# Understanding Drivers' Trust After Software Malfunctions and Cyber Intrusions of Digital Displays in Automated Cars

*William Payre [1], Jaume Perelló-March [1], Giedre Sabaliauskaite [2], Hesamaldin Jadidbonab [2], Siraj Shaikh [2], Hoang Nguyen [2], Stewart Birrell [1]*

*[1] National Transport Design Centre*

*Coventry University, Coventry, CV1 2TT, UK*

*[2] Systems Security Group, Centre for Future Transport and Cities*

*Coventry University, Coventry, CV1 5FB, UK*

## ABSTRACT

The aim of this paper is to examine the effect of explicit (i.e., ransomware) and silent (i.e., no turn signals) failures on drivers' reported levels of trust and perception of risk. In a driving simulator study, 38 participants rode in a conditionally automated vehicle in built-up areas and motorways. They all experienced both failures. Not only levels of trust decreased after experiencing the failures, especially after the explicit one, but also some of the scores were low. This could mean cyber-attacks lead to distrust in automated driving, rather than merely decreasing levels of trust. Participants also seemed to differentiate connected driving from automated driving in terms of perception of risk. These results are discussed in the context of cyber intrusions as well as long- and short-term trust.

**Keywords**: Trust, Automation, Automotive, Cyber Security, Driving, Digital Display, Perception of Risk

## INTRODUCTION

### Background

Contemporary computerised automated systems, such as automated driving, posit challenges in terms of cyber security and user trust (Seetharaman et al., 2021). Indeed, confidence in how secure and reliable these systems are has dropped because of software vulnerabilities and connectivity (Schoettle and Sivak, 2014). Current threats against modern vehicular platforms stem from three main trends. The first one is the increased complexity of vehicle software (e.g., raise in lines of code and electronic control units; Antinyan, 2020). The second trend is the increased connectivity resulting in vehicles being more attractive and accessible to offenders (e.g., wireless networks and communication interfaces connected to the Internet of Things environment; Sheehan et al., 2019). The third trend is the increased content value with respect to personal and sensitive information (e.g., locations visited, journeys, route; see Deng et al., 2020). Software, connectivity and data storing systems built over digital platform could be exposed to security vulnerabilities and failure (ISO/SAE DIS 21434, 2020), which means that user trust in automotive systems remains an open question.

## State of the art

Past work examined the best practices on how to display in-vehicle information in order to optimise drivers' trust in the system (Wintersberger et al, 2020). One example is that displaying the status and actions of the system enhances transparency and supports appropriate levels of trust in the system (Carsten and Martens, 2019). Nevertheless, little is known on the consequences of failing to communicate reliable information on the vehicle status and operations on trust. This is prominent with respect to silent failures, whereby the system fails to inform the driver of its limit and inability to operate reliably (Louw et al., 2019). This lack of evidence is worrying from the safety perspective since automation failures can influence drivers' level of trust in automated driving systems (Payre et al., 2016; 2017; 2021). As a result, disuse (e.g. no use), misuse (e.g. unsafe operation as reported by the National Transportation Safety Board Tesla crash report, 2017) or abuse (e.g. take advantage of the limits) of such system may arise (Parasuraman et al., 1997). Recent evidence suggests that the subjective level of trust should be aligned with the capabilities of the automation to mitigate the undesirable effect of overtrust (i.e. using the automated system despite its unreliability) and distrust (i.e. not using the system although it is reliable; Khastgir et al., 2018). The name of this process is trust calibration (Lee and See, 2004).

A critical factor affecting the calibration of trust is perceived risk (Hoff & Bashir, 2015; Lee & See, 2004), defined as "the likelihood and consequences of error" (Riley, 1996). Perceived risk correlates negatively with trust (Riley, 1996) and two sub-types have been recently identified: relational and situational perceived risk (Stuck et al., 2021). Perceived relational risk refers to the driver's attitude towards the automated vehicle modulated by

experience. Perceived situational risk refers to the drivers' attitude towards the driving task or context, modulated by the probability of potential negative outcomes. According to Stuck et al., the trustworthiness of the conditionally automated vehicle defines relational risk, whilst the potential negatives outcomes from a task define situational risk. This means that each sub-type of perceived risk has unique relationships with trust (Li et al., 2019).

Despite numerous studies have shown what and how to present information to support trust calibration, there has been little discussion to comprehend whether, how and when screen failures influence drivers' trust in automated systems and perceived risk. Addressing this research gap, we will investigate the effect of the type of in-vehicle screen failure and its timing on individuals' trust and perceived risk.

The objective of this study is to understand the extent to which software malfunctions and cyber intrusions of digital displays occurring in conditionally automated vehicles (SAE level 3) impact drivers' trust – and by extension road safety if they decide to resume manual control from the vehicle. We have generated a set of realistic use-cases where screen safety failures, caused by either software failure or cyber-attack (i.e. ransomware) may result in not displaying information relevant to the operation and status of the automated driving system. Furthermore, a cyber-attack may lead to a security breach warning or ransomware popping up the screen. These use-cases are integrated in a bespoke driver in-the-loop simulator allowing the collection of attitudinal data. It was expected that: Hypothesis 1 (H1) a): Trust decreases after the software fails to display relevant information (i.e., silent failure) and b) after the cyber-attack (i.e., explicit failure); H2: The explicit failure has a greater effect on trust than the silent failure; H3: Perceived risk would be higher for the explicit failure than the silent.

## Method, material and experimental procedure

The sample consisted of 38 participants, including 22 males and 16 females. One female participant dropped out from study because of simulator sickness. They were 36.2 year-old on average (SD = 12.5) with 0 to 43 years of driving experience either in the UK or in the European Union (M = 15.7, SD = 13.1). They drove on average 7734 miles a year (min = 0, max = 20000, SD = 5891). They were free to withdraw from the study at any time.

The experiment was carried out in a driver in-the-loop simulator with a full-body vehicle and three degrees of freedom. A 7" touchscreen display allowed drivers to activate the automated driving mode, see the status of vehicle (i.e., manual vs. automated mode) and create a user profile including name, surname, email and password (Figure 1). These details would later be used to replicate the ransomware attack, following the method described in (Wolf and Lambert, 2017).

Figure 1. (left) Driver's view of the dashboard and HMI used for activating the automation (right) Driver in-the-loop simulator and its 4.75 meter, 270 degree curved screen.

Participants were semi-randomly assigned to each experimental group: early vs. late failure i.e. they experienced the screen malfunction of the SAE level 3 car either 2 min or 10 min after the system was engaged. The scenario consisted of 15 miles of suburban roads and motorway. After a familiarisation trial where participants drove the car and engaged the automation, they completed three counter-balanced 12 min trials, namely *no failure*, *silent failure* and *explicit failure*. The within-subject factor was the type of failure with two conditions: silent (i.e., turn signal failed to activate when automated car performed an overtaking manoeuvre) or explicit (i.e., ransomware). Participants were prompted to engage in a word search task upon self-activation of the automated driving system. They were allowed to resume manual control and reengage the system at any time. After each trial, participants filled in a survey including questions on trust, perception of risk and demographics. The study lasted approximately 2 hours and participants were given a £20 voucher as a compensation for their time.

## Measures

Participants were administered the trust in automation scale (TAS) (Körber, 2018) before and after the study, to assess dispositional trust. After each trial, situational trust was measured using the Situational Trust Scale – Automated Driving (STS-AD, Holthausen, 2020). They also indicated their opinion on a bespoke 5-point Likert scale (ranging from 1: *strongly disagree* to 5: *strongly agree*), for the following two items: *I would recommend someone else to trust this conditionally automated vehicle* (Recommend) and *I think it is necessary to trust vulnerable conditionally automated vehicles* (Trust vulnerable). The reason underlying the use of bespoke items was to assess to what extent individuals would recommend others to trust the system, and the effect of the vulnerability of the system on self-reported trust.

Perception of risk (POR) was assessed using an adapted version of the three-item scale for perceived relational risk (1: *strongly disagree* to 5: *strongly*

*agree*; Rajaonah et al., 2008; Li et al., 2019): *The conditionally automated vehicle is risky* (POR1); *Using the conditionally automated vehicle increases the risk of having a road accident* (POR2); and *Using the conditionally automated vehicle forces me to take a lot of risks* (POR3). For this measure, two participants' scores were not analysed due to data loss when extracting the data.

## Results

Concerning the TAS, five paired-samples T-tests indicated a significant effect of the conditions on the level of trust for three dimensions (Table 1 and Figure 2).

Table 1. Paired t-test values for the trust in automation scale measured before and after the experiment. Asterisks indicate significant effects.

| Dependent variable | t | df | p value | M (SD) before | M (SD) after |
|---|---|---|---|---|---|
| Reliability/Competence | 2.2 | 36 | .035* | 3.95 (.47) | 3.69 (.69) |
| Understanding/Predictability | 2.6 | 36 | .014* | 4.35 (.56) | 4.01 (.81) |
| Familiarity | -1 | 36 | .334 | 2.16 (1.06) | 2.30 (1.23) |
| Intention of developers | 2.1 | 36 | .040* | 4.30 (.70) | 4.05 (.81) |
| Propensity to trust | 1.2 | 36 | .255 | 3.53 (.53) | 3.41 (.69) |
| Trust in Automation | 1 | 36 | .345 | 3.76 (.73) | 3.55 (1.12) |

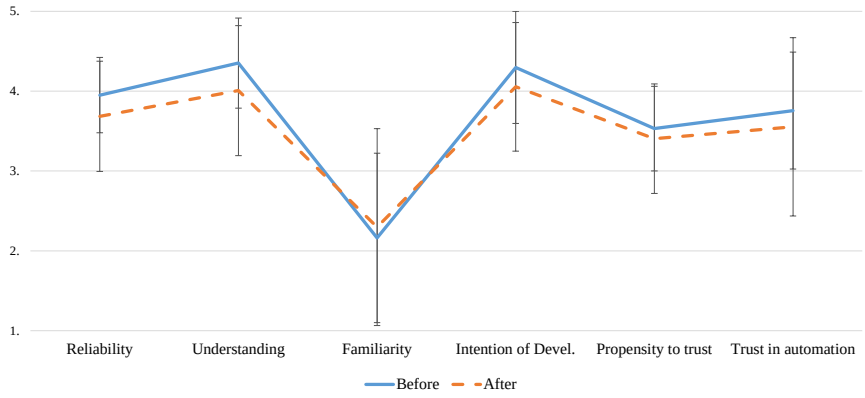


Figure 2. Mean and SD values for the factors from the TAS scale before and after the trial

The experimental conditions did not significantly affect STS-AD.

With respect to the two bespoke trust items, two repeated measures ANOVA showed that there was a significant effect of the condition on the *Recommend*

($F(2, 33) = 4.96$, $p = .013$, $\eta^2_p = .23$) and *Trust vulnerable* items ($F(2, 33) = 6.3$, $p = .005$, $\eta^2_p = .28$). Mauchly's Test of Sphericity indicated that the assumption of sphericity for both comparisons had not been violated, $\chi^2(2) = 4.34$, $p = .114$; $\chi^2(2) = 1.00$, $p = .61$. A post hoc pairwise comparison using the Bonferroni correction showed that the *Recommend* and *Trust vulnerable* scores were lower ($p < .05$) after the explicit failure ($M_{rec} = 2.89$, $SD_{rec} = 1.2$; $M_{vul} = 2.77$, $SD_{vul} = 1.3$) compared to both the control ($M_{rec} = 3.49$, $SD_{rec} = .92$; $M_{vul} = 3.26$, $SD_{vul} = .98$) and silent failure ($M_{rec} = 3.54$, $SD_{rec} = .98$; $M_{vul} = 3.4$, $SD_{vul} = 1.1$) conditions.

Perception of risk scores for did not significantly vary after experiencing the failures and were moderate (Table 2). However, an interaction effect was found between the order of the *explicit failure* condition and the perception of risk during the *explicit failure* condition (Pillai's Trace = .29, $F(4, 68) = 2.91$, $p = 0.03$, $\eta^2_p = 0.15$). A further exploration of pairwise comparisons showed the ratings for *Using the conditionally automated vehicle forces me to take a lot of risks* ($M = 3.21$, $SD = 1.25$, $p = 0.02$) were significantly greater than *Using the conditionally automated vehicle increases the risk of having a road accident* ($M = 2.43$, $SD = 1.09$, $p = 0.02$) when *explicit failure* was the second condition.

Table 2. Perception of risk scores after each experimental condition.

| Dependent variable | M control (SD) | M silent (SD) | M explicit (SD) |
|---|---|---|---|
| POR 1 (risky) | 2.92 (1.12) | 2.8 (.94) | 2.89 (1.13) |
| POR 2 (road accidents) | 2.86 (1.21) | 2.6 (1.01) | 2.76 (1.14) |
| POR 3 (take risks) | 2.86 (1.29) | 2.51 (1.01) | 2.81 (1.29) |

## Discussion

This driving simulator research explored the effects of software malfunctions and ransomware on trust during conditionally automated driving. The findings evidenced that silent and explicit failures occurring in a conditionally automated vehicle had a detrimental effect on the following factors of the TAS: Reliability/Competence, Understanding/Predictability, and Intention of Developers. This was congruent with previous work that showed a decrease in dispositional trust in automation due to a lack of Reliability/Competence (Fu et al., 2020; Kraus et al., 2020; Liu et al., 2021), Understanding/Predictability and negatively perceived Intention of Developers (Alonso and De La Puente, 2018; Kraft et al., 2020; Kraus et al., 2020). H1a and b were supported.

No significant variations in situational trust (STS-AD) were observed, which was not congruent with results from previous research. However, results from the two bespoke items indicated that trust scores after the *explicit failure* were

significantly lower than the *silent failure* and *control* conditions. It is important to stress that these trust scores were in essence low (i.e., lower than the median point of the scale) meaning that drivers distrusted the system. This finding supported H2 and was aligned with a number of studies identifying cyber threats as a critical factor affecting negatively trust in automated vehicles (Khan et al., 2020; Seetharaman et al., 2021). Furthermore, suspicion stemming from uncertainty, perception of malintent, and cognitive effort in trying to generate an explanation for the cyber-attack, may explain the lower level of trust after the ransomware compared to the other two conditions (Bobko et al., 2014). This claim seems to be supported by the interaction between the perception of risk after the *explicit failure* and the order of the *explicit failure* condition. When the cyber-attack occurred in second position, participants already had some experience to know that automated driving was reliable, but not enough to ignore the fact that they were hacked. Such uncertainty could explain why participants reported that using the AD resulted in taking more risks in general than taking more risks to have a road accident. It could mean that they made a distinction between being hacked and the risk of having a road accident. In other words, drivers seemed to differentiate an automated driving system from a connected driving system. AD may increase the risk of being hacked (i.e., connected vehicle) but because the vehicle was handling the driving task safely despite the cyber-attack (i.e., automated driving), drivers could focus on sorting out the cyber-threat. According to this interaction, having to deal with the cyber-threat when driving manually would have been more risky than in automated mode. Hence, suffering a cyber-attack while AD is engaged could be safer than while driving manually. Other interpretations of these findings are also possible. H3 was partially supported.

Overall, these results suggested that system malfunctions in conditionally automated driving had a negative effect on both dispositional, longer-term trust and learned, short-term trust. The ransomware explicit failure had a stronger negative effect on trust than the silent failure. A possible explanation for this is that the ransomware was conspicuous whereas the non-activation of the turn signals was not. Therefore, the ransomware was possibly perceived as riskier compared to the silent failure. The ransomware might have also stressed how vulnerable the automated driving system was. These data should be interpreted with caution because they are declarative only. Furthermore, the ransomware could have been perceived as a spam rather than a genuine cyber-attack, which may have tempered the effect of the explicit failure on trust and perception of risk. Further research should investigate the effect of both explicit and silent failures on driving behaviour, performance and safety.

## CONCLUSIONS

The novelty of this study is that two types of connected and automated driving failures and their effect on subjective trust have been examined: ransomware (i.e. explicit) vs. non-activation of the turn signals (i.e. silent). What is also innovative is that the explicit failure originates from an external source being outside the vehicle-driver system. Results showed that drivers trusted less the system after experiencing a ransomware cyber-attack than after a silent failure or no failure at all. Findings of the present study are important for designers and decision makers because they shed light on how external malevolent actors can impair trust in automated vehicles, provided that drivers should trust this vulnerable technology.

## ACKNOWLEDGMENTS

This work was supported by the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1).

## REFERENCES

Alonso, V., and De La Puente, P. (2018). System transparency in shared autonomy: A mini review. In Frontiers in Neurorobotics (Vol. 12, Issue November, p. 83).

Antinyan, V. (2020, November). Revealing the complexity of automotive software. In Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (pp. 1525-1528).

Bobko, P., Barelka, A. J., & Hirshfield, L. M. (2014). The construct of state-level suspicion: A model and research agenda for automated and information technology (IT) contexts. *Human Factors*, *56*(3), 489–508.

Carsten, O., and Martens, M. H. (2019). How can humans understand their automated cars? HMI principles, problems and solutions. *Cognition, Technology & Work, 21*(1), 3-20.

Deng, F., Lv, Z., Qi, L., Wang, X., Shi, M., & Liu, H. (2020). A big data approach to improving the vehicle emission inventory in China. *Nature Communications, 11*(1), 1-12.

Fu, E., Hyde, D., Sibi, S., Johns, M., Fischer, M., and Sirkin, D. (2020). Assessing the Effects of Failure Alerts on Transitions of Control from Autonomous Driving Systems. *IEEE Intelligent Vehicles Symposium, Proceedings*, 1956–1963.

Hoff, K.A. and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), pp.407-434.

Holthausen, B.E. (2020). Development and validation of the situational trust scale for automated driving (STS-AD) (Doctoral dissertation, Georgia Institute of Technology).

International Organization for Standardization. ISO/SAE DIS 21434 Road Vehicles - Cybersecurity engineering (2020).

Khan, S. K., Shiwakoti, N., Stasinopoulos, P., & Chen, Y. (2020). Cyber-attacks in the next-generation cars, mitigation techniques, anticipated readiness and future directions. *Accident Analysis and Prevention*, *148*.

Khastgir, S., Birrell, S., Dhadyalla, G. and Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies, 96*, pp.290-303.

Körber, M. (2018, August). Theoretical considerations and development of a questionnaire to measure trust in automation. In Congress of the International Ergonomics Association (pp. 13-30). Springer, Cham.

Kraft, A. K., Maag, C., Cruz, M. I., Baumann, M., & Neukum, A. (2020). Effects of explaining system failures during maneuver coordination while driving manual or automated. *Accident Analysis and Prevention*, *148*, 105839

Kraus, J., Scholz, D., Stiegemeier, D., and Baumann, M. (2020). The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency. *Human Factors*, *62*(5), 718–736.

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

Li, M., Holthausen, B.E., Stuck, R.E. and Walker, B.N. (2019, September). No risk no trust: Investigating perceived risk in highly automated driving. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 177-185).

Liu, P., Jiang, Z., Li, T., Wang, G., Wang, R., and Xu, Z. (2021). User experience and usability when the automated driving system fails: Findings from a field experiment. *Accident Analysis & Prevention*, *161*, 106383.

Louw, T., Kuo, J., Romano, R., Radhakrishnan, V., Lenné, M. G., and Merat, N. (2019). Engaging in NDRTs affects drivers' responses and glance patterns after silent automation failures. *Transportation Research Part F: Traffic Psychology and Behaviour, 62*, 870-882.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Academy of Management*, *20*(3), 709–734.

National Transportation Safety Board (2017). Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Williston, Florida, May 7, 2016. Available at: <https://dms.ntsb.gov/pubdms/> (accessed 07 Feb 2022).

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

Payre, W., Birrell, S. and Parkes, A. (2021). Although autonomous cars are not yet manufactured, their acceptance already is. *Theoretical Issues in Ergonomics Science, 22*(5), 567-580.

Payre, W., Cestac, J., Dang, N. T., Vienne, F., and Delhomme, P. (2017). Impact of training and in-vehicle task performance on manual control recovery in an automated car. *Transportation Research Part F: Traffic Psychology and Behaviour, 46*, 216-227.

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.

Rajaonah, B., Tricot, N., Anceaux, F. and Millot, P. (2008). The role of intervening variables in driver–ACC cooperation. *International journal of human-computer studies, 66*(3), pp.185-197.

Riley, V. (1996). Operator reliance on automation: Theory and data. In *Automation and human performance: Theory and applications* (pp. 19-35). CRC Press.

Seetharaman, A., Patwa, N., Jadhav, V., Saravanan, A.S. and Sangeeth, D. (2020). Impact of Factors Influencing Cyber Threats on Autonomous Vehicles. *Applied Artificial Intelligence*, pp.1–28.

Sheehan, B., Murphy, F., Mullins, M., and Ryan, C. (2019). Connected and autonomous vehicles: A cyber-risk classification framework. *Transportation Research Part A: Policy and Practice, 124*, 523-536.

Stuck, R.E., Tomlinson, B.J. and Walker, B.N. (2021). The importance of incorporating risk into human-automation trust*. Theoretical Issues in Ergonomics Science*, pp.1-17.

Wintersberger, P., Nicklas, H., Martlbauer, T., Hammer, S., and Riener, A. (2020, September). Explainable automation: Personalized and adaptive uis to foster trust and understanding of driving automation systems. In 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 252-261).

Wolf, M. and Lambert, R. (2017). Hacking trucks-cybersecurity risks and effective cybersecurity protection for heavy duty vehicles. Automotive-Safety & Security 2017-Sicherheit und Zuverlässigkeit für automobile Informationstechnik.