task manually, it means $P_{trust} < P_{man}$. Thus, the first terms of (1) and (2) can be estimated by observing the user's reliance behavior. As for the second terms, P_{auto} could be calculated with the sensor models and algorithms used to implement the system, and P_{man} could be estimated by using the parameters of a target task and environmental conditions. Therefore, the second terms of (1) and (2) can be also estimated.

Trust Calibration Cue

The second element of the proposed method involves the idea of giving a cognitive cue to users when over-trust or under-trust is detected. This cue is called a "trust calibration cue" (TCC). The four types of TCCs (visual, audio, verbal, and anthropomorphic) were originally proposed in (Okamura & Yamada, 2018). These were designed to be intuitive and effective warning signals (Laughery & Wogalter, 2014). Many studies examined the information associated with trust. de Visser et al. (2014) proposed a design guideline for trust cues, which are information elements used to make a trust assessment about a system. Unlike our TCC, their trust cues were used to display information specific to trust dimensions and stages.

Adaptive Trust Calibration

With the detection framework and TCCs described above, we propose a method of adaptive trust calibration as follows. Details of the detection algorithm in the step 3 will be described in the next section.

Method Adaptive Trust Calibration

```
1: while collaboration tasks are performed do
2:     Observe a user's behavior of reliance on a system.
3:     Evaluate the expression (1) and (2) in the framework.
4:     if over-trust or under-trust is detected then
5:         Present a TCC to the user.
6:     end if
7: end while
```

If our method can effectively mitigate over-trust or under-trust, the following are hypothesized:

[H0] the manual choice rates increase if TCCs are presented in cases of over-trust or decrease if TCCs are presented in cases of under-trust.

[H1] users with TCCs perform better and more robustly than the users without TCCs.

[H2] adaptively presenting TCCs could trigger the trust calibration process more effectively than continuously maintaining system transparency in a conventional way.

Method

Apparatus and Materials

We developed a drone simulator based on an open-source 3D map library CesiumJS(The Cesium Consortium, 2018). Figure 1[1] shows a screen image of the simulator running in the Chrome browser.

Pothole Inspection Task

A pothole is a bowl-shaped depression in the surface of a road and can be a possible cause of traffic accidents. The participants of the experiments were asked to inspect road images from a drone to check if there were any potholes.

A route with 24 checkpoints (CKPs) was defined in the simulated environment. Each CKP was shown as a small yellow circle on the screen. When the drone came close to one of the CKPs on the route, the message shown in Figure 2 (A) popped up and asked the participants to choose

[1]The map images in this manuscript are from the Geospatial Information Authority of Japan (https://maps.gsi.go.jp) CC BY 4.0.

Figure 1: Drone simulator

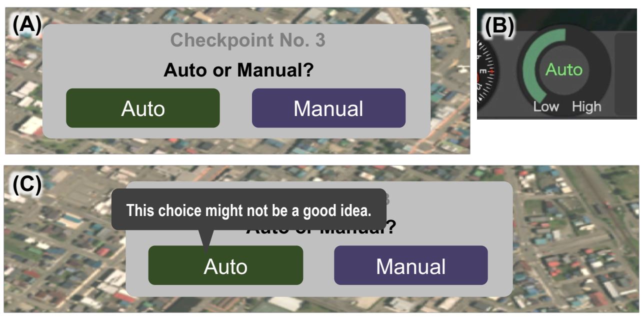

Figure 2: (A) Pop-up for selection, (B) reliability indicator, and (C) verbal TCC

whether to rely on the drone's automatic inspection or to inspect the CKP manually. The indicator shown in Figure 2 (B) always displayed the reliability of the automatic pothole inspection. When the participants selected the "Auto" button, an automatic-inspection result was shown for three seconds with a road image. When the "Manual" button was selected, a road image of the area around the CKP was displayed, and the participants had to make a pothole report manually. Both cases with potholes are shown in Figure 3. Potholes were artificially rendered as irregular shapes in a dark brown color on the road images in the pop-up window. As the pothole inspection task is a remote sensing task, it would be quite difficult for an autonomous system to know the correct answer at the time of each inspection in practical situations. Therefore, the correct answer was not presented to the participants.

We used the verbal TCC shown in Figure 2 (C) because a preliminary experiment revealed that the verbal TCC showed a strong effect on changing participants' behavior. When the

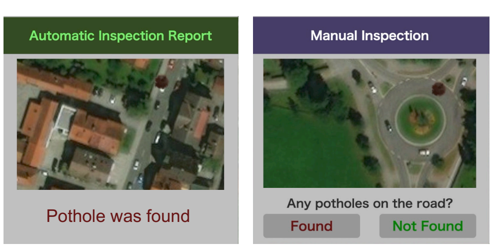

Figure 3: Pothole inspection windows

framework detected over-trust or under-trust depending on the choice a participant made, this TCC was presented right after the choice was made (pushing a button).

Participants and Scenarios

A total of seventy participants (51 male, 19 female) took part in the experiment online. Their ages ranged from 25 to 75 years old ($M = 44.2, SD = 10.3$). They were recruited through a cloud-sourcing service provided by Yahoo! Japan. We defined the ABA/BAB scenarios of under-trust (A) and over-trust (B) by manipulating the weather conditions in order to evaluate the proposed method for both bidirectional trust changes. The performance of the automatic pothole inspection P_{auto} was configured on the basis of signal detection theory (SDT) (Stanislaw, 1999). SDT describes the detection of signals in noisy environments. Noise and signals are represented as two overlapping density distributions. The distance between the two curves represents the sensitivity d' of a system.

In the A condition, good weather conditions were simulated. The screen brightness was 100%, and there were no sound effects except for the sound of the drone flying. P_{auto} and the corresponding sensitivity d' defined in SDT were manipulated to be 0.88 and 2.35, respectively, indicating that the system has a very high discrimination ability. In contrast, the weather conditions were bad in the B condition. A thunderstorm was simulated with a blurred and dark (40% brightness on average) screen and with sound effects. P_{auto} dropped to 0.50, and the corresponding sensitivity d' became 0.1, meaning that the reliability of the automatic pothole inspection had greatly deteriorated. In both ABA/BAB scenarios, each condition continued until eight CKPs were inspected so that the total number of CKPs was twenty four. Participants were randomly assigned to one of four groups: NoTCC-ABA group (without TCC in the ABA scenario), TCC-ABA (with a verbal cue in the ABA scenario) group, NoTCC-BAB group, and TCC-BAB group. Hereinafter, two groups with a common attribute are called the TCC groups, the NoTCC groups, the ABA groups, and the BAB groups.

Estimation of P_{man} and Manipulation Check

Although providing a general estimation model of P_{man} is beyond the scope of this paper, P_{man} under the conditions of the current experiment can be estimated as follows. Geirhos et al. (2018) demonstrated that human image recognition is still better than the top-performing deep neural networks in the case of image degradation such as Gaussian blur or additive Gaussian noise. This finding could provide a basis for estimating the second terms of the proposed framework in the experiment because the pothole inspection became an image recognition task with blurred and noisy road images when the weather conditions turned worse. We assumed that P_{auto} would fluctuate more widely than P_{man} under changing weather conditions, and we estimated that the inequality $P_{auto} > P_{man}$ was true during the good weather period and false during the bad one.

Algorithm *Adaptive Trust Calibration*

```
Initialize:
Total number of checkpoints(CKPs): M = the number of CKPs.;
Over-trust flag list: OT[1], ..., OT[M] are initialized with zero;
Under-trust flag list: UT[1], ..., UT[M] are initialized with zero;
Number of current CKP: i ⇐ 1;

while i ≦ M and not time-over do
  if the drone reached a CKP then
    Estimate P_man and P_auto;
    if choice behavior is AUTO and P_man > P_auto then
      OT[i] ⇐ 1;
      if i ≧ 3 and (OT[i−2] + OT[i−1]]) ≧ 1 then
        Over-trust is detected and TCC is presented to the user;
      end if
    else if choice behavior is MANUAL and P_man < P_auto then
      OU[i] ⇐ 1;
      if i ≧ 3 and (OU[i−2] + OU[i−1]]) ≧ 1 then
        Under-trust is detected and TCC is presented to the
        user;
      end if
    end if
    i ⇐ i + 1;
  end if
end while
```

We checked the validity of this estimation by measuring the manual success rates (P_{man}) in a pre-experiment. Thirty-two participants [25 male, 7 female, mean age 42(SD=12)] were recruited through a cloud-sourcing service provided by Yahoo! Japan. None of them joined the main experiment. They inspected the prepared CKPs manually in accordance with the same procedure of the main experiment. The results indicated that the mean of the manual success rates and the sensitivity d' was 0.83 ($SE = 0.02$) and 1.85 for the A condition and 0.79 ($SE = 0.02$) and 1.69 for the B condition. One sample t-test revealed that $P_{auto} > P_{man}$ in the A condition [$t(47) = -2.26, p = 0.01, Cohen'sd = 0.33$] and $P_{auto} < P_{man}$ in the B condition [$t(47) = -13.66, p < 0.01, Cohen'sd = 1.97$]. We concluded that the estimation was valid under the conditions of the main experiment.

Procedures

The online experiment started with an **instruction phase**. The participants were given an instruction stating that the goal of the experiment was to inspect 24 CKPs within 20 minutes. They were told that the average success rate of manual pothole inspection was around 75%. They also learned that the reliability of the drone's automatic inspection, which is continuously displayed on the indicator, was very high, although it could fluctuate depending on the weather conditions. Next, in the **training phase**, the participants started a practice flight of the drone and learned how to inspect the CKPs. This phase was finished after the first three CKPs were inspected, and **the main phase** of the experiment was started with either condition A or B depending on the scenario of the group. The algorithm *Adaptive Trust Calibration* based on the proposed method was applied.

A simple moving average of three CKPs was used in the algorithm to capture the participants' behavior changes in each condition with eight CKPs. If the participants completed the 24th inspection or the elapsed time exceeded 20 minutes, the main phase was finished.

In this experiment, the three things were measured as the dependent variables. TCC rates are the rates of the frequency at which TCCs were presented to the participants at each CKP, indicating how our method was working during the experiment. Manual rates are the mean values of the manual choice ratio for each condition, showing how the participants relied (or did not rely) on the drone's automatic inspection and therefore indicating their trust calibration status. The sensitivity d' demonstrates the performance of human-autonomous system collaborative tasks.

Results

Seventy participants completed all 24 CKPs within the time limit. Of the 70, 17 were in the NoTCC-ABA group, 18 in the TCC-ABA group, 21 in the NoTCC-BAB group, and 14 in the TCC-BAB group. The average time taken to finish the main phase of the experiment was 9 minutes 5 seconds, which means 22.5 seconds per CKP.

TCC Rates

Within each condition, the TCC rates showed a similar trend in which the values were initially higher and then decreased along the CKP series. For example, for the B condition in the group TCC-ABA, the mean of the TCC rates from CKP 11 to 13 was $0.48(SE = 0.11)$, which then significantly decreased to $0.19(SE = 0.08)$, that is, the mean value from CKP 14 to 16 [$t(17) = 4.53, p < 0.01, Cohen'sd = 0.99$]. The TCC rates for all B conditions ($M = 0.31$ $SE = 0.03$) were significantly higher than those for all A conditions ($M = 0.15$ $SE = 0.02$) [$t(514) = 4.69, p < 0.01, Cohen'sd = 0.39$].

Manual Rates

We evaluated the proposed method by comparing the eight-CKP mean values of the manual rates for each condition so that we could capture the accumulated effects of presenting TCCs. Table 1 shows the means of the manual rates for each condition. C1, C2, and C3 mean A, B, and A for the ABA groups, B, A, and B for the BAB groups.

We conducted a one-way ANOVA [within-subjects design; independent variable: the scenario conditions of three levels, A, B, and A (B, A, and B), dependent variable: manual rate] for each group. All post-hoc analyses were done by using the Holm-Bonferroni method. Figure 4 illustrates the results of the ANOVA. **ABA groups:** The NoTCC-ABA group did not show any significant difference in manual rates among the three conditions [$F(2,32) = 0.20$, $p = 0.82$, $\eta_p^2 = 0.01$]. In comparison, the TCC-ABA group showed significant differences [$F(2,34) = 6.50$, $p < 0.01$, $\eta_p^2 = 0.28$]. The post-hoc analysis indicated that the manual rate significantly increased from the first A condition to the B condition [$t(17) = 3.56$, $adjusted.p < 0.01$], and the rate for the second A condition then significantly decreased [$t(17) = 2.45$, $adjusted.p =$

Table 1: Means of manual rates

Condition	C1	C2	C3
NoTCC-ABA	0.23 (0.08)	0.28 (0.09)	0.26 (0.07)
TCC-ABA	0.19 (0.06)	0.50 (0.06)	0.22 (0.07)
NoTCC-BAB	0.46 (0.08)	0.32 (0.08)	0.63 (0.09)
TCC-BAB	0.45 (0.09)	0.22 (0.08)	0.71 (0.06)

(Standard errors in parentheses.)

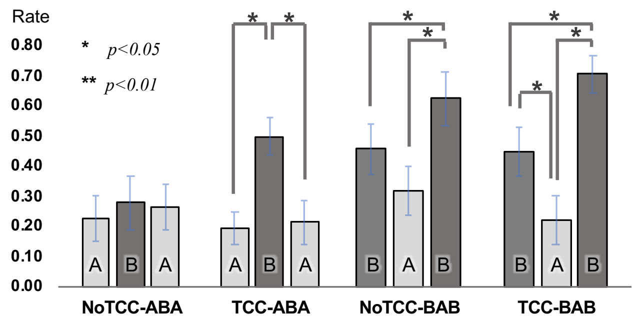

Figure 4: Manual rates

Table 2: Means of the sensitivity d'

Condition	C1	C2	C3
NoTCC-ABA	1.67 (0.05)	1.02 (0.11)	1.74 (0.10)
TCC-ABA	1.46 (0.12)	1.29 (0.09)	1.80 (0.04)
NoTCC-BAB	0.53 (0.21)	1.39 (0.12)	0.67 (0.26)
TCC-BAB	0.88 (0.20)	1.47 (0.10)	0.73 (0.21)

(Standard errors in parentheses.)

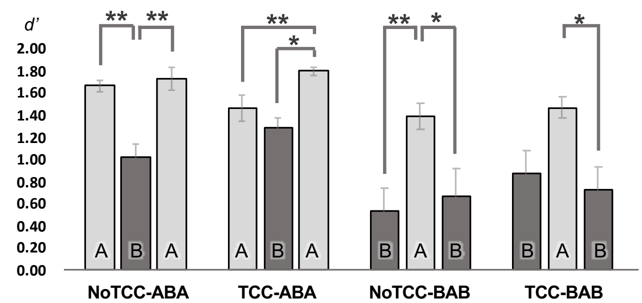

Figure 5: Sensitivity d'

0.03]. **BAB groups:** The NoTCC-BAB group showed significant differences [$F(2,40) = 6.41$, $p < 0.01$, $\eta_p^2 = 0.24$]. The post-hoc analysis showed that the manual rate for the A condition did not change significantly from the first B condition [$t(20) = 1.46$, $adjusted.p = 0.16$]; however, the rate for the second B condition significantly increased [$t(20) = 3.14$, $adjusted.p = 0.02$]. The TCC-BAB group showed significant differences [$F(2,26) = 14.48$, $p < 0.01$, $\eta_p^2 = 0.53$]. The post-hoc analysis indicated that the manual rate for the A condition significantly decreased from the first B condition [$t(13) = 2.65$, $adjusted.p = 0.02$], and the rate for the second B condition then increased significantly [$t(20) = 4.47$, $adjusted.p < 0.01$].

Sensitivity d' (Performance)

Table 2 shows the means of the sensitivity d' for each condition. We conducted the same one-way ANOVA and the results are illustrated in Figure 5. **ABA groups:** For the NoTCC-ABA group, a significant effect was found [$F(2,32) = 14.8$, $p < 0.01$, $\eta_p^2 = 0.48$]. The post-hoc analysis indicated that the mean value of d' significantly decreased from the first A condition to the B condition [$t(16) = 5.26$, $adjusted.p < 0.01$] and then significantly increased from the B condition to the second A condition [$t(16) = 4.05$, $adjusted.p < 0.01$]. For the TCC-ABA group, a significant effect was found [$F(2,34) = 7.52$, $p < 0.01$, $\eta_p^2 = 0.31$]. The post-hoc analysis indicated that the mean value of d' significantly increased from the B condition to the second A condition [$t(17) = 5.44$, $adjusted.p < 0.01$] and also showed a significant difference between the first A condition and the second A condition [$t(17) = 2.61$, $adjusted.p = 0.04$]. **BAB groups:** For the NoTCC-BAB group, a siginificant effect was found [$F(2,40) = 7.45$, $p < 0.01$, $\eta_p^2 = 0.27$]. The post-hoc analysis revealed that the mean value of d' significantly increased from the first B condition to the A condition [$t(20) = 3.76$, $adjusted.p < 0.01$] and significantly decreased from the A condition to the second B condition [$t(20) = 2.98$, $adjusted.p = 0.01$]. For the TCC-BAB group, the significant effect was found [$F(2,26) = 4.75$, $P = 0.02$, $\eta_p^2 = 0.27$]. The post-hoc analysis indicated that the mean value of d' for the A condition marginally increased from that for the first B condition [$t(13) = 2.46$, $adjusted.p = 0.06$]. The mean value of d' for the second B condition significantly decreased from that for the A condition [$t(13) = 3.13$, $adjusted.p = 0.02$].

Discussion

In the ABA scenario, the manual rates for the B condition increased significantly from the first A condition in the ABA-TCC group, while no significant change was observed in the ABA-NoTCC group. These results indicate that the participants got into the state of over-trust in the B condition and that TCCs successfully promoted the participants in the ABA-TCC group to calibrate their trust properly. Similarly, the results of the BAB scenario indicate that the participants under-trusted the system in the A condition and only the participants with TCCs managed to adjust their trust. These results support hypothesis **H0**. Note that the manual rate for the second B condition in NoTCC-BAB group were significantly higher than that for the A condition. This implies that the 16 tasks would be enough for the participants to learn the system and the environment so that they could calibrate their trust better.

Regarding the performance, the sensitivity d' for the first A condition of the NoTCC-ABA group significantly dropped, while that of the TCC-ABA group did not change significantly. The sensitivity d' of the NoTCC-BAB group showed a significant difference between the first B condition and the A condition, while that of the TCC-BAB group stayed at a higher level with no significant change. These results indicate that the participants in the groups with TCCs performed better and more robustly; therefore, hypothesis **H1** was confirmed to be true.

Although the reliability information was continuously displayed with the indicator, the participants of the NoTCC groups did not significantly change their choice behaviors at the first change in weather when the automatic reliability greatly deteriorated. In contrast to this, the participants of the TCC groups successfully altered their choice behaviors accordingly at the first change in weather. The results for the NoTCC groups were not in line with the previous studies that emphasized the importance of continuous system transparency. One possible interpretation is that it might not have been easy for the participants to rectify an improper trust status once they fell into the categories of over-trust or under-trust. Adams, Bruyn, and Houde (2003) suggests that calibration can only occur in response to new evidence that may change the users' prevailing recognition, while no new evidence can be learned without changing the current behavior first. The TCCs successfully played the role of a new trigger to solve this cognitive dilemma (Llinas, Bisantz, Drury, Seong, & Jian, 1999). TCCs were presented adaptively to the trust calibration status so that it would be easier for the participants to understand the implication of the cues. We believe that the results demonstrate the effectiveness of the adaptive presentation method and confirmed hypothesis **H2**.

Finally, the several limitations of our study suggest the need for further experiments and future research. The current framework focuses on performance-related factors to detect over-trust and under-trust. However, automation could be beneficial beyond providing better performance, such as to faster task completion, lighter workloads, and fewer risks. For example, Naujoks, Wiedemann, and Schö mig (2017) discussed the desire to do non-driving-related tasks during autonomous driving, which leads the driver to select autonomous mode. The proposed framework could be integrated with such factors by considering the utilities of choices. The second terms of the framework $P_{man} \gtrless P_{auto}$ could be replaced with $EU(auto) \gtrless EU(man)$, where $EU(x) = U_s(x) * P(x) + U_f(x) * (1 - P(x))$, $P(x)$ is either P_{auto} or P_{man}, and $U_s(x)$ and $U_f(x)$ are the utility functions of choice x if a result is a success and a failure, respectively. Further research should be done to investigate ways to define these utility functions. In the proposed detection algorithm, a binary decision is made with a simple moving average value of three CKPs. Future research should explore a different way of representing the over-trust or under-trust status, such as defining the status as a probability depending on the degree of over- or under-trust. Future research should explore different task difficulties or complexities as well as different types of tasks, such as autonomous driving, decision aids, and interactive games. In the current experiment, a simple pop-up dialogue was used to observe the participants' behavior. Further studies should investigate the continuous measurement of behaviors that could work well with real-time tasks. For example, a driver's intention to use automatic driving could be inferred with a touch sensor on a steering wheel to check if the driver's hands are on the wheel. Further experiments should be done to evaluate our method with different types of TCCs as well as different presentation timings to investigate the requirements of effective cues.

Conclusion

Previous studies on trust calibration mainly examined the factors contributing to system transparency. Not many studies provided a practical model of trust calibration. In the current study, we investigated a method to detect the improper status of trust calibration and notify the users of the miscalibration. We proposed a formal framework to define the status of trust calibration with a performance-centric view of trust, which makes it possible to measure the calibration status by observing human behaviors. We also examined giving a cognitive cue to notify the human users of miscalibration status. We developed a method of adaptive trust calibration by combining the framework with the cognitive cue. The empirical study results demonstrated that the proposed method successfully detected the miscalibration and helped the participants change their behaviors to achieve better performances. Although we used a simple image screening task in the evaluation, the proposed method, which is based on the task-independent framework, could be applied to other collaborative applications of human and autonomous systems. Despite several limitations, we believe that our proposed method could contribute to better interaction designs for collaboration with autonomous systems.

Acknowledgments

References

Adams, B. D., Bruyn, L. E., & Houde, S. (2003). Trust in automated systems literature review. *DRDC Toronto No. CR-2003-096. Defence Research and Development Canada*.

Chambers, A. B., & Nagel, D. C. (1985). PILOTS OF THE FUTURE : HUMAN OR COMPUTER ? *Communication of the ACM*, *28*(11).

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 251–258).

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). A design methodology for trust cue calibration in cognitive agents. In *Proceedings of the International Conference on Virtual, Augmented and Mixed Reality* (pp. 251–262).

de Visser, E. J., Peeters, M. M., Malte, F. J., Kohn, S., Tyler, H. S., Pak, R., & Neerincx, M. A. (2019). Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, 1–20.

Gac, J., & Lee, J. D. (2006). Extending the decision field theory to model operators' reliance on automation in supervisory control situations. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, *36*(5), 943–959.

Geirhos, R., Medina Temme, C. R., Rauber, J., Schü tt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS '18)* (pp. 7549–7561).

Haeuslschmid, R., Buelow, M. V., Pfleging, B., & Butz, A. (2017). Supporting Trust in Autonomous Driving. In *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 319–329). Limassol, Cyprus. doi: 10.1145/3025171.3025198

Helldin, T. (2014). *Transparency for Future Semi-Automated Systems*. Unpublished doctoral dissertation, Orebro University.

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 210–217).

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors*, *58*(3), 509–519.

Hoff, K., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), 407–434. doi: 10.1109/MIS.2013.24

Laughery, K. R., & Wogalter, M. S. (2014). A three-stage model summarizes product warning and environmental sign research. *Safety Science*, *61*, 3–10.

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

Llinas, J., Bisantz, A., Drury, C., Seong, Y., & Jian, J. Y. (1999). Studies and analyses of aided adversarial decision making. Phase 2: Research on human trust in automation. *AFRL-HE-WP-TR-1999-0216*.

McGuirl, J. M., Sarter, N. B., John M. McGuirl, & Nadine B. Sarter. (2006, dec). Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors*, *48*(4), 656–665. Retrieved from `http://www.sagepub.com/journalsPermissions.navhttp://journals.sagepub.com/doi/10.1518/001872006779166334` doi: 10.1518/001872006779166334

Muir, B. M. (1987). Trust between humans and machines. *International Journal of Man-Machine Studies*, *27*, 527–539.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922.

Naujoks, F., Wiedemann, K., & Schö mig, N. (2017). The importance of interruption management for usefulness and acceptance of automated driving. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 254–263).

NHTSA. (2017). Automatic vehicle control systems - investigation of Tesla accident. *National Highway Traffic Safety Administration, PE 16-007*, 13.

Okamura, K., & Yamada, S. (2018). Adaptive trust calibration for supervised autonomous vehicles. In *Adjunct Proceedings of the 10th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 92–97).

Parasuraman, R., & Dietrich H. Manzey. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in Close Relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112. doi: 10.1037/0022-3514.49.1.95

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors*, *58*(3), 377–400. doi: 10.1177/0018720816634228

Stanislaw, H. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. Retrieved from `https://link.springer.com/content/pdf/10.3758{\%}2FBF03207704.pdf` doi: 10.3758/BF03207704

The Cesium Consortium. (2018). *CesiumJS - Geospatial 3D mapping and virtual globe platform.* Retrieved from `http://cesiumjs.org`

Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T. M., Dixon, C., & Tielman, M. L. (2020). Taxonomy of trust-relevant failures and mitigation strategies. *ACM/IEEE International Conference on Human-Robot Interaction*, 3–12. doi: 10.1145/3319502.3374793