Gaze Behaviour as a Measure of Trust in Automated Vehicles

Francesco Walker, University of Twente, The Netherlands, f.walker@utwente.nl, **Willem Verwey**, University of Twente, The Netherlands, **Marieke Martens**, University of Twente, The Netherlands

ABSTRACT

Extensive evidence shows that drivers' intention to use self-driving technology is strongly modulated by trust, and that the benefits promised by automated vehicles will not be achieved if users do not trust them. It follows that vehicles should be designed to achieve optimal trust. However, there is no consensus as to how this should be assessed. To date, most studies have relied on self-reports, an approach subject to problems. In the driving simulator study reported here, we used drivers' gaze behaviour during simulated Highly Automated Driving to investigate whether this could provide a more objective measure. The results indicated a negative relationship between self-reported trust and monitoring behaviour: The higher their self-reported trust, the less participants monitored the road, and the more attention they paid to a non-driving related secondary task. These findings suggest that gaze behaviour in a secondary task situation provides a more objective indicator of driver trust than self-reports.

Keywords: trust in automation, eye movement behaviour, human factors, autonomous driving, automated vehicle, secondary task.

1 INTRODUCTION

Studies show that more than 90% of road accidents are due to human errors (e.g. NHTSA, 2008). Automated vehicles are expected to drastically reduce the number of accidents. They are also expected to increase travelling comfort, allowing users to engage in all sort of activities while the car takes care of driving, and giving back the freedom to travel to individuals who have lost the ability to drive due to age or disability (Payre, Cestac, & Delhomme, 2016; Urmson & Whittaker, 2008). Nevertheless, none of these benefits will be achieved if users' do not trust the automated system.

Extensive evidence indicates that acceptance, perceived usefulness and intention to use self-driving technology are strongly mediated by trust (Choi & Ji, 2015; Ghazizadeh et al., 2012; Parasuraman & Riley, 1997). Importantly, trust evolves dynamically over time. This means that drivers' intention to use the technology depends not just on their initial trust levels, but also on their experience with the automated system and on its perceived reliability (Körber, Baselar, & Bengler, 2018; Sauer, Chavaillaz, & Wastell, 2016; Lee & See, 2004).

To understand how trust impacts the use of Highly Automated Vehicles we need effective trust metrics. Up till now, most studies have relied on self-reports. However, questionnaires are not a continuous measure. This means they cannot capture real-time changes in user trust, and are hard to use outside an experimental context (Hergeth, Lorenz, Vilimek, & Krems, 2016). Analysis of drivers' gaze behaviour could potentially provide a continuous and objective measure of trust.

To date, there have been few studies on the relationship between driver trust during Highly Automated Driving (HAD) and eye-movement behaviour. A few (e.g. Körber et al., 2018; Hergeth et al., 2016; Helldin, Falkman, Riveiro, & Davidsson, 2013) suggest that participants with a high level of trust tend to monitor the road less. Others (e.g. Gold et al., 2015) have failed to find such a link. The aim of our study is to generate new evidence that can help to resolve the question whether eye movement behaviour may provide a reliable indicator for the trust of drivers in automated vehicles. Importantly, if drivers' trust could be objectively measured in real-time, vehicle behaviour and display information could then be tuned accordingly.

In the present study, we investigated the influence of the vehicle reliability on drivers' monitoring behaviour. Videos were used to simulate HAD. In the videos, the simulated vehicle performed longitudinal and lateral vehicle control, and applied the brakes when cyclists or pedestrians were crossing the road. Participants were asked to pay attention to the road and perform a secondary task, but *only* if they trusted the way the system was handling the driving. We compared two groups of participants, where each participant was seated in the mock-up of the driving simulator. One group viewed videos of a car handling the driving task perfectly; a second group viewed videos of a car struggling with the driving task (i.e. it tended to drift towards the centre of the road and braked abruptly when approaching crossing pedestrians or cyclists). In line with Hergeth et al. (2016) and Korber et al.'s

(2018) results, we expected to find a negative relationship between trust and monitoring frequency: The less drivers trust the system, the more they view the road, and vice versa.

2 METHODS

2.1 Participants

Thirty participants, all students of the University of Twente, were selected for the experiment and participated in exchange for money (6 euro) or study credits. Six participants were excluded from the analysis: 5 due to the poor quality of their eye-tracking data (i.e. pupil detected in less than 79% of the frames), and 1 because the post interview showed that the participant had previous experience with a Level 2 (SAE, 2014) automated vehicle. The other participants reported no previous experience with automated vehicles. All participants had their driver's licence for at least one year and reported normal vision. None wore glasses, and none reported that they commonly suffered from motion sickness. The 14 female and 10 male participants selected for the final analysis were all between 18 and 24 years of age (M = 20.46; SD = 1.414).

2.2 Videos – Pilot study

A Go-Pro Hero 4 Session camera, placed centrally on the hood of an Audi A4, was used to film 14 two-minute driving scenarios, encountered while driving in the surroundings of the University of Twente. We then performed a pilot study in the driving simulator of the University of Twente, in which 10 participants viewed the videos in a different random order. One participant was excluded due to motion sickness. The remaining participants were between 19 and 28 years of age (M = 21.7; SD = 2.3). They were asked to imagine being in an autonomously driving vehicle. After each video they were asked to verbally rate the vehicle's performance on a 7-point Likert scale. Of the 14 videos, the 3 with the highest scores (i.e. M > 5.7) and the 3 with the lowest scores (i.e. M < 2.8) were used in the main experiment. Wilcoxon signed ranks tests showed that the scores of the 3 highest rated videos significantly differed from the 3 scores with the lowest ratings (all *p*'s = .005).

2.3 Tasks

Participants were assigned to one of two groups: The Perfect Vehicle group viewed forward looking videos recorded from a car handling the driving task perfectly (it always kept in lane, and braked comfortably in front of crossing cyclists or pedestrians). The Poor Vehicle group viewed videos of a car struggling with the driving task, with a tendency to drift towards the centre of the road, and to brake abruptly in front of crossing cyclists and pedestrians. Perfect Vehicle and Poor Vehicle participants were both confronted with one crossing cyclist or pedestrian per video. Although group assignment was random, the randomization procedure ensured the same distribution of participants by gender and initial trust level in each group.

Participants of both groups performed a non-driving-related secondary task (NDRT) during simulated driving. This was presented on a Dell Latitude 7370 laptop, and was programmed using Psychopy software (Peirce, 2009). As in Verwey (2000), the NDRT consisted of a simple addition task. In each trial, a target number, between 20 and 87, was presented in red at the centre of the screen. Below the target number, the screen showed an addition in the form <number>+7. As previously, the number always lay in the range from 20 to 87. Participants were asked to indicate whether the result of the addition was higher or lower than the target number. At the end of each trial, system feedback (the word "correct" for correct responses, the word "incorrect" for incorrect responses) was presented centrally on the laptop screen. Participants had no time limit to perform the task. Responses were recorded using the right and left arrows of a standard keyboard, which participants held on their lap for the entire duration of the experiment. The task is presented in Figure 1.

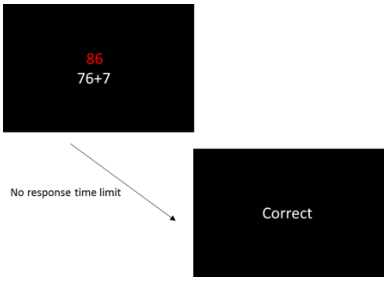

Figure 1. Secondary addition task. Correct trial: By pressing the left arrow, the participant correctly indicated that 76+7 is lower than the target number "86".

To perform the NDRT participants allocated their attention away from the driving simulator screen, and could not use peripheral vision to watch the road while they were performing the secondary task. The time participants were watching the NDRT screen served as an indicator of trust.

2.4 Trust questionnaire

Participants' trust in the simulated vehicle was also measured using a modified version of the Empirically Derived (ED) Trust Scale (Jian, Bisanz, & Drury, 2000). This uses a 7-point Likert scale (1 = totally disagree; 7 = totally agree), to indicate level of agreement with seven statements (based on Verberne, Ham, & Midden, 2012). Participants filled in the trust questionnaire before and after completing the experiment (see section "Procedure"). Assessing trust before the start of the experiment was important to ensure that participants in each group were equally distributed in terms of their initial trust.

2.5 Apparatus

2.5.1 Driving simulator

The driving simulator of the University of Twente consists of a mock-up equipped with steering wheel, pedals and indicators. Videos, displayed through Psychopy software (Peirce, 2009), are projected on a screen of 7.8 x 1.95 meters. The screen has a total resolution of 3072*768 pixels (~10ppi).

2.5.2 Mobile eye-tracker

Participants' eye-movements were recorded using Tobii Pro Glasses 2. The head-mounted mobile eye-tracker weighs 45 grams and is equipped with four eye cameras. It tracks movements of both eyes, and uses an additional scene camera to track the external world. The glasses were connected to a 130 x 85 x 27 mm recording unit with a weight of 312 grams. The eye-tracker was wirelessly connected to a Dell tablet, running Windows 10 operating system and the Tobii Pro Glasses Controller software (Tobii Pro Glasses 2, 2018).

2.6 Procedure

Participants filled out a pre-measurement trust questionnaire online at home. Then, on the testing day, they were welcomed to the driving simulator room and told that they would view videos while sitting in the simulator mock-up. After filling in an informed consent form, the mobile eye-tracker was calibrated following the procedure recommended by Tobii (i.e. Tobii Pro Glasses 2, 2018). After calibration, participants sat down in the mock-up of the driving simulator. On screen instructions informed them that the experiment would be divided into two phases. Phase 1 would be a "trust development session", in which they could develop a general idea of how the system worked by watching its behaviour. They were told that in this phase they should focus on the vehicle's behaviour. In phase 2 the car would behave as in phase 1, though the scenarios would be different. The phase 1 video was then presented on screen. At the end of the video, participants were asked "How do you think that the car coped with the driving task?". Responses were expressed out loud using a 7-point Likert scale, with 1 indicating "very badly" and 7 "very well".

After participants' responses had been recorded, instructions for phase 2, the "active session", were displayed on screen: Participants were told to imagine they were in a self-driving car, and to perform the secondary task whenever they felt that the car was handling the driving safely. Once again, participants were reminded that the car would behave as in phase 1. Participants were given time to practice the secondary task before the start of the phase. After practice, two videos were presented. At the end of phase 2, participants were asked to rate how they thought the car had coped with the driving task. The order of the videos was counterbalanced across participants.

Participants were then asked to leave the mock-up, remove the eye-tracker, and fill in the post trust questionnaire. Although the items of the post-measurement were the same as those in the pre-measurement questionnaire, they were rephrased to refer to participants' trust towards the vehicle they had just experienced, and not their general trust in self-driving cars. The experiment ended with a final questionnaire, in which participants were asked to provide information on their level of education and previous experience with automated vehicles. Responses were used to check that participants had no previous experience with automated vehicles, either as drivers or as passengers.

2.7 Analysis

Our analysis assessed participant eye-movement behaviour during phase 2 onto two regions of interest (ROI): 1) the road, i.e., the central section of the driving simulator screen (size 2.6 x 1.95 meters); 2) the laptop, on which the NDRT was presented. Specifically, we assessed fixation count (i.e. number of fixations made in each ROI) and fixation duration (i.e. total time spent viewing each ROI).

3 RESULTS

3.1 Monitoring behaviour and trust

Analysis of the results for the two groups showed that 65% (SD = .099) of the Perfect Vehicle group's fixations were on the road, as opposed to 83% (SD = .084) for the Poor Vehicle group. A Mann Whitney U test revealed that the difference was highly significant (U = 12, p = .001). Moreover, members of the Perfect Vehicle group spent 70% (SD = .117) of their time looking at the road, as opposed to 84% (SD = .113) for the Poor Vehicle group. Again, the difference between the two groups was highly significant (U = 24, p = .007). These results are strong evidence that monitoring behaviour could be an effective measure of trust. See Figure 2.

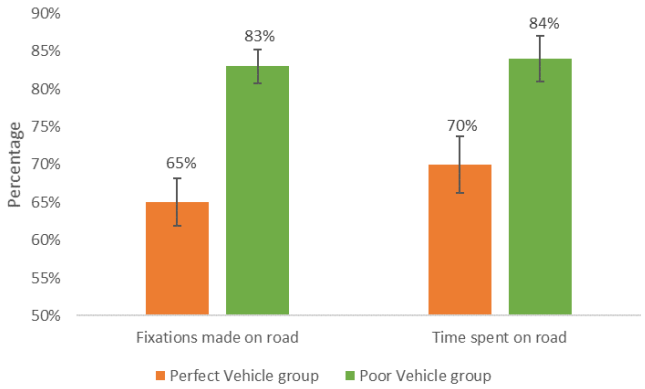

Figure 2. Participant monitoring behaviour. The Poor Vehicle group made more fixations (83% vs. 65%) and spent more time (84% vs. 70%) on the road compared to the Perfect Vehicle group. Error bars represent standard error of the means.

To confirm the relationship between trust and monitoring behaviour, we used a Pearson correlation to correlate the two measures. Self-reported trust scores collected at the end of the study and pooled across the two groups correlated strongly both with fixation count (r = -.658, p < .001) and with fixation duration (r = -.611, p = .001): In brief, the higher the trust, the less participants fixated the road. See Figure 3.

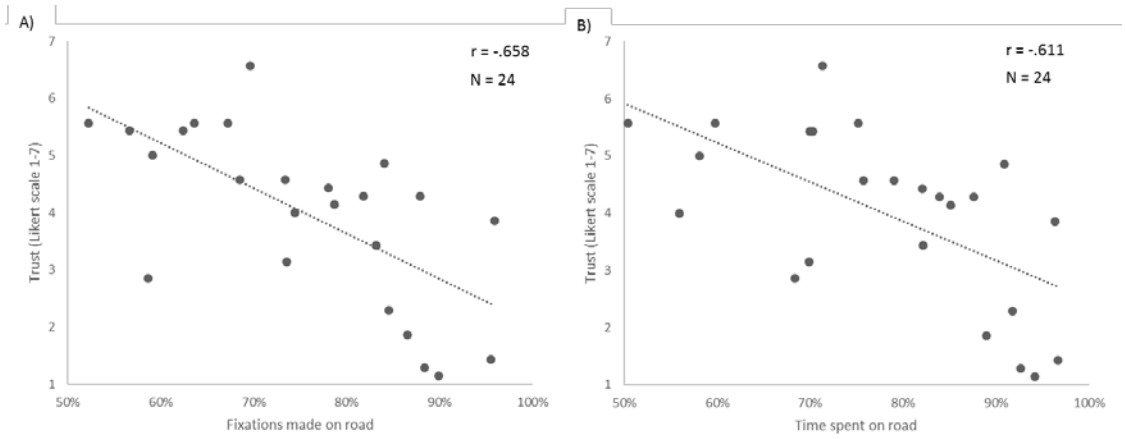

Figure 3. Correlation between self-reported post trust (1: low; 7: high) and monitoring behaviour. The higher the trust, the less participants fixated on the road. A) Fixations made on road; B) Time spent on road.

3.2 Validation

Two independent annotators used Tobii Pro Lab software to classify the same set of fixations on the road and on the laptop. Interrater reliability was estimated using a Pearson correlation. Fixation durations and fixation counts computed from the fixations mapped by judge 1 correlated strongly with the fixations mapped by judge 2 (durations: r = .998, $p < .001$; counts, r = .999, $p < .001$).

To test whether the videos presented to the Perfect Vehicle group were associated with higher self-reported trust than those presented to the Poor Vehicle group, we analysed the data with a Mann-Whitney U test. As expected, the Perfect Vehicle group had higher scores. This was true both for the videos presented during phase 1 (Perfect Vehicle: M = 6.2, SD = .632; Poor Vehicle: M = 4.93, SD = 1.328; U = 30.5) and phase 2 (Perfect Vehicle: M = 5.3, SD = .949; Poor Vehicle: M = 3.71, SD = 1.541; U = 29), $p = .016$. These findings are further evidence that monitoring behaviour can indeed provide a valid measure of driver trust.

4 DISCUSSION

Extensive evidence shows that adoption of automated technology is strongly modulated by user trust. Up to now, however, driver trust in self-driving technology has usually been investigated through self-reports – a technique that is not continuous, and difficult to use in real-world scenarios. To better understand the link between driver trust and the behaviour of automated vehicles, we need trust measures that are more objective and easier to use. Here, we investigated the potential of mobile eye-tracking technology in a secondary task situation. Participants were divided into two groups. The Perfect Vehicle group viewed videos of a simulated Highly Automated Vehicle (SAE, 2014) which coped perfectly with the driving task, while the videos presented to the Poor Vehicle group showed a vehicle that tended to swerve towards the centre of the road and braked abruptly in front of crossing cyclists or pedestrians. Participants' eye-movements to the road and the secondary task were recorded, and self-reported trust was measured before the start (for control purposes) and at the end of the experiment.

Comparison of the two measures showed that the higher the drivers' self-reported trust scores collected at the end of study, the less time they spent viewing the road, and vice versa. These results confirm the negative relationship between trust and monitoring behaviour (i.e. fixation count and duration) during Highly Automated Driving (SAE, 2014), previously reported by Hergeth et al. (2016) and Korber et al. (2018).

The results also confirm the success of our manipulation. The Perfect Vehicle group made more fixations and spent more time on the secondary task compared to the Poor Vehicle group. Participants were instructed to perform the secondary task only when they thought that the automated system was handling the driving task safely. It appears, therefore, that the Perfect Vehicle group trusted their simulated vehicle more than the Poor Vehicle group. This was confirmed by participants' answers to the question "How do you think that the car coped with the driving task?". Both in phase 1 and phase 2, the Perfect Vehicle group reported higher scores than the Poor Vehicle group.

One limitation of our study is that participants who thought that the automated vehicle was not handling the driving task safely were instructed to keep viewing the road. In a real-world scenario, it is likely that drivers who felt uncomfortable with the car's behaviour would disengage the automated system and take back manual control. Here, due to the video-based nature of the study, drivers could not disengage the automated system. It is also likely that participants in this simulation-based study did not feel as threatened by the vehicle's behaviour as they would have felt in a real-world situation. Nonetheless, results clearly show different eye-movement behaviour between the two groups.

To summarize, our study represents a step forward in the development of more objective, non-invasive and continuous measurements of drivers' trust in self-driving technology. Such measurements are particularly important in real-world scenarios, where they could be used to assess real-time changes in driver trust. If drivers over- or underestimate the capabilities of their vehicles, automated systems could react by modifying their behaviour or by providing them with additional information concerning the vehicles' performance. As suggested by Hergeth et al. (2016), future research should further investigate the interactions between self-reported trust, monitoring behaviour and driver behaviour. This is a necessary step towards the use of gaze behaviour as an objective and continuous measurement of driver trust.

REFERENCES

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.

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. Cognition, Technology & Work, 14(1), 39-49.

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.

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.

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.

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: manipulating trust in automation and reliance in automated driving. Applied ergonomics, 66, 18-31.

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

National Highway Traffic Safety Administration. (2008). National motor vehicle crash causation survey: Report to congress. National Highway Traffic Safety Administration Technical Report DOT HS, 811, 059.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human factors, 39(2), 230-253.

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual control recovery. Human factors, 58(2), 229-241.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. Frontiers in neuroinformatics, 2, 10.

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.

Taxonomy, S. A. E. Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems. Technical report, SAE International, 2014.

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

Urmson, C., & Whittaker, W. (2008). Self-driving cars and the urban challenge. IEEE Intelligent Systems, 23(2).

Verberne, F. M., Ham, J., & Midden, C. J. (2012). Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. Human factors, 54(5), 799-810.

Verwey, W. B. (2000). On-line driver workload estimation. Effects of road situation and age on secondary task measures. Ergonomics, 43(2), 187-209.