Adjunct Proceedings of the 10th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18), September 23–25, 2018, Toronto, Canada.

Adaptive Trust Calibration for Supervised Autonomous Vehicles

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

Seiji Yamada
National Institute of Informatics and SOKENDAI
Tokyo, Japan
seiji@nii.ac.jp

ACM.
AutomotiveUI '18 Adjunct,, September 23–25, 2018, Toronto, ON, Canada
ACM 978-1-4503-5947-4/18/09.
https://doi.org/10.1145/3239092.3265948

Abstract
Poor trust calibration in autonomous vehicles often degrades total system performance in safety or efficiency. Existing studies have primarily examined the importance of system transparency of autonomous systems to maintain proper trust calibration, with little emphasis on how to detect over-trust and under-trust nor how to recover from them. With the goal of addressing these research gaps, we first provide a framework to detect a calibration status on the basis of the user's behavior of reliance. We then propose a new concept with cognitive cues called trust calibration cues (TCCs) to trigger the user to quickly restore appropriate trust calibration. With our framework and TCCs, a novel method of adaptive trust calibration is explored in this study. We will evaluate our framework and examine the effectiveness of TCCs with a newly developed online drone simulator.

Author Keywords
Trust Management; Trust Calibration; Supervised Autonomous Vehicle

CCS Concepts
•Human-centered computing → HCI theory, concepts and models;

Figure 1: Driver-less shuttle (c) Rama, Cc-by-sa-2.0-fr

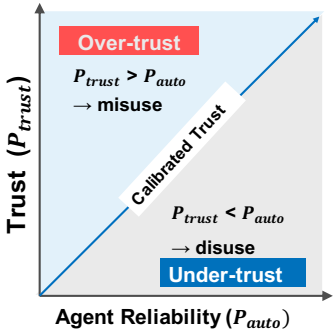

Figure 2: Over- and under-trust

Introduction

Unmanned autonomous vehicle services supervised by human operators are receiving increased attention. Applications of such vehicles includes driver-less shuttles (Figure 1) to transport people at school campuses or office parks and unmanned aerial vehicles for aerial images, delivery, and military purposes. Until perfectly automated technologies are realized, human interventions are inevitable. One key aspect of such interventions is the operators' trust in the autonomous agent. While the agent's reliability can change due to various reasons in the vehicle's environment, the operators sometimes fail to calibrate their trust in the agent accordingly and will fall into the category of over-trust or under-trust. The poor trust calibration often leads to serious safety issues [19].

Trust Calibration

Calibration of trust has been defined as "the correspondence between the person's trust in the agent and the agent's capabilities" [14]. When the well-calibrated trust is achieved without any over-trust or under-trust, the total human-agent performance will be safely maximized. Extensive research has been conducted examining the factors that influence a human operator's trust in automation [7, 10, 22]. Many studies [4, 15] have emphasized the importance of system transparency to maintain proper trust calibration. Studies on visualizing car uncertainty during automated driving [8, 11] have indicated that providing good transparency by constantly presenting the system information is important to maintain continuous trust calibration. Nevertheless, there is still a possibility of poor trust calibration resulting in over-trust or under-trust. Yanco et al. [21] proposed a model to measure the evolution of trust over the use of a system. Few studies, however, have focused on how to detect if the calibration is appropriate or not, how to swiftly recover from over-trust or under-trust.

Adaptive Trust Calibration

This study will focus on the problem of over-trust or under-trust by exploring the following two research questions;1) Can we **detect** if the user is over-trusting or under-trusting the agent ? and 2) How can we assist the user to **promptly recover** from over-trust or under-trust ?

1) A Framework to Detect Over-trust and Under-trust

We propose a framework to detect the status of the trust calibration on the basis of the reliance behavior of the user. Suppose we have a scenario of human-agent collaboration, in which a set of tasks needs to be done manually by a user or automatically by an agent. The user should make successive decisions whether to rely on the agent or do each task manually. In our framework, three parameters P_{auto}, P_{trust}, and P_{self} are defined as follows:

- P_{auto}: Probability of the successful result of the task done by the agent. This is called "reliability of the agent.".
- P_{trust}: User's estimation of P_{auto}. This is the user's trust in the agent.
- P_{self}: User's self-confidence. This is the trust the user has in their own ability to perform the task manually.

P_{auto} varies depending on the various conditions of the agent. P_{trust} also changes accordingly and quickly becomes equal to P_{auto} if trust calibration is performed appropriately. As in Figure 2, over-trust occurs if $P_{trust} > P_{auto}$, and under-trust occurs if $P_{trust} < P_{auto}$. Although these are straightforward definitions, it is difficult to measure P_{trust} without explicitly asking the user, since it is conceptually defined by describing the internal state of the user. To make the status of trust calibration measurable, we define over-trust and under-trust using the third parameter P_{self} in addition to P_{trust} and P_{auto} as follows:

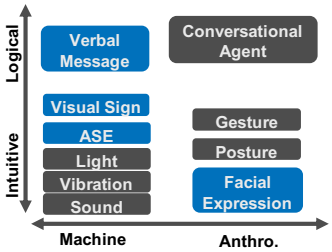

Figure 3: Possible TCCs

Visual-sign TCC
A red warning sign in the shape of reverse triangle [1].
Audible TCC
Sound-based artificial subtle expression [9], which can convey the confidence level of the agent.
Anthropomorphic TCC
A drone animation with a cartoon-like face parts (See Figure 6) to show the agent's state.
Verbal TCC
A warning text message displayed as a tooltip when the Yes button is selected.

Table 1: 4 TCCs to be evaluated in our experiment

Over-trust: the user's estimation of the agent reliability is higher than the user's self-confidence even though the actual reliability is lower than the self-confidence.

$$(P_{trust} > P_{self}) \wedge (P_{self} > P_{auto}) \quad (1)$$

Under-trust: the user's estimation of the agent reliability is lower than the user's self-confidence even though the actual reliability is higher than the self-confidence.

$$(P_{trust} < P_{self}) \wedge (P_{self} < P_{auto}) \quad (2)$$

Several studies [5, 6, 22] demonstrated that the reliance behavior can be explained by the relationship of the user's trust in the agents and the user's self-confidence. When the user makes a decision to rely on an agent, it is reasonable to say that this behavior indicates $Ptrust > Pself$. If the user decides not to rely on the agent but to do the task manually, that indicates $Ptrust < Pself$. In this way, by observing the user's reliance behavior, the trust calibration status can be measured, if the sign of the value $P_{self} - P_{auto}$ can be estimated. While [9, 20] have investigated gaze behavior as a trust measure which seems to be versatile, we focus on the reliance behavior because it is much easier to monitor and valid for our current target.

2) Trust Calibration Cues for a Prompt Recovery

The recovery from over-trust or under-trust is inherently difficult because the user's decision is based on what is just recognized, which the user believes is reasonable. In this study, we will explore a new idea of giving a cognitive cue to the user when over-trust or under-trust is detected. This cue is expected to trigger the user to promptly notice what has been happening in the environment and to calibrate the trust on the basis of the new findings. We call this cognitive cue a "trust calibration cue" (TCC). Based on the findings from the studies on trust [3, 4, 12] as well as the research on warning messages [1, 13], we have categorized possible TCCs into two axes (Figure 3); one is from intuitiveness to logical and the other is from mechanical to anthropomorphic. Table 1 shows the four specific TCCs to be evaluated in our upcoming experiment. We expect that Audible TCC and Visual-sign TCC are to be intuitive, Anthropomorphic TCC is to be familiar, and Verbal TCC is to be used as a baseline in a control group. These TCCs will be evaluated in terms of time sensitivity [16] and accuracy of the calibrations.

Experiment Setup

Scenario

We have designed an experimental scenario of a supervised drone, which can automatically detect potholes on road surfaces with its onboard camera. A participant of the experiment is asked to fly the drone along a predefined route toward a goal. The drone will occasionally make reports on potholes and the participant should make decisions whether to rely on the drone's automatic report or to manually check the road image, because the reliability of the drone's pothole detection can fluctuate depending upon the conditions of weather and sunshine.

Simulation Testbed

We have developed a 3D drone simulator based on an open-source JavaScript WebGL library CesiumJS [2] and Bing Map API [17]. A screen image of the simulator is shown in Figure 4. The left-hand pane of the screen displays the on-board camera image, and the right-hand pane shows a 3D drone on the navigation map with flight indicators. If the drone comes close enough to one of the predefined points on the route, a message pops up (Figure 5) in which the drone notifies the participant if there are any road potholes within the area around the checkpoint. The participants need to select Yes to accept the report or No to check the image by themselves. All the actions of the participants are recorded with timestamps.

Figure 5: Pop up message

Figure 6: Anthropomorphic TCC

Figure 4: A screen capture of the 3D drone simulator

Assumptions

We assume P_{auto} can be calculated on the basis of the internal sensor's sensitivity. Since the robustness of human image recognition is higher than that of the agent's recognition, P_{auto} would fluctuate more widely than P_{self} with changes in the weather conditions. These assumptions make it possible to estimate the sign of $P_{self} - P_{auto}$.

Procedure

60 participants on the web are invited to fly a drone over a 20-km route in a rural area in Hokkaido, Japan, along National Route 12. The experiment will be performed in the following three phases. In **Phase 1**, the participants learn to use the drone simulator. They are explicitly told that the average success rate of manual pothole detection is 75%, so that they can adjust their initial self-confidence P_{self} accordingly. Next, in **Phase 2**, all participants are intentionally trained to strongly trust the drone at a reliability of P_{auto}=100%, since we focus on the over-trust issue in this experiment. After the manipulation checks of P_{trust} and P_{self}, **Phase 3**, the main part of the experiment, is performed. The participants are divided into the four groups corresponding to one of the TCCs described in Table 1. The P_{auto} is artificially decreased from 100% to 50%, which changes the sign of $P_{self} - P_{auto}$. The participants are expected to become over-trusting the drone. If the participants select Yes on the pop-up notification windows, the chi-square frequency test is used to determine if the counts of the observed behavior significantly differ from the one that we would expect by chance. If this test passes, the participants are judged to be in the over-trust status. Right after the detection, the corresponding TCC is presented in each group. In each experiment, the drone will make 30 reports regarding checkpoints randomly selected from 63 predefined ones on the route. After finishing Phase 3, participants will be asked to complete a trust survey. We will evaluate the effectiveness of our framework with the detection results in the first half of Phase 3, and examine how TCCs influence the reliance behaviors in the last half of Phase 3. The independent variable is a TCC with four levels, and the dependent variable is the selection behavior.

Conclusion

The experiment will be performed soon using Yahoo cloud sourcing. By examining our proposed adaptive trust calibration with the drone experiment, we hope to clarify what is necessary to quickly recover from the situations that result from poor trust calibration. Although there are many assumptions made, we still believe the finding of this study will contribute to better user-interface designs for supervised autonomous vehicles. We also expect that further investigation on other important factors [18] in autonomous driving, such as reliance behaviors of drivers with non-driving-related activities, will help to enhance our framework to be applicable to autonomous vehicle in general.

Acknowledgements

This study was partially supported by JSPS KAKENHI "Cognitive Interaction Design" (No. 26118005).

REFERENCES

1. Mark A Changizi, Matt Brucksch, Ritesh Kotecha, Kyle McDonald, and Kevin Rio. 2014. Ecological warnings. *Safety Science* 61, C (Jan. 2014), 36–42. DOI: http://dx.doi.org/10.1016/j.ssci.2013.07.012
2. The Cesium Consortium. 2018. CesiumJS - Geospatial 3D Mapping and Virtual Globe Platform. (2018). http:://cesiumjs.org
3. Andrew J Cowell and Kay M Stanney. 2005. Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. *International Journal of Human-Computer Studies* 62, 2 (Feb. 2005), 281–306. DOI: http://dx.doi.org/10.1016/j.ijhcs.2004.11.008
4. Eward J. de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A design methodology for trust cue calibration in cognitive agents. *Proceedings of the International Conference on Virtual, Augmented and Mixed Reality* (2014), 251–262. DOI: http://dx.doi.org/10.1007/978-3-319-07458-0_24
5. Peter de Vries, Cees Midden, and Don Bouwhuis. 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies* 58, 6 (2003), 719–735. DOI: http://dx.doi.org/10.1016/S1071-5819(03)00039-9
6. Ji Gao and John D Lee. 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 (2006), 943–959. DOI: http://dx.doi.org/10.1109/TSMCA.2005.855783
7. Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors* 53, 5 (2011), 517–527. DOI: http://dx.doi.org/10.1177/0018720811417254
8. Tove Helldin, Goran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2013), 210–217. DOI: http://dx.doi.org/10.1145/2516540.2516554
9. Sebastian Hergeth, Lutz Lorenz, Roman Vilimek, and Josef F. Krems. 2016. Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors* 58, 3 (May 2016), 509–519. DOI: http://dx.doi.org/10.1177/0018720815625744
10. Kevin Hoff and Masooda Bashir. 2015. Trust in Automation：Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. DOI: http://dx.doi.org/10.1177/0018720814547570
11. Malte F Jung, David Sirkin, Turgut M Gür, and Martin Steinert. 2015. Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems* (2015), 2201–2210. DOI: http://dx.doi.org/10.1145/2702123.2702479

12. Takanori Komatsu, Seiji Yamada, Kazuki Kobayashi, Kotaro Funakoshi, and Mikio Nakano. 2010. Artificial Subtle Expressions: Intuitive Notification Methodology of Artifacts. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems* (2010), 1941–1944. DOI: http://dx.doi.org/10.1145/1753326.1753619
13. Kenneth R Laughery and Michael S Wogalter. 2014. A three-stage model summarizes product warning and environmental sign research. *Safety Science* 61, C (Jan. 2014), 3–10. DOI: http://dx.doi.org/10.1016/j.ssci.2011.02.012
14. John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.
15. Joseph B Lyons, Garrett G Sadler, Kolina Koltai, Henri Battiste, Nhut T Ho, Lauren C Hoffmann, David Smith, Walter Johnson, and Robert Shively. 2017. *Shaping trust through transparent design: theoretical and experimental guidelines*. Vol. 499. Springer. 127–136 pages. DOI: http://dx.doi.org/10.1007/978-3-319-41959-6_11
16. Stephanie M Merritt, Kelli Huber, Jennifer LaChapell-Unnerstall, and Deborah Lee. 2014. Continuous calibration of trust in automated systems. *Air Force Research Laboratory Technical Report* (2014). DOI:http://dx.doi.org/10.21236/ADA606748
17. Microsoft. 2018. Bing Maps API Documentation. (2018). https://www.microsoft.com/en-us/maps/documentation
18. Brittany E Noah, Philipp Wintersberger, Alexander G Mirnig, Shailie Thakkar, Fei Yan, Thomas M Gable, Johannes Kraus, and Roderick McCall. 2017. First Workshop on Trust in the Age of Automated Driving. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct* (Sept. 2017), 15–21. DOI: http://dx.doi.org/10.1145/3131726.3131733
19. Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. *Proceedings of the ACM/IEEE International Conference on Human Robot Interaction* (2016), 101–108.
20. Francesco Walker, Willem Verwey Verwey, and Marieke Martens. 2018. Gaze Behaviour as a Measure of Trust in Automated Vehicles. *Proceedings of the the 6th Humanist Conference* (June 2018).
21. Holly A Yanco, Munjal Desai, Jill L Drury, and Aaron Steinfeld. 2016. *Methods for developing trust models for intelligent systems*. Springer. 219–254 pages. DOI: http://dx.doi.org/10.1007/978-1-4899-7668-0_11
22. Jessie X Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (2017), 408–416. DOI: http://dx.doi.org/10.1145/2909824.3020230