Driving Assessment Conference

2015 Driving Assessment Conference

Jun 25th, 12:00 AM

Effects of Take-Over Requests and Cultural Background on Automation Trust in Highly Automated Driving

Sebastian Hergeth
BMW Group Research and Technology, Munich, Germany

Lutz Lorenz
BMW Group Research and Technology, Munich, Germany

Josef F. Krems
Chemnitz University of Technology, Chemnitz, Germany

Lars Toenert
BMW China Services, Beijing, China

Follow this and additional works at: http://ir.uiowa.edu/drivingassessment

EFFECTS OF TAKE-OVER REQUESTS AND CULTURAL BACKGROUND ON AUTOMATION TRUST IN HIGHLY AUTOMATED DRIVING

Sebastian Hergeth[1,2], Lutz Lorenz[1], Josef F. Krems[2], Lars Toenert[3]
[1]BMW Group Research and Technology, Munich, Germany
[2]Chemnitz University of Technology, Chemnitz, Germany
[3]BMW China Services, Beijing, China
Email: Sebastian.Hergeth@bmw.de

Summary: Appropriate automation trust is a prerequisite for safe, comfortable and efficient use of highly automated driving systems (HADS). Earlier research indicates that a drivers' nationality and Take-Over Requests (TOR) due to imperfect system reliability might affect trust, but this has never been investigated in the context of highly automated driving. A driving simulator study ($N = 80$) showed that TORs only temporarily lowered trust in HADSs, and revealed similarities in trust formation between German and Chinese drivers. Trust was significantly higher after experiencing the system than before, both for German and Chinese participants. However, Chinese drivers reported significantly higher automation mistrust than German drivers. Self-report measures of automation trust were not connected to behavioral measures. The results support a distinction between automation trust and mistrust as separate constructs, short- and long-term effects of TORs on automation trust, and cultural differences in automation trust.

INTRODUCTION

Highly automated driving systems (HADS) can increase driving comfort, safety, and efficiency, but only if drivers trust them appropriately. They provide longitudinal and lateral vehicle control in certain conditions and for a limited amount of time (Trimble, Bishop, Morgan, & Blanco, 2014). During that time, the driver does not have to monitor the vehicle and can fully engage in non-driving related tasks (NDRT). However, the driver is required to take over vehicle control if requested by the system. These take-over requests (TOR) are necessary because perfect reliability of HADSs is virtually impossible, for example due to the environmental complexity involved or sensor failures. But how are TORs perceived by the driver, and how does that affect trust in HADSs?

TORs indicate that the system cannot continue to provide complete vehicle control. On the one hand, they could be interpreted as automation failures. Trust is best understood as a dynamic process, and automation failures typically decrease trust (Lee & See, 2004). On the other hand, drivers might perceive TORs as warnings. This makes an important difference: Lees and Lee (2007) proposed a differentiation of warning types along *purpose*, *process* and *performance* dimensions. In a study of automotive collision warning systems, they showed that intended, comprehensible and useful warnings fostered trust. Studies on TORs (e.g. Gold, Damböck, Lorenz, & Bengler, 2013) have modeled TORs exactly as such accurate warnings. People tend to rely on automation they trust, and disregard automation they do not trust (Chien, Lewis, Semnani-Azad, & Sycara, 2014). So far, the effects of TORs on automation trust are unclear.

Among other variables that might affect automation trust, cultural differences have been

frequently cited (Chien et al., 2014). For example, Sherman, Helmreich, and Merritt (1997) found significant cultural differences amongst pilots in attitudes regarding flight deck automation. They hypothesized that pilots from more hierarchical national cultures (e.g. asian countries) may be more favorable towards automation than pilots from more individualistic societies (e.g. western countries). This emphasizes that findings regarding trust in automation should be validated when they are transferred from one culture to another (Lee & See, 2004), especially when comparing Western drivers such as Germans to Asian drivers such as Chinese.

The goal of this study was to investigate the effects of TORs and cultural background on automation trust in highly automated driving. We hypothesized (i) that TORs would affect automation trust and (ii) that Chinese drivers would be more inclined to trust HADS than German drivers.

METHOD

Participants

Ninety-one employees of the BMW Group voluntarily participated in the study. Due to simulation errors during the experiment, 11 German participants were excluded from the analysis and replaced, resulting in the planned sample size of $N = 80$.

National subsamples were compared along several potential confounding factors (see Beggiato & Krems, 2013). A Chi-square test for independence indicated no significant association between nationality and gender balance, $\chi^2(1) = 0.82$, $p = .366$, *odds ratio* $= .66$. There was no significant age difference between the German and Chinese participants, $t(55.31) = -1.87$, $p = .067$, $r = .24$. There was also no significant difference in experience with advanced driver assistance systems between German and Chinese participants, $t(77.91) = 1.40$, $p = .165$, $r = .16$. German participants did report significantly more driving experience (years of owning a drivers license) than Chinese participants, $t(52.28) = 2.79$, $p = .007$, $r = .36$. In practical terms however, the absolute difference was rather unsubstantial (4.07 years).

Table 1. Participants

	n participated	*n* considered	Gender ♂	Gender ♀	Age M	Age SD	ADAS Experience M	ADAS Experience SD	Driving Experience M	Driving Experience SD
German	51	40	21	19	27.83	8.66	3.28	1.72	10.31	8.51
Chinese	40	40	25	15	30.65	4.05	2.73	1.78	6.24	3.57

Note. ADAS = Advanced Driver Assistance Systems, six-point scale (1 = Not at all; 6 = Very often)

Experimental Design

The study employed a two-factor mixed between-within design, with cultural background as the between-subjects factor (Chinese; German). Each participants' trust was measured repeatedly over the course of the experiment (before and after the experimental session, and eight times during the experimental session), with time of measurement forming the within-subjects factor. As a self-report measure of automation trust, we used the automation trust scale developed by Jian, Bisantz, and Drury (2000). This 12-item questionnaire has been adopted successfully in studies with German (Beggiato & Krems, 2013) and Chinese (Ritz, 2004) participants.

Additionally, we used single-item automation trust ratings (e.g. Brown & Galster, 2004) to assess drivers' trust during the experimental session. As behavioral measures of automation trust, resumption lags and take-over times were recorded. Rajaonah, Anceaux, and Vienne (2006) found that trust in an active cruise control was positively correlated with the time spent using the device. Conversely, low trust in a HADS might be related to the time *not* spent using the device. Borrowing a term from multitasking, resumption lags were defined as the time between a TOR and the re-activation of the HADS by the driver. Regarding take-over times, studies with rear-end collision warning systems indicate that reaction times to warnings extend with increasing trust (Abe, Itoh, & Tanaka, 2002). Take-over time was defined as the first manual braking or steering input after a TOR exceeding 2^0 steering wheel angle or 10% braking pedal position.

Apparatus

The study was conducted in BMW Group laboratories in Germany and China. The static driving simulator had six visual channels including rear visibility (see Figure 1).The three forward channel Plasma monitors, each at a resolution of 1920 × 1080 and with 127 cm screen size, were rendered at 60 frames/s and provided a horizontal field of view of 78°. A display with the same specifications right behind the vehicle's rear seats provided an image for the rearview mirror. The two side mirror rear channels accommodated 800x600 TFT displays.

Figure 1. Simulator Setup

Figure 2. Simulator Interior

Procedure

Data were collected in single ninety-minute experiments. The virtual driving scenario for all sessions was a standard three-lane freeway with a hard shoulder. At the beginning of each experiment, participants were briefed on the driving simulator and the HADS. The HADS provided lateral and longitudinal control, including lane changes and overtaking. Participants were told that the system would not require monitoring during highly automated driving and that any necessary driver intervention would be announced with sufficient time to react by a TOR (combined sinusoidal sound and visual icon). Manual braking or steering shut off the automation. To create a real-world situation with drivers engaging in NDRTs during highly automated driving, a modified Version of the Surrogate Reference Task (ISO/TS 14198, 2012) was presented on a tablet mounted in the center console (see Figure 2). At the beginning of each NDRT, the experimenter asked participants to rate their Single-Item Automation Trust in the HADS ("On a scale from 0% to 100%, how much do you trust the system?"). In a training session, participants were familiarized with the simulator, the HADS, the NDRT, and the TOR. Participants then completed an a priori Automation Trust Scale questionnaire. In the following

experimental session, the first single-item automation trust rating was collected after 2 minutes of highly automated driving. Immediately afterwards, the NDRT was presented for 45 seconds. This process was repeated intermittently every 2½ minutes, resulting in a total of eight times of measurement (see Figure 3). During the second presentation of the NDRT, approximately 5½ minutes into driving, a suddenly occurring accident in the cars' own lane triggered the first TOR (time to collision = 7s). After clearing the accident, participants reactivated the HADS. If they failed to do so, the experimenter prompted them after 45 seconds. A second TOR occurred during the fourth presentation of the NDRT. The situation was identical in both TORs and varied only in traffic density, counterbalanced between participants. During the subsequent last four NDRTs, no TORs occurred. The sequence of events of TORs and NDRTs is displayed in Figure 3. After the experimental session, participants filled out the Automation Trust Scale for a second time, some scales that are outside the scope of this report, and a demographic questionnaire. The experiment concluded with an open interview to collect additional information.

RESULTS

Self-Report Measures

Single-item automation trust ratings. A mixed between-within subjects analysis of variance was conducted   to assess the effect of cultural background on single-item automation trust ratings across time of measurement (see Figure 3). Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of time of measurement, $\chi^2(27) = 288.56$, $p <$ .001. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .49$). Results showed a significant main effect of time of measurement on single-item automation trust ratings, $F(3.40, 264.91) = 9.23$, $p < .001$, $r = .19$, indicating that ratings changed during the experimental session. Contrasts revealed that both the first TOR, $F(1, 78) = 4.73$, $p = .033$, $r = .24$, and second TOR $F(1,78) = 6.74$, $p = .011$, $r = .28$, significantly lowered single-item automation trust ratings. In addition, single-item automation trust ratings at the last time of measurement were significantly higher than at the first time of measurement, $F(1, 78) = 36.01$, $r = .56$, $p < .001$. This means that automation trust increased during the experimental session, and decreased only temporarily after TORs. There was no significant main effect of cultural background on single-item automation trust ratings, $F(1, 78) = 0.13$, $p = .719$, $r = .00$. There was also no significant interaction effect between the time of measurement and cultural background on single-item automation trust, $F(3.40, 264.91) = 0.36$, $p = .812$, $r = .04$.

Automation Trust Scale. To evaluate the impact of cultural background on participants' automation trust scale ratings before and after the experimental session, mixed between-within subjects analysis of variance were performed. In line with empirical evidence (Spain, Bustamante, & Bliss, 2008), trust as measured by the Automation Trust Scale was treated as a multi-dimensional construct. Separate scores were calculated for the subscales automation trust and automation mistrust. Results showed a significant main effect for time of measurement on automation trust, with trust being significantly higher after the experimental session than before, $F(1,78) = 7.08$, $p = .009$, $r = .29$. There was no significant main effect of cultural background on automation trust ratings, $F(1, 78) = 2.66$, $p = .107$, $r = .18$. There was also no significant interaction effect between the time of measurement and cultural background on automation trust, $F(1, 78) = 0.73$, $p = .397$, $r = .07$.

For *mis*trust ratings, there was a significant main effect of cultural background, $F(1, 78) = 15.51$, $p < .001$, $r = .41$. Automation mistrust ratings of Chinese participants were significantly higher than those of German participants, both before and after the experimental session (see left side of Figure 4). Results showed no significant main effect of the time of measurement on automation mistrust, $F(1,78) = 0.01$, $p = .921$, $r = .01$. There was also no significant interaction effect between the time of measurement and cultural background on automation mistrust, $F(1, 78) = 2.01$, $p = .160$, $r = .16$.
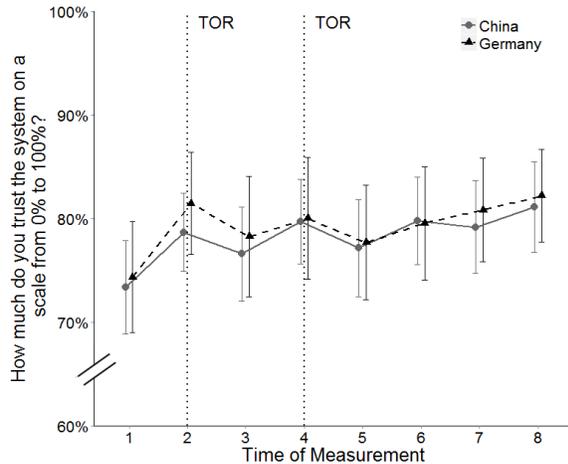

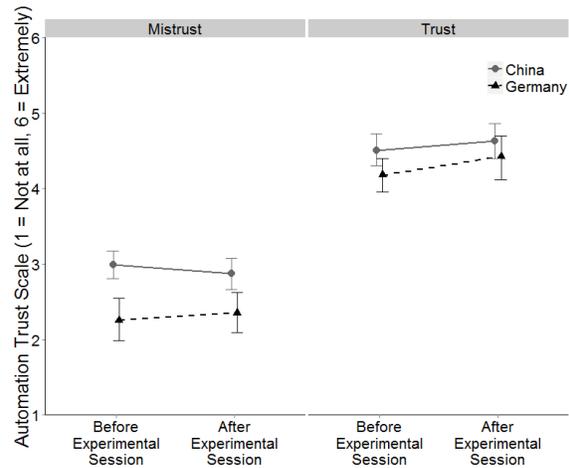**Figure 3. Single-item trust ratings during the experimental session**

Figure 4. Automation trust scale ratings before and after the experimental session

Behavioral Measures

Single-item automation trust ratings did not significantly predict *take-over times* for Chinese, $b = 2.93$, $t(78) = -.18$, $p = .855$, or German participants, $b = 4.24$, $t(78) = -1.25$, $p = .215$. Single-item automation trust ratings could not explain a significant proportion of variance in take-over times for the Chinese, $R^2 = .03$, $F(1, 78) = .00$, $p = .855$, or German participants, $R^2 = .02$, $F(1, 78) = 1.56$, $p = .215$.

Resumption lags longer than 45 seconds were excluded from analysis (China = 25, Germany = 28), because participants were prompted to reactivate the system at that time after a TOR. Single-item automation trust ratings preceding TORs could not significantly predict subsequent resumption lags for the Chinese, $b = 18.62$, $t(63) = -0.06$, $p = .952$, or German participants, $b = 23.62$, $t(60) = 1.04$, $p = .303$. Single-item automation trust did not explain a significant proportion of variance in resumption lags after TORs for the Chinese participants, $R^2 = .00$, $F(1, 63) = .00$, $p = .952$, or German participants, $R^2 = .02$, $F(1, 60) = 1.08$, $p = .303$.

CONCLUSION

The objective of this study was to investigate the effects of TORs and cultural background on trust in HADSs.

Results show that regardless of cultural background, TORs temporarily lowered automation trust (see Figure 3). This indicates that in line with earlier research and theoretical considerations (Lee

& See, 2004), TORs were indeed perceived as automation failures. On the other hand, both single-item automation trust ratings and automation trust scale ratings were significantly higher at the end of the experimental session than in the beginning. These seemingly contradictory findings might be explained if short- and long-term effects of TORs are differentiated. TORs may have undermined trust temporarily because they illustrated that the system was not perfectly reliable. In the long run however, TORs might have helped participants to understand the system, thereby increasing trust. According to Lee & See (2004), trust can be fostered - among else - by observing system behavior (performance dimension) and understanding the underlying mechanisms (process dimension). Another interpretation is provided by Beggiato and Krems (2013), who suggest that automation failures do not decrease trust if they are known in advance. Participants in the current study were instructed beforehand about possible automation failures. Additionally, TORs in the current study were exclusively accurate warnings, which presumably foster trust within the framework of Lees and Lee (2007).

Spain, Bustamante, and Bliss (2008) suggested that when using the system trust scale, trust should be considered as a two-dimensional construct. The current results support this distinction. Although not significant, there seems to be a difference in trust formation between German and Chinese participants: While German participants' mistrust slightly increased during the experimental session, Chinese participants' mistrust decreased. Additionally, Chinese participants' mistrust was significantly higher both before and after the experimental session (see Figure 4). A possible explanation is that the driving strategy of the automation (e.g. headway distances or cooperation with other road users) was evaluated differently. Single-item trust ratings however, were again remarkably similar between cultural backgrounds (Figure 3). It should be kept in mind that since participants were all BMW Group employees, they possibly shared an organizational culture that might have been as influential as their cultural background.

The lacking connection between automation trust and resumption lags (cf. Rajaonah et al., 2006) could be explained by limited variability, as drivers in the current study were required to reactivate the system within a given timeframe. Since there was a tight time window for driver intervention after a TOR to avoid an accident, the same explanation could apply to take-over reactions. In addition, Itoh, Abe,and Tanaka (1999) hypothesized that decisions are made quickly both if an operator trusts or distrusts an automation strongly. These effects might have cancelled each other out and blurred the relation between self-report measures of trust and take-over times.

The current study has extended earlier findings on automation trust to HADSs. Above that, it has contributed to the understanding of cultural differences in automation trust, and investigated select behavioral measures of automation trust. The findings suggest a differentiation between automation trust and mistrust, short- and long-term effects of TORs on automation trust, and cultural differences in automation trust. Future studies should investigate how other types of TORs, for example incomprehensible ones, affect trust in HADSs. It is also unclear how drivers react to TORs if they are not informed about possible automation failures beforehand. In addition, there remains a lack of reliable and valid behavioral measures of trust in highly automated driving. Resumption lags and take-over reaction times should be examined in situations when drivers can choose if and how they act, for example after uncritical TORs. Finally, no study so far has investigated the long-term development of trust in HADS, and whether single-item ratings are also suitable to assess automation *mis*trust.

REFERENCES

Beggiato, M., & Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research Part F: Traffic Psychology and Behaviour*, *18*, 47–57. doi:10.1016/j.trf.2012.12.006

Brown, R. D., & Galster, S. M. (2004). Effects of Reliable and Unreliable Automation on Subjective Measures of Mental Workload, Situation Awareness, Trust and Confidence in a Dynamic Flight Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(1), 147–151. doi:10.1177/154193120404800132

Chien, S. Y., Lewis, M., Semnani-Azad, Z., & Sycara, K. (2014). An Empirical Model of Cultural Factors on Trust in Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*, (1), 859–863.

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). "Take over!" How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1938–1942. doi:10.1177/1541931213571433

ISO/TS 14198 (2012). *Road vehicles - Ergonomic aspects of transport information and control systems - Calibration tasks for methods which assess driver demand due to the use of in-vehicle systems*. International Organization for Standardization, Geneva, Switzerland.

Itoh, M., Abe, G., & Tanaka, K. (1999). Trust in and use of automation: Their dependence on occurrence patterns of malfunctions. *IEEE SMC'99 Conference Proceedings, USA, 3*, 715-720. doi:10.1109/ICSMC.1999.823316

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. doi: 10.1207/S15327566IJCE0401_04

Lees, M., & Lee, J. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics*, *50*(8), 1264–1286. doi:10.1080/00140 130701318749

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. doi:10.1518/hfes.46.1.50

Rajaonah, B., Anceaux, F., & Vienne, F. (2006). Study of driver trust during cooperation with adaptive cruise control. *Le Travail Humain*, *2*(69), 99–127. doi: 10.3917/th.692.0099

Ritz, F. (2004). *Einflussfaktoren auf die Teamleistung im interkulturellen Vergleich*. (Doctoral Thesis, Technische Universität Berlin, Germany). Retrieved from http://opus4.kobv.de/opus4-tuberlin/files/814/ritz_frank.pdf

Sherman, P. J., Helmreich, R. L., & Merritt, A. C. (1997). National Culture and Flight Deck Automation: Results of a Multination Survey. *The International Journal of Aviation Psychology*, *7*(4), 311–329. doi:10.1207/s15327108ijap0704

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an Empirically Developed Scale for System Trust: Take Two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1335–1339. doi:10.1177/154193120805201907

Trimble, T. E., Bishop, R., Morgan, J. F., & Blanco, M. (2014, July). *Human factors evaluation of level 2 and level 3 automated driving concepts: Past research, state of automation technology, and emerging system concepts*. (Report No. DOT HS 812 043). Washington, DC: National Highway Traffic Safety Administration.