Gaze Behavior, Skin Conductance, and Trust in Automation

Jingming Wang

Human Factors & Engineering Psychology

Supervised by:

Verwey, Willem, prof. dr. ing.

Walker, Francesco, MSc.

Abstract

The study investigated the possibility of using gaze behavior and skin conductance to measure trust in automation. Specifically, we divided participants into either Perfect Vehicle Group or Poor Vehicle Group, while in each group, the simulated automated vehicle performed in a perfect or poor way with driving tasks. Gaze behavior and skin conductance were measured throughout the experiment, while self-reported trust was measured after every driving phase. We hypothesized that (1) The Perfect Vehicle Group would have higher self-reported trust than the Poor Vehicle Group; (2) Gaze behavior would be negatively associated with the self-reported trust; (3) Skin conductance level would be negatively associated with the self-reported trust. The results showed that participants had higher self-reported trust in the automated vehicle with perfect performance, and gaze behavior was negatively associated with the self-reported trust. We concluded that gaze behavior appeared to be a reliable indicator of trust in automation.

Keywords: trust in automation, gaze behavior, skin conductance

Gaze Behavior, Skin Conductance, and Trust in Automation

Automated vehicles have many promising benefits. For individual drivers, automated vehicles can exempt them from routine driving tasks and increase their comfort (Gold, Körber, Hohenberger, Lechner, & Bengler, 2015). For society, the wide application of automated vehicles can reduce gas emission and fuel consumption, at the same time, increase traffic efficiency and road safety by reducing human errors, such as distraction and speeding (Gold et al., 2015; Payre, Cestac, & Delhomme, 2014). However, the realization of the benefits mentioned above requires not only sophisticated technology, but also acceptance from individual drivers.

In the Automation Acceptance Model proposed by Ghazizadeh, Lee and Boyle (2012), trust is one of the crucial contributors to the acceptance of automated systems. Trust is a social psychological concept that originally describes interpersonal relationships. A widely accepted definition of trust is provided by Rousseau et al. (1998): "Trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviour of another" (p. 395). In other words, trust is a strategy individuals use to reduce complexity and uncertainty and is tolerant of risk (Earle, 2010).

Previous studies suggest that trust also applies in human-automation relationship. In many cases, the introduction of automation does not replace humans, but instead changes human's role and tasks. For example, the introduction of automation in the cockpit changes the pilot's role from controlling into monitoring. Thus, the performance depends on the cooperation between human and automated system. Previous studies suggest that trust in automation is one of the major contributors to the adoption of automation (e.g., Bailey & Scerbo, 2007; Hoff & Bashir, 2015; Lee & Moray, 1994; Lee & See, 2004). More specifically on the topic of automated vehicles, trust is also found to influence users' intention to use automated vehicles (Choi & Ji, 2015). Moreover, trust not only predicts whether the

automated system will be used, but also *how* it will be used. When users have low levels of trust in automation, they are more likely to disuse the system by underutilizing its capabilities. On the contrary, when the trust level is high or even excessive, they tend to over-rely on the system. For example, users are less likely to monitor the automated system to a necessary degree and have an insufficient level of situation awareness. When the situation exceeds the system's boundary and requires manually operation, users cannot make sense of the situation and thus may need longer reaction time and have poorer responding performance (Carlson, Desai, Drury, Kwak, & Yanco., 2014; Bagheri & Jamieson, 2004; Körber, Baseler, & Bengler, 2018; McGuirl & Sarter, 2006).

Thus, to successfully design the interaction between human and automated system, so that the system could be accepted and used properly, it is important to investigate what influences users' trust and how to build appropriate trust levels. Körber et al. (2018) stresses that trust in automation is influenced by both characteristics of the automated system (e.g., the reliability of the system) and the users themselves (e.g., the tendency of the user to trust). In the case of automated vehicles, most users are laypersons with limited knowledge or experience with automated vehicles, and thus the use of automated vehicles is a novel situation for them (Casner, Hutchins, & Norman, 2016). In these circumstances, the information provided about the reliability of the automated vehicle may have a great impact on users' trust in the automated vehicle (Hoff & Bashir, 2015). Previous studies show that reliance and trust towards automation are influenced by reliability information provided with explicit statements (Madhavan & Wiegmann, 2005), by whether any limitations are mentioned during introduction (Biassoni, Ruscio, & Cicer, 2016), and by the level of reliability of the automated vehicle experienced during training sessions (Sauer, Chavaillaz, & Wastell, 2016). Moreover, the effect is stronger when the information is provided by actual experience in training session than by only verbal instructions about the reliability of the

automated vehicle, which is indicated by the findings that the initial reliance and trust levels established by verbal instructions are vulnerable to later experience during the experiment (Bahner, Hüper, & Manzey, 2008; Sauer et al., 2016).

By far, self-report is the most widely used method to measure trust in automation. Although self-report might be one of the most straightforward and easy methods to measure trust, it has some limitations. For example, self-report cannot capture the temporary changes of trust, and is not viable in many applied settings (Hergeth, Lorenz, & Vilimek, 2016). Gaze behavior and skin conductance are two promising alternatives for self-report.

The first possible alternative is gaze behavior. In highly automated driving, drivers are expected to engage in non-driving-related tasks (NDRTs) instead of monitoring the driving situation continuously. As mentioned above, trust in automation influences users' reliance on automation (e.g., Carlson et al., 2014). When trust is excessive, users should over-rely on the automated system and thus monitor less on the driving situation and focus more on NDRTs. On the contrary, when trust level is low, users should rely less on the system but actively monitor the driving situation by themselves. The link between trust in automation and reliance on automation gives rise to the hypothesis that gaze behavior, an observable indication of reliance, could be used to measure users' trust in automation (Lee & See, 2004; Parasuraman & Manzey, 2010). The evidence for this hypothesis is mixed. Some studies support the hypothesis (e.g. Hergeth et al., 2016; Körber, Baseler, & Bengler, 2018; Walker, Verwey, & Martens, 2018). For example, Hergeth and colleagues (2016) asked participants to perform a visually demanding NDRT during highly automated driving. They found that there was a consistent relationship between gaze behavior and self-reported trust: Participants who reported higher trust also tended to monitor the driving situation less. However, there are also studies which fail to find the link (Gold et al., 2015) or even find the opposite relationship (Helldin, Falkman, Riveriro, & Davidsson, 2013). For instance, Helldin and colleagues (2013)

found that participants who were provided with uncertainty information about the car's ability had lower trust levels towards the automated system and took over control of the car faster when needed than the other group without uncertainty information. Nevertheless, they also spent more time performing other activities (e.g., reading newspapers) than the other group.

Another option is to use skin conductance, which reflects changes in the skin's ability to conduct electricity (Grubin & Madsen, 2005). Changes in skin conductance is not under conscious control but modulated by sympathetic nervous system. One factor that has been associated with skin conductance repeatedly is stress. When people are under stressful situation, the sweat excretion will increase and thus skin conductivity will follow (Costa, Roe, & Taillieu, 2001; Santarchangelo et al., 2012). There are few studies investigating the relationship between skin conductance and trust. Khawaji and colleagues (2015) examined the possibility to use skin conductance as indictors of interpersonal trust since trust was closely associated with stress. They found that skin conductance level was significantly affected by both trust and cognitive load in the text-chat environment and concluded that skin conductance was a promising tool to measure interpersonal trust when cognitive load was low. Akash and colleagues (2018) further evaluated the possibility of using skin conductance to measure trust in automation. Participants were told to evaluate a sensor's performance. The sensor was designed to detect obstacles on the road. Participants had multiple trials with the sensor and were asked to respond with trust or distrust after every trial. Feedback of the sensor's performance would be provided after participants' response. Then researchers built a model of trust using features from EEG and skin conductance response signal. They concluded that physiological measures (EEG and skin conductance) were promising real-time indicators of human trust in machine.

The present study aimed to validate the reliability of these alternative measures, namely gaze behavior and skin conductance, to measure trust in automation. Specifically, the

associations between self-reported trust, gaze behavior, and skin conductance were investigated. The experiment consisted of 3 driving phases and used a 2 (group; between-subject) x 3 (phases; within-subject) mixed design. Participants were divided into two groups, the Perfect Vehicle Group and the Poor Vehicle Group, according to the self-driving car's performance. In all three phases, different driving scenarios were displayed while participants were instructed to imagine themselves being in a self-driving car and to perform the NDRTs only if they trusted the behavior of the self-driving car. In phases 1 and 2, the self-driving car performed differently for two groups (handling driving tasks perfectly for Perfect Vehicle Group, handling driving tasks poorly for Poor Vehicle Group). Phase 3 was designed to investigate whether the developed trust levels would last. Both groups received poor vehicle behavior in this phase. The skin conductance and gaze behavior were measured in all three phases, and self-reported trust was measured after every phase. We hypothesized that (a) The Perfect Vehicle Group would have higher self-reported trust, monitor more on NDRT and less on the road, and have lower skin conductance level than the Poor Vehicle Group in phase 1 and phase 2; (b) The Perfect Vehicle Group would have lower self-reported trust, monitor less on the NDRT and more on the road, and have higher skin conductance level in phase 3 than in phase 2, while the Poor Vehicle Group would remain unchanged; (c) Higher self-reported trust would be related to less monitoring behavior towards driving situation and more monitoring behavior towards NDRTs. In other words, trust would be negatively related to gaze behavior; (d) Higher self-reported trust would be negatively associated with skin conductance level.

Method

Participants

36 participants were recruited to participate in the study and they were rewarded with

study credits or money (6 euros). 10 participants were excluded from analysis (because of the poor data quality either for eye-tracking or skin conductance), resulting in a sample size of 26 (13 per group). The final sample consisted of 10 females, and 16 males, aged from 19-36 (M =24.27, SD = 4.37). They all at least had a driver's license for 2 years (M=5.96, SD=4.18). All participants had no experience with self-driving cars before, and did not wear glasses.

Material and apparatus

Video material. There were two kinds of videos: High reliability and low reliability. High reliability referred to videos in which the self-driving car coped perfectly with driving tasks (i.e., slow down before zebra and leave cyclists and pedestrian a safe distance to crossing), while low reliability referred to videos in which the self-driving car performed poorly with the driving tasks (i.e., drifted toward the center of the road and braked abruptly when there was a cyclist or pedestrian crossing the road). Different videos of the same kind had different driving scenarios but same level of reliability. The video material had been used in a previous study (Walker, Verwey, & Martens, 2018).

The videos were presented in the driving simulator of the University of Twente. Participants were instructed to imagine being in a self-driving car. In phase 1 and phase 2, two videos of high reliability were presented for the Perfect Vehicle Group each phase, while the Poor Vehicle Group watched videos of low reliability. In phase 3, two low reliability videos were presented for both groups. The presentation order of two videos in every phase were counter-balanced across participants.

Driving simulator. The driving simulator of the University of Twente consists of a mock-up equipped with steering wheel, pedals and indicators. Videos were displayed through Psychopy software (Peirce, 2009) and were projected on a screen (7.8 x 1.95 meters, the

resolution is 3072*768 pixels, ~10ppi, as show in Figure 1).

Figure 1. The screen of the driving simulator

Tasks

Surrogate reference task (SuRT). The task that was used to simulate Non-driving-related tasks (NDRTs) in an automated driving scenario is called Surrogate Reference Task (ISO 14198, 2012). As shown in Figure 2, participants were asked to select a single, larger circle (diameter 47 px) in a scatter of 50 distractor circles (diameter 40 px). Participants could start a new trial by clicking the "start" button (in the middle of the three buttons at the bottom of the screen). The task was shown on an Apple iPad. The placement of the iPad required participants to allocate their visual focus completely away from driving scene when performing the task.

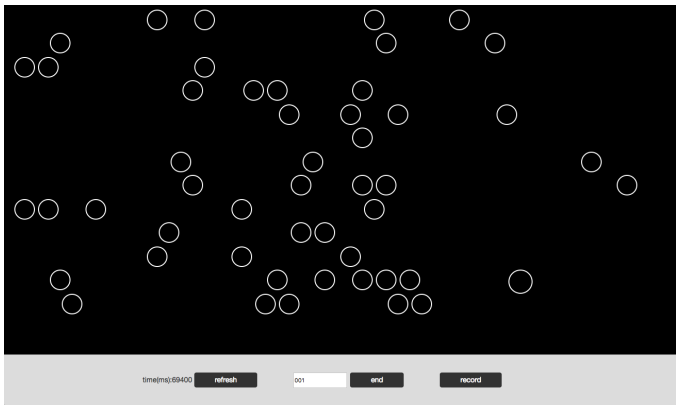

Figure 2. The Surrogate Reference Task (SuRT); example screen

Measures

Self-reported trust. Trust level towards the self-driving car was measured through a 7-point Likert scale (1= *totally disagree*, 7= *totally agree*). The scale consisted of seven

questions and was modified from one of the most commonly used self-report scales of trust in automation (Jian, Biasantz, Drury & Llinas, 2000). Trust score was calculated by averaging the sum of seven items (scores for item 1 and item 5 were reversed). A Higher score indicated a higher trust level. Trust scores were collected before experiment and after every driving phase during the experiment. The Cronbach's alpha of this scale was 0.90, 0.83, 0.87, 0.92 for pre-test and after every phase respectively.

Gaze behavior. Participants wore a Tobii Pro Glasses 2, a mobile eye-tracker during all phases of the experiment. The eye-tracker is equipped with two cameras for each eye to capture eye movements, and a wide-angle full HD scene camera to record the live view. The glasses were connected to a recording unit and wirelessly to a Dell tablet with the Tobii Pro Glasses Control software (Tobii Pro Glasses 2, 2018).

We used Tobii Pro Lab to analyze gaze behavior. First, three driving phases were selected separately as three time of interest (TOI) from the entire recording. As show in Figure 3, we manually defined two areas of interest (AOI): the road (the central screen), and the display of the SuRT (iPad).

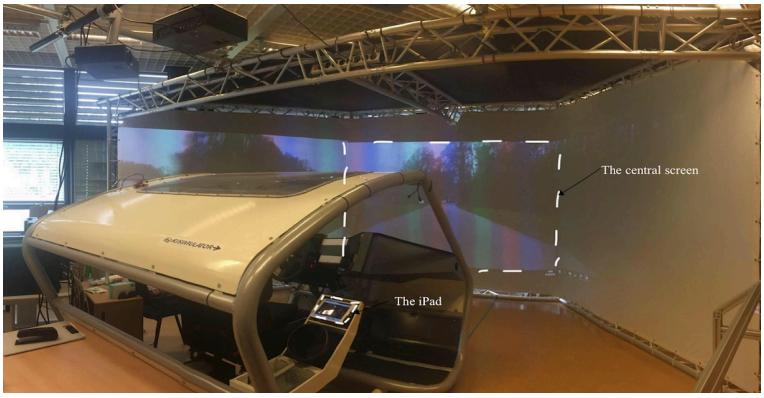

Figure 3. Two areas of interest (AOI) are mapped with white dotted line

Fixations outside of the two AOIs were counted as fixations on other area. We calculated fixation duration (fixation time on an AOI/total fixation time during the TOI) and fixation frequency (fixation count on an AOI/total fixation count during the TOI) on driving scene (the road) and NDRT (the display of the SuRT).

Skin conductance. Skin conductance was recorded using the Empatica E4 wristband. The E4 is a wearable wireless watch-like device designed for continuous, real-time data acquisition. Electrodermal activity data was acquired through two ventral wrist electrodes and was sampled at 4 Hz. The units of skin conductance are the microsiemens (μS). The E4 also contains sensors to measure heart rate, temperature, and acceleration (Further information can be found on the website: https://www.empatica.com/research/e4/). Participants wore the wristband on the left hand in all three phases to minimize movement (as they need to use right hand to perform NDRT during driving phases).

We analyzed skin conductance level response. Recordings during the one minute interval before phase 1 started was considered as baseline, as there were no stimulus and little movements during that period. Skin conductance values during the baseline and three driving phases were selected for further analysis. They were first z-transformed for inter-personal comparisons. Then the z-transformed values were averaged separately for baseline and three phases. Baseline values were subtracted from values of the three phases. This resulted in one value of skin conductance level response for every phase.

Procedure

A few days before the experiment, participants completed the pre-test questionnaire consisting of demographic questions and the trust scale to measure initial trust level in self-driving car (see in Appendix A). Participants were assigned to either of the group based on their initial trust level. The on-site experiment lasted about 40 minutes. Participants were first welcomed and signed the informed consent form. Then they put on the wristband to measure skin conductance and eye-tracker to measure gaze behavior, and sat in the simulator. They practiced the SuRT before phase 1. The experiment was composed of three driving phases. Participants were instructed to imagine themselves being in a self-driving car and to perform the NDRTs only if they trusted the behavior of the self-driving car. Phase 1 was equivalent to

a practice phase, in which participants got used to the simulator and NDRT, and developed an expectation towards the self-driving car (high trust for the Perfect Vehicle Group and low trust for the Poor Vehicle Group). After phase 1, they filled out the trust scale (see in Appendix B). During phase 2, different driving scenarios were presented, but the car behaved in the same way as in phase 1. Participants completed the trust scale again after phase 2. In phase 3, both groups of participants viewed the same low reliability condition videos in which the car performed the driving tasks poorly. After phase 3, they filled out the trust scale for the last time in the simulator. After that, eye-tracker and wristband were removed and participants were asked to complete an exit questionnaire (see in Appendix C).

Results

Self-report trust

The pre-test trust level was balanced between groups (Perfect Vehicle Group: M=4.02, SD=1.29; Poor Vehicle Group: M=4.09, SD=1.31).

Repeated-measure ANOVA was adopted for analysis of self-report trust. The sphericity test showed a non-significant result and the residuals were normally distributed. As in Figure 3, results showed a significant effect of phases ($F\ (3,\ 72)$ =6.21, p=.001, η^2= .21) and an interaction between phases and group ($F\ (3,72)$ =20.00, p<.001, η^2= .46).

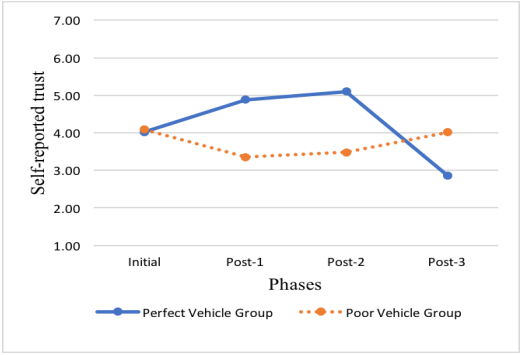

Figure 3. Mean values of trust score for each group and every phase

Since the interaction was significant, simple effects tests were further conducted. As

the effects of group in different phases showed, while two groups had no significant difference in initial trust in automation ($F\ (1,24) = .02$, $p = .90$, $\eta^2 = .001$), the Perfect Vehicle Group had higher trust levels after phase 1 ($F\ (1,24) = 23.02$, $p < .001$, $\eta^2 = .49$) and phase 2 ($F\ (1,24) = 21.83$, $p < .001$, $\eta^2 = .48$). In contrast, the Perfect Vehicle Group had lower trust level than the Poor Vehicle Group after phase 3 ($F\ (1,24) = 5.76$, $p = .03$, $\eta^2 = .19$).

In terms of the effect of phase within each group, results are displayed in Table 1. In the Perfect Vehicle Group, participants had higher trust level after phase 1 and phase 2 compared to initial trust level. In contrast, their trust level decreased significantly after phase 3. Trust in the Poor Vehicle Group decreased significantly after phase 1, and remained the same after phase 2. Surprisingly, their trust level increased after phase 3, compared to after phase 2.

Table 1. *The Effect of Phase Within Each Group*

Group	Self-reported trust	Mean-difference (The latter phase minus the former)	*P*
Perfect Vehicle Group	Pre-test—Phase 1	.86*	.02
	Phase 1—Phase 2	.22	.31
	Phase 2—Phase 3	-2.23**	.00
Poor Vehicle Group	Pre-test—Phase 1	-.74*	.05
	Phase 1—Phase 2	.132	.54
	Phase 2—Phase 3	.53*	.04

** significant at the 0.01 level.
* significant at the 0.05 level.

Gaze behavior

Because gaze data did not fit the assumptions of mixed Anova, non-parametric methods were used. A Mann Whitney U test showed that gaze duration of the two groups was significantly different only in phase 2 (SuRT: $U = 35.50$, $p = .01$; Road: $U = 42.00$, $p = .03$). For gaze frequency, there was also a significant difference only in phase 2 (SuRT: $U = 38$, $p = .02$; Road: $U = 46$, $p = .05$). Gaze behavior on road were shown in Figure 4.

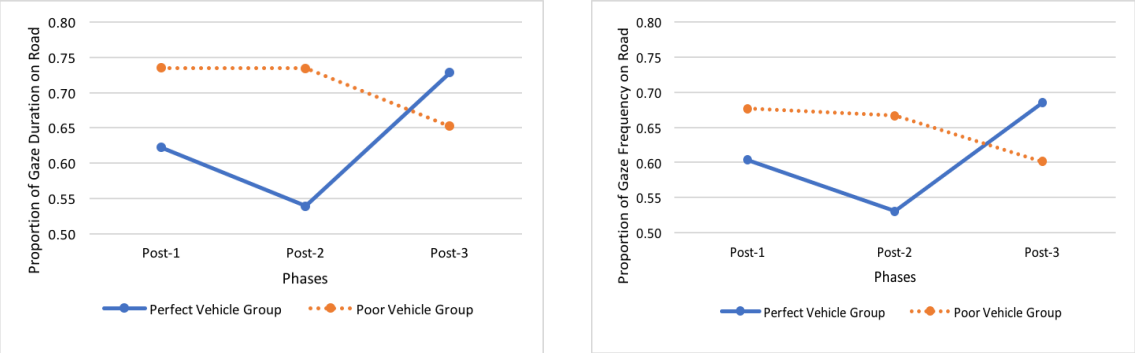

Figure 4. Mean proportion of gaze behavior on road for each group and every phase

As shown in Table 2, a Friedman test (non-parametric test for one-way Anova of repeated measures) showed the Perfect Vehicle Group's gaze behavior was significantly different among three phases. However, for the Poor Vehicle Group, there was a significant difference only in gaze duration on road.

Table 2. *Friedman test of gaze behavior for two group*

	Gaze Duration		Gaze Frequency	
	SuRT	Road	SuRT	Road
Perfect Vehicle Group	15.85**	15.18**	10.31*	12.28*
Poor Vehicle Group	2.00	6.62*	1.17	5.22

** significant at the 0.01 level.
* significant at the 0.05 level.

Skin conductance level (SCL) response

Like with the gaze behavior data, non-parametric methods were adopted also for skin conductance data. The Mann Whitney U tests showed that there was no significant difference in skin conductance level for the two groups in all three phases. The Friedman tests showed that neither group had significant differences among the three phases.

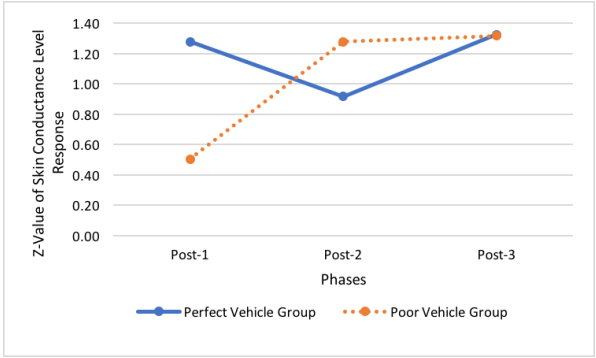

Figure 5. Mean SCL response (z-transformed) for each group and every phase

Correlations

Pearson correlation was conducted for self-reported trust, gaze behavior, and skin conductance level response. Results for gaze behavior showed a consistent correlation between gaze behavior and self-report trust. Specifically, in every phase, both gaze duration and gaze frequency to the road were positively associated with self-reported trust after that phase, while gaze behavior to SuRT were negatively related to self-repot trust. Moreover, gaze behavior in phase 2 had a strong correlation to self-reported trust after phase 1 as well (see Table 3 and Table 4).

However, skin conductance level response was negatively associated with self-reported trust only in phase 2 (see Table 5).

Table 3. *Correlation between self-report trust and gaze duration*

	Phase 1 Duration		Phase 2 Duration		Phase 3 Duration	
	SuRT	Road	SuRT	Road	SuRT	Road
Post-1	.651**	-.599**	.707**	-.643**	.398*	-0.329
Post-2	.404*	-0.379	.642**	-.602**	0.282	-0.245
Post-3	0.071	-0.109	-0.005	-0.042	.592**	-.625**

** Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

Table 4. *Correlation between self-report trust and gaze frequency*

	Phase 1 Frequency		Phase 2 Frequency		Phase 3 Frequency	
	SuRT	Road	SuRT	Road	SuRT	Road
Post-1	.606**	-.532**	.681**	-.562**	0.341	-0.295
Post-2	0.283	-0.259	.587**	-.514**	0.182	-0.17
Post-3	0.116	-0.131	0.016	-0.039	.555**	-.578**

** Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

Table 5. *Correlation between self-report trust and skin conductance level response*

	SCR-Phase 1	SCR-Phase 2	SCR-Phase 3
Post-1	0.083	-0.221	-0.012
Post-2	-0.06	-.389*	-0.134
Post-3	-.455*	-.441*	-0.299

* Correlation is significant at the 0.05 level (2-tailed).

Discussion

The main goal of the study was to investigate whether gaze behavior and skin conductance could be reliable measures of trust in automation. We hypothesized that gaze behavior and skin conductance level response would be negatively associated with self-reported trust. Specifically, we divided participants into either the Perfect Vehicle Group or the Poor Vehicle Group, with the simulated self-driving car performed either perfectly or poorly with driving tasks. As expected, participants in the Perfect Vehicle Group developed higher trust level to the self-driving car than the Poor Vehicle Group. Further, the moderate to strong association between gaze behavior and self-reported trust indicated that gaze behavior could be used as a reliable measure of trust in automation.

Our main finding was that gaze behavior could be used to measure trust in automation. Gaze duration and gaze frequency on road were negatively related to self-reported trust, while gaze behavior on SuRT were positively associated with self-reported trust. These results are in line with previous findings (Hergeth et al., 2016; Körber, Baseler, & Bengler, 2018; Walker, Verwey, & Martens, 2018). As the simulated self-driving car performed in the same way in phase 1 and phase 2, it was reasonable for self-reported trust after phase 1 to be associated with gaze behavior in phase 2. However, the correlation coefficients of them were even larger than coefficients of gaze behavior in phase 2 and self-reported trust after phase 2. This result may indicate that the trust level participants hold before a phase starts plays an important role in later gaze behavior.

Participants had different initial trust levels towards self-driving car, which was balanced between two groups. We manipulated the reliability of the car's performance. As expected, results showed that participants in the Perfect Vehicle Group had higher self-reported trust than the Poor Vehicle Group after both phase 1 and phase 2, which confirmed the success of our manipulation. The result is consistent with a previous study which used the

same way of manipulation (Walker, Verwey, & Martens, 2018). The result also confirms previous findings that users' initial trust in automation is vulnerable to later experience (Bahner, Hüper, & Manzey, 2008; Sauer et al., 2016). However, for gaze behavior, the two groups showed no significant difference in phase 1. One possible explanation is that trust level needs a process to develop. Since the two group had balanced initial trust, their gaze behavior may also not differentiate at the beginning of phase 1. As they got to know more about the self-driving car's behavior, they developed different trust levels and the gaze behavior changed together with trust level. Another explanation is that participants simply needed some time to understand the rule of the experiment. In both cases, this finding suggests that a practice phase is necessary. Moreover, we added a third phase in which both groups experienced a self-driving car with poorly reliable behavior. As expected, the Perfect Vehicle Group's high trust level built through the previous two phases dropped to a low level, which was even lower than trust level of the Poor Vehicle Group. The poor behavior of the self-driving in phase 3 seems to be more intolerable with the comparison of the previous perfect behavior. The trust in the self-driving car which was built through the previous two phases was easily destroyed in phase 3. This finding shows that trust in automation is similar to interpersonal trust in the way that they are both easy to lose (Earle, 2010). However, the trust level for the Poor Vehicle Group elevated significantly in phase 3, compared to the previous two phases, even though the self-driving car performed in a similar manner throughout the three phases. One possible explanation is that after longer exposure to the poorly behaved vehicle, participants have developed a trust level that close to the true reliability of the simulated self-driving car because they truly understand the car's behavior. In other words, although the car's behavior was not perfect, they trusted it would not crash. For example, several participants mentioned in the exit questionnaire that they noticed although the car tended to drift to the center of the road, it would go back to its own side

when there was car coming in the opposite direction.

For skin conductance level, there was no consistent correlation with self-reported trust, except in phase 2 there was a weak negative correlation. This was inconsistent with previous findings (Khawaji, Zhou, Chen, & Marcus, 2015). One possible explanation is that the arousal level caused by the experiment setting is smaller than by a real setting. Another possibility is that the implementation of measurement may be flawed. For example, we used the time interval before phase 1 started as baseline, during which participants were reading instructions. Moreover, we used the E4 wristband to measure skin conductance, which might be different from stationary sensor measured via finger (Ollander, Godin, Campagne, & Chararbonnier, 2016). One prerequisite of using the E4 wristband to get reliable data is good connection with the skin, which means the wristband needs to be rather tightly. However, during the experiment, there was no objective standard whether the wristband was tight enough but left for participants themselves to judge.

Taking together, the study adds evidence to the application of using gaze behavior as an indicator of trust in automation. Gaze behavior is a valuable tool to measure trust in applied settings for many reasons. For one thing, gaze behavior is a promising tool to measure temporary trust in automation. It is possible to look at the gaze behavior in a shorter time frame and see how gaze behavior changes. Moreover, gaze behavior may be more closely related to take over quality than self-reported trust when manually operation is required. For example, some users may switch from SuRT and road constantly, while others focus on road at the beginning and only switch to SuRT after they understand the car's behavior. These differences may have direct influence in situational awareness and thus influence take over quality. The findings in phase 3 also facilitate our understanding of building trust in automation. Previous studies suggest that we need to build appropriate trust level so that users don't over-rely or disuse the automated system (Carlson, Desai, Drury, Kwak, & Yanco.,

2014; Bagheri & Jamieson, 2004). However, the study suggests that trust in automation is volatile. For the Perfect Vehicle Group, the developed trust level was easily destroyed by exposure to one unreliable phase. This suggests that the trust level of users may be destroyed by one single accident in real-life. On the contrary, it's also worth noting that the Poor Vehicle Group's trust level increased as the exposure time became longer. This finding raises another issue about how to keep users' trust in automation in an appropriate level.

The study also has some limitations. Firstly, this study was a video-based experiment that happened in a simulator room. As mentioned in some participants' exit questionnaires, they felt less responsible than actually driving on the road. Thus, it is likely that the gaze behavior and skin conductance in real driving scenario will differ in this study. Future research could benefit from adding options for participants to disengage automated driving and resume manually control. It would make it closer to real situations and moreover provide implications of take-over behavior in a self-driving car. Secondly, as discussed above, the implementation of skin conductance measurement might be flawed. Future studies could use stationary sensors or at least set an objective standard to check if the connection of wristband is tight enough to ensure data quality.

In conclusion, gaze behavior appears to be a promising reliable indicator of trust in automation. It's an objective, non-invasive and continuous measure that viable in many settings. Our study not only adds evidence to the use of gaze behavior as a measurement of trust in automation, but also make some interesting implications into how to build appropriate trust level of users. Future research should take one step forward to investigate the possibility of using gaze behavior to measure temporary changes in trust level. It would also be interesting to investigate the interactions among self-reported trust, gaze behavior and take over behavior.

References

Akash, K., Hu, W. L., Jain, N., & Reid, T. (2018). A Classification Model for Sensing Human Trust in Machines Using EEG and GSR. arXiv preprint arXiv:1803.09861.

Bagheri, N., & Jamieson, G. A. (2004). Considering subjective trust and monitoring behaviour in assessing automation-induced "complacency." In D. A. Vicenzi, M. Mouloua, & P. A. Hancock (Eds.), *Human performance, situation awareness and automation: Current research and trends (Volume II)* (pp. 54–59). Mahwah, NJ: Lawrence Erlbaum Associates.

Bahner, J. E., Hüper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66*(9), 688-699. https://doi.org/10.1016/j.ijhcs.2008.06.001

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science, 8*(4), 321-348. https://doi.org/10.1080/14639220500535301

Biassoni, F., Ruscio, D., & Ciceri, R. (2016). Limitations and automation. The role of information about device-specific features in ADAS acceptability. *Safety Science, 85*, 179-186. https://doi.org/10.1016/j.ssci.2016.01.017

Carlson, M. S., Desai, M., Drury, J. L., Kwak, H., & Yanco, H. A. (2014). Identifying factors that influence trust in automated cars and medical diagnosis systems. In *AAAI Symposium on The Intersection of Robust Intelligence and Trust in Autonomous Systems* (pp. 20–27). Palo Alto, CA: Association for the Advancement of Artificial Intelligence. Achieved from http://www.aaai.org/ocs/index.php/SSS/SSS14/paper/download/7729/7725

Casner, S. M., Hutchins, E. L., & Norman, D. (2016). The challenges of partially automated driving. *Communications of the ACM, 59*(5), 70-77. doi: 10.1145/2830565

Choi, J. K., & Ji, Y. G. (2015). Investigating the Importance of Trust on Adopting an Autonomous Vehicle. *International Journal of Human–Computer Interaction, 31*(10), 692-702. doi: 10.1080/10447318.2015.1070549

Costa, A. C., Roe, R. A., & Taillieu, T. (2001). Trust within teams: The relation with performance effectiveness. *European journal of work and organizational psychology, 10*(3), 225-244. https://doi.org/10.1080/13594320143000654

Earle, T. C. (2010). Trust in risk management: a model‐based review of empirical research. *Risk analysis, 30*(4), 541-574. https://doi.org/10.1111/j.1539-6924.2010.01398.x

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. *Cognition, Technology & Work, 14*(1), 39-49. doi: 10.1007/s10111-011-0194-3

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in automation–Before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing, 3*, 3025- 3032. https://doi.org/10.1016/j.promfg.2015.07.847

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013, October). 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). ACM.

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. https://doi.org/10.1177/0018720815625744

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on

factors that influence trust. *Human Factors, 57*(3), 407-434. https://doi.org/10.1177/0018720814547570

ISO 14198. (2012). PD 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: BSI.

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. https://doi.org/10.1207/S15327566IJCE0401_04

Khawaji, A., Zhou, J., Chen, F., & Marcus, N. (2015). Using galvanic skin response (GSR) to measure trust and cognitive load in the text-chat environment. *In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1989-1994). ACM. doi: 10.1145/2702613.2732766

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics, 66*(Supplement C), 18-31. doi: https://doi.org/10.1016/j.apergo.2017.07.006

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies, 40*(1), 153-184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Madhavan, P., & Wiegmann, D. A. (2005). Cognitive anchoring on self-generated decisions reduces operator reliance on automated diagnostic aids. *Human factors, 47*(2), 332 -341. https://doi.org/10.1518/0018720054679489

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors,*

48(4), 656-665. https://doi.org/10.1518/001872006779166334

Ollander, S., Godin, C., Campagne, A., & Charbonnier, S. (2016). A comparison of wearable and stationary sensors for stress detection. In *IEEE International Conference on Systems, Man, and Cybernetics, October.* doi: 10.1109/SMC.2016.7844917

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. doi: 10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors, 39*(2), 230-253. https://doi.org/10.1518/001872097778543886

Payre, W., Cestac, J., & Delhomme, P. (2014). Intention to use a fully automated car: Attitudes and a priori acceptability. *Transportation research part F: traffic psychology and behaviour, 27*, 252-263. https://doi.org/10.1016/j.trf.2014.04.009

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in neuroinformatics*, 2, 10. https://doi.org/10.3389/neuro.11.010.2008

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of management review, 23*(3), 393-404. https://doi.org/10.5465/amr.1998.926617

Santarcangelo, E. L., Paoletti, G., Balocchi, R., Scattina, E., Ghelarducci, B., & Varanini, M. (2012). Watching neutral and threatening movies: Subjective experience and autonomic responses in subjects with different hypnotizability levels. *International Journal of Psychophysiology, 84*(1), 59-64. https://doi.org/10.1016/j.ijpsycho.2012.01.010

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics, 59*(6), 767-780. https://doi.org/10.1080/00140139.2015.1094577

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics, 1*(1), 89–129. doi:10.1518/155723405783703082

Tobii Pro Glasses 2 (2018). Eye tracking specifications [company website]. Retrieved from: https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/#Specifications

Walker, F., Verwey, W., & Martens, M. (2018). Gaze behavior as a measure of trust automated vehicles. *Proceedings of the 6th Humanist Conference, the Hague, Netherlands, 13-14 June.* Retrieved from: http://www.humanist-vce.eu/fileadmin/contributeurs/humanist/TheHague2018/29-walker.pdf

Appendix A

Pre-test Questionnaire

PART 1:

Name and Surname:

Age:

Gender:

Nationality:

Years of driving experience (i.e. Years that passed since when you first got your driving licence):

On average, how often do you drive on European roads (including Dutch roads)?

- Never
- Once per month
- Once per week
- Twice per week
- Every day

PART 2:

Through this brief questionnaire we would like to measure your attitudes toward self-driving cars.

We understand that your knowledge on self-driving cars might be limited, so please answer based on your ideas and expectations.

Please respond as truthfully as possible, and keep in mind that there is no “correct” answer.

Your privacy is protected according to Dutch law.

Please circle your answer.

1 = not at all

7 = extremely

1. I am cautious about self-driving cars

1 – 2 – 3 – 4 – 5 – 6 – 7

2. Self-driving cars are reliable

1 – 2 – 3 – 4 – 5 – 6 – 7

3. I would entrust my car to self-driving functions for lane changing, automatic braking, etc.

1 – 2 – 3 – 4 – 5 – 6 – 7

4. I can count on self-driving cars

1 – 2 – 3 – 4 – 5 – 6 – 7

5. Self-driving cars can have harmful consequences

1 – 2 – 3 – 4 – 5 – 6 – 7

6. I trust self-driving cars

1 – 2 – 3 – 4 – 5 – 6 – 7

7. I assume that self-driving cars will work properly

1 – 2 – 3 – 4 – 5 – 6 – 7

Appendix B

Post-test Questionnaire

Answer the questionnaire keeping in mind the behaviour of the car you have been "driving" during this experiment.

Please respond as truthfully as possible, and keep in mind that there is no "correct" answer.

Your privacy is protected according to Dutch law.

Please circle your answer.

1 = not at all; 7 = extremely

1. I was cautious about the self-driving car

1 – 2 – 3 – 4 – 5 – 6 – 7

2. The self-driving car was reliable

1 – 2 – 3 – 4 – 5 – 6 – 7

3. I would entrust my car to the tested self-driving functions (lane changing, automatic braking, etc.)

1 – 2 – 3 – 4 – 5 – 6 – 7

4. I could count on the self-driving car

1 – 2 – 3 – 4 – 5 – 6 – 7

5. This self-driving car can have harmful consequences

1 – 2 – 3 – 4 – 5 – 6 – 7

6. I trusted the self-driving car

1 – 2 – 3 – 4 – 5 – 6 – 7

7. The self-driving car worked properly

1 – 2 – 3 – 4 – 5 – 6 – 7

Appendix C

Exit Questionnaire

Educational level (Bsc, Msc, etc.):

Main transport mode:

- Car/motorbike
- Public transport
- Bicycle
- Walking
- Other

On a scale from 1 to 7, how interested are you in self-driving cars? (please circle your answer)

1 = Not at all

7 = Extremely

1 – 2 – 3 – 4 – 5 – 6 – 7

Did you have the feeling that the car was behaving in a safe or in an unsafe manner? Why?

Please write here any further comments: