2019 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
AUTONOMOUS GROUND SYSTEMS (AGS) TECHNICAL SESSION
AUGUST 13-15, 2019 - NOVI, MICHIGAN

ERROR TYPE, RISK, PERFORMANCE, AND TRUST: INVESTIGATING THE IMPACTS OF FALSE ALARMS AND MISSES ON TRUST AND PERFORMANCE

Huajing Zhao[1], Hebert Azevedo-Sa[1], Connor Esterwood[2], X. Jessie Yang, PhD[1,3], Lionel Robert, PhD[1,2], Dawn Tilbury, PhD[1,4]

[1] Robotics Institute
[2] School of Information
[3] Department of Industrial and Operations Engineering
[4] Department of Mechanical Engineering

University of Michigan, Ann Arbor, MI

ABSTRACT

Semi-autonomous vehicles are intended to give drivers multitasking flexibility and to improve driving safety. Yet, drivers have to trust the vehicle's autonomy to fully leverage the vehicle's capability. Prior research on driver's trust in a vehicle's autonomy has normally assumed that the autonomy was without error. Unfortunately, this may be at times an unrealistic assumption. To address this shortcoming, we seek to examine the impacts of automation errors on the relationship between drivers' trust in automation and their performance on a non-driving secondary task. More specifically, we plan to investigate false alarms and misses in both low and high risk conditions. To accomplish this, we plan to utilize a 2 (risk conditions) × 4 (alarm conditions) mixed design. The findings of this study are intended to inform Autonomous Driving Systems (ADS) designers by permitting them to appropriately tune the sensitivity of alert systems by understanding the impacts of error type and varying risk conditions.

Citation: H. Zhao, H. Azevedo-Sa, C. Esterwood, X. J. Yang, L. Robert, D. Tilbury, "Error Type, Risk, Performance, and Trust: Investigating the Different Impacts of false alarms and misses on Trust and Performance", In *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium* (GVSETS), NDIA, Novi, MI, Aug. 13-15, 2019.

1. INTRODUCTION

When operating semi-autonomous vehicles, drivers are expected to take advantage of the automated aids to increase their productivity and improve their safety. Trust in Automation (TiA) is a fundamental factor to allow drivers to leverage the features provided by Autonomous Driving Systems (ADSs) [29]. We seek to investigate the

impacts that false alarms and misses have on trust and performance in a semi-autonomous driving context. In addition, this study will investigate the impact of varying external risk levels represented by different road conditions. Understanding these impacts is important because the consequences of different ADS errors should be taken into account when designing these systems.

ADSs are an increasingly pervasive force in the modern world and are building the path for the use of Autonomous and Semi-Autonomous Vehicles (AVs and SAVs) [31]. As a result, one can begin to see the role of the future human driver as more fluid than that of today's driver. Bringing the concept of human–automation teaming [30] to the driver–SAV context, future drivers are more likely to conduct secondary non-driving-related tasks (NDRT) while teaming with ADSs. Effective human–automation teaming and, consequently, driver–SAV teaming, require the human agent to monitor the automation and eventually help it to conclude specific tasks [4-7].

The fact that no system is perfect requires human drivers to frequently adjust their level of dependence [3]. Several studies have looked at dependence and the role that trust plays in how drivers respond to system errors [3, 8, 9]. In this literature, errors have largely been classified into two types: *false alarms* and *misses*, in accordance with the definitions of signal detection theory (SDT) [10-12].

In this paper we approach the subjects of error type, trust, and performance. In addition, we investigate the moderating impact that risk might have on this relationship. The next sections are ordered as follows: Section 2 introduces a theoretical basis from the literature related to automation error types, trust, risk and performance; Section 3 presents the hypotheses of this study and the rationales behind them; Section 4 brings details about the experiment to be conducted; and Section 5 discusses the possible implications of the expected results for future work.

2. BACKGROUND

2.1 Error Type & Task Performance

Generally, automation errors can be classified as false alarms or misses. A false alarm occurs when the automation alerts the human operator that it has detected something when in fact nothing is actually there. Misses occur when the automation fails to alert the operator that it has detected something when in fact it should have detected something. False alarms are a Type I error while misses are a Type II error.

Prior research on automation has found significant differences between the impact of false alarms and that of misses on task performance. Although not specifically examining an ADS technology, Wickens et al. [28] found that the performance on operating the automation degraded more when operators were given a higher rate of false alarms, whereas the performance on a non-driving secondary task degraded more with higher rates of misses. Consistent with these findings and in an ADS context, Sanchez et al. [26] found that false alarms led to lower performance on the primary operating or driving task, while misses led to lower NDRT performance.

Based on the evidence of these two studies, it appears that both error types have an overall negative impact on performance but that misses seem to have a stronger negative impact on NDRT performance than false alarms. How trust and risk could act as moderators in this relationship is discussed in Section 2.2.

2.2 Error Type, Trust & Risk

Trust: Trust has been investigated rigorously in relation to human–automation interaction. An extensive review of trust in the human–automation domain was conducted by Lee and See [17]. This work highlighted the critical components for trust formation, as well as the three bases of human–automation trust: performance, process, and purpose. In addition, many other authors have consistently presented vulnerability as a fundamental aspect of trust [18].

In relation to trust and error type, Chancey et al. [3] found that false alarms had a stronger impact on trust than misses. Furthermore, trust appeared to moderate the relationship between error type and two different responsive behaviors. From the results of this study, we can conclude that both error types impact trust and they do so differently.

Risk: In our study, risk is defined as the subjective degree of uncertainty associated with a given situation. Risk has been seen as an essential component for trust, where trust was characterized as a willingness to be vulnerable to another party [18-20]. If trust can be seen as the act of being vulnerable, then risk is likely to be a moderating factor in trust-building phenomena. For example, in a low-risk situation, an individual might be more willing to be vulnerable—i.e. trusting—than in a high-risk situation.

Chancey et al. [3] attempted to explain the different impacts that error type has on trust using risk as an interacting factor, but they failed to produce significant results. Petersen et al. [27], however, investigated risk in relation to trust in ADSs and found that risk as related to automation errors impacts trust significantly; these authors did not investigate misses, opting to include only false alarms in their study.

An ADS literature review uncovered a lack of investigation into risk as a moderating factor in the relationships involving trust and performance. We seek to address this issue by introducing false alarms and misses as different imperfect alarm conditions for an ADS. In addition, our study will add real-world consequences for participants when the automation fails, an attempt to create a more salient trustor/trustee relationship. Overall we intend to (1) evaluate the impacts of different ADS error types (false alarms and misses) on trust and performance and (2) investigate the role of risk as a potential moderator of these relationships. The following section presents our hypotheses and expected outcomes for this study.

3. HYPOTHESES AND EXPECTED RESULTS

3.1 Hypotheses 1 (H1) and 2 (H2)

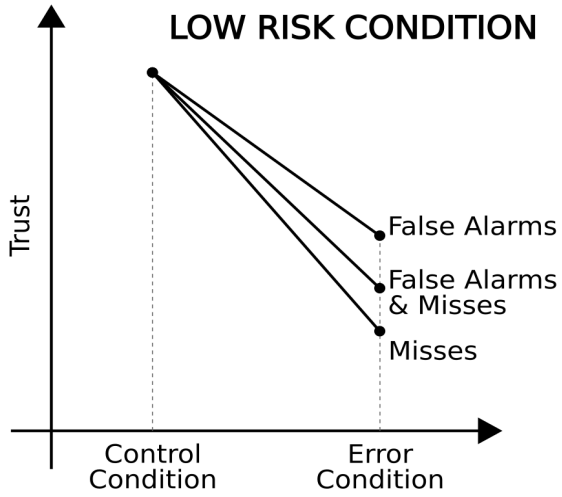
We hypothesize that people regard misses as more harmful to their safety than false alarms, leading to a larger drop in trust. In short, false alarms can be a nuisance but misses can actually lead to crashes.

H1: Under both high- and low-risk conditions, misses have a stronger negative effect on trust than false alarms.

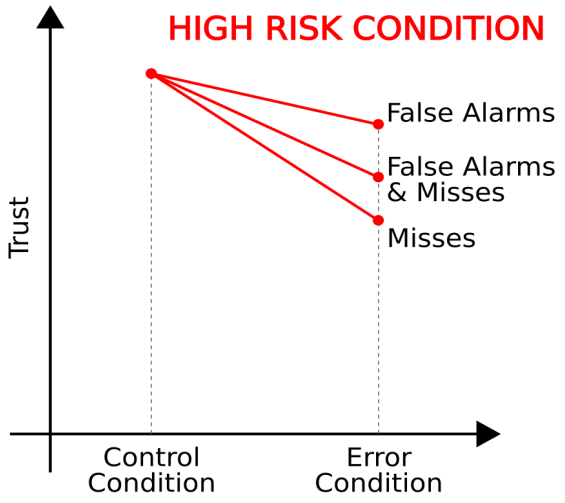
In this study we plan to manipulate risk conditions by varying the type of road paths: straight or curvy. Straight roads will represent our low-risk condition while curvy roads will represent our high-risk condition. We speculate that people might perceive the straight road condition to be easier for the automation to handle when compared to the curvy roads. Therefore, participants should have higher expectations for the automation capabilities in low-risk situations, while they might be more lenient with automation error in high-risk situations.

H2: Both types of errors reduce trust more in the low-risk condition than in the high-risk condition.

Figure 1 shows a pictorial representation of hypotheses H1 and H2.



(a) Low-risk Condition



(b) High-risk Condition

Figure 1. H1 and H2 – Under both high-risk and low-risk conditions, *misses* have a stronger negative effect on trust than *false alarms*. The negative impact of error types—both individually and combined—is stronger in the low-risk condition.

3.2 Hypotheses 3 (H3) and 4 (H4)

In addition to H1 and H2, we plan to consider the impacts of automation errors on performance. We expect that subjects will perform worse in the presence of automation errors but that misses will be more impactful than false alarms. We expect to find that misses are more likely to induce crashes or at least require more attention from the drivers. In other words, when drivers realize that the automation misses some of the obstacles, they will be compelled to pay more attention to the driving task. Moreover, the road shape will represent an additional difficulty to drivers, and misses in a high external risk condition (curvy roads) will be more prejudicial to NDRT performance. Our third and fourth hypotheses summarize these suppositions.

H3: Both types of errors reduce NDRT performance but misses reduce NDRT performance more than false alarms.

H4: The negative effects of both types of errors on NDRT performance are more profound in the high-risk condition than in the low-risk condition.

Figure 2 presents hypotheses H3 and H4.

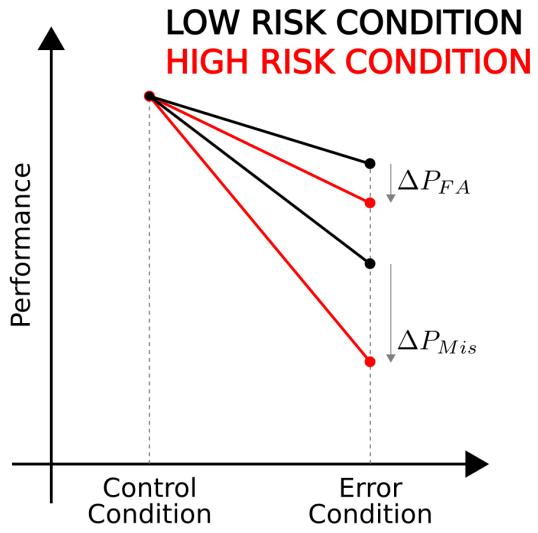


Figure 2: H3 and H4 - high risk has a stronger negative impact on performance than low risk, and misses have a stronger negative impact on performance than false alarms. Here, ΔP_{FA} denotes the difference of trust decrease due to false alarms between high risk and low risk conditions, and ΔP_{Mis} denotes that due to misses between high risk and low risk conditions. The hypotheses expect $|\Delta P_{FA}| < |\Delta P_{Mis}|$.

4. METHOD

4.1 Participants and Compensation

We plan to recruit 80 participants. Participants are to be compensated according to their NDRT final score. The base rate of compensation will be $15. Each participant will be eligible for a cash bonus up to $35. There will be four bonus levels, assigned according to participants' performance: Bronze ($0 bonus), Silver ($5 bonus), Gold ($15 bonus), and Platinum ($35 bonus).

4.2 Experimental Task

The simulation part of the experiment was designed and will be implemented with the *Autonomous Navigation Virtual Environment Laboratory (ANVEL) Simulator* [21], and the NDRT will be implemented as an adapted version of the Surrogate Reference Task [4], with The Psychology Experiment Building Language (PEBL) [22].

Subjects are to operate a simulated vehicle with ADS features (i.e. automatic lane keeping, cruise control, and collision avoidance systems). In parallel, they are to perform a visual search NDRT where they need to find a "Q" character among many "O" characters. Participants will gain 1 point for each correctly chosen "Q" and lose 5 points each time the emergency brake (collision avoidance system) gets activated. Figure 3 shows the experimental setup while Figures 4 and 5 show the driving and non-driving tasks, respectively.

Figure 3: Experimental Setup.

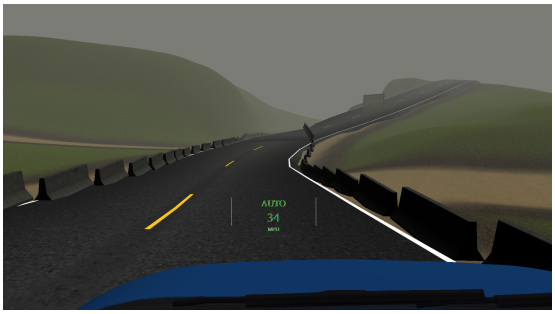


Figure 4: Driving Task.

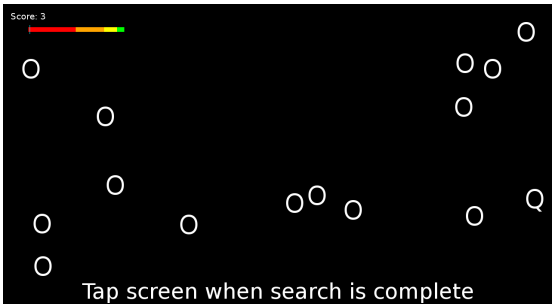


Figure 5: Non-driving Task.

4.3 Experimental Design

A 2 (high vs. low external risk) x 4 (error types) mixed design is proposed. The external risk conditions will be manipulated by the road condition: straight roads (low external risk) and curvy roads (high external risk). Additionally, there will be four alarm conditions: control condition, where there will be no errors; false alarm condition, where the system will provide sound alarms with the message "*Stopped vehicle ahead. Take control now!*" but there will be no obstacles on the road; misses condition, where the system will be unable to recognize and warn the driver about an obstacle on the road; and the combined false alarm and misses condition, where both false alarms and misses will be present. Each participant will be randomly assigned to one of the four alarm conditions, in both straight and curvy road conditions (representing the low and high external risk conditions, respectively). These conditions will be counterbalanced with a Latin square design to minimize learning and ordering effects. All eight conditions are shown in Table 1. Each participant is to experience both external risk conditions but only one alarm condition, configuring a total of 20 participants for each pair of *a* and *b* conditions.

Dependent variables include participants' subjective responses, behavioral responses, and task performance, as well as vehicle dynamics data.

Subjective data are to be gathered through surveys before, during, and after each drive, including trust perception, risk perception, and workload perception. Behavioral responses (dynamic eye movement tracking) and performance will also be collected. From our previous experience, we have verified that eye gaze *monitoring ratio* is the most important and significant measure of trusting behaviors in our setup. Monitoring ratio is the ratio of time drivers spend looking at the driving scene to the time they spend looking elsewhere [23].

Vehicle dynamics data will provide the characteristics of the state of the vehicle during the simulation time. It is possible to gather data from the vehicle's pose, velocity, acceleration, steering, and pedal inputs, and other similar metrics available from the simulation environment.

4.4 Experimental Procedure

Initially, all participants will complete the consent forms and a pre-experiment survey

Table 1: Manipulated independent variables defining each condition in the experiment. Each participant will experience both External Risk conditions and one Alarm Condition (for example, Conditions 3a and 3b).

		EXTERNAL RISK	
		LOW	HIGH
ALARMS	CONTROL	Condition 1a	Condition 1b
	FALSE ALARMS	Condition 2a	Condition 2b
	MISSES	Condition 3a	Condition 3b
	FALSE ALARMS & MISSES	Condition 4a	Condition 4b

related to their personal information, experience with ADSs, their mood, and their initial propensity to trust in automation. After the survey, the experimenters will explain the tasks and give details about the simulated vehicle control and the dynamics of the experiment. All participants will have the opportunity to complete a training session before the actual experiment begins. In sequence, they will have the eye tracker fitted and calibrated, and will complete two trials (one for each external risk condition). After the trials and at the end of the experiment, participants will be asked to complete post-trial surveys related to their trust in automation and perceived workload.

These surveys will be administered electronically. Each experiment will last approximately 60 minutes.

5. DISCUSSION

This study is expected to take place during the summer of 2019. We intend to analyze the gathered data and obtain the results in the fall. The hypotheses and the methodology were designed considering the information synthesized from the results of our previous studies [4].

The ultimate goal of our overarching project is to examine and understand the factors that influence drivers' trust in ADSs, and possibly define a framework to measure and manipulate humans' and autonomies' trust levels. We believe that such a framework can enhance drivers' safety and effectiveness by optimizing driving and secondary task performances. Within this framework we intend to develop techniques to identify opportunities for shifting the control authority between the driver and the vehicle, i.e. (1) predict when the driver is likely to give or take control of the driving to the vehicle's autonomy and (2) predict when the vehicle's autonomy should give or take control of the driving to the driver.

If the presented hypotheses hold true, we should be able to control drivers' trust levels by introducing simulated imperfections in the ADS behavior or even by providing drivers more information to increase their situational awareness. As the main goal for our future work, we aim to use this scheme for optimizing NDRT performance levels.

We expect to contribute to the existing literature on ADS trust, risk, and human–automation teaming as well as explore the connection of all these factors to signal detection theory. This study should expand our knowledge about the impacts of error types and risks on drivers' trust and performance in semi-autonomous driving. The results of this study are intended to inform the design and development of ADSs by helping to determine the operational

requirements for the reliability of those systems' alarms.

6. ACKNOWLEDGMENTS

This research is supported in part by the Automotive Research Center (ARC) at the University of Michigan, with funding from government contract DoD-DoA W56HZV-14-2-0001, through the U.S. Army Combat Capabilities Development Command (CCDC)/Ground Vehicle Systems Center (GVSC). We greatly appreciate the guidance of Victor Paul (GVSC), Ben Haynes (GVSC), Dariusz Mikulski (GVSC) and Jason Metcalfe (ARL) in helping design the study. The authors would also like to thank Quantum Signal, LLC, for providing ANVEL software and their development support.

7. DISCLAIMER

Reference herein to any specific commercial company, product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the Department of the Army (DoA). The opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or the DoA and shall not be used for advertising or product endorsement purposes

8. REFERENCES

[1] Ajzen, I., *The theory of planned behavior.* Organizational behavior and human decision processes, 1991. **50**(2): p. 179-211.

[2] Sorkin, R.D. and D.D. Woods, *Systems with Human Monitors: A Signal Detection Analysis.* Human–Computer Interaction, 1985. **1**(1): p. 49-75.

[3] Chancey, E.T., et al., *Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence.* Human factors, 2017. **59**(3): p. 333-345.

[4] Petersen, L., et al., *Effects of Augmented Situational Awareness on Driver Trust in Semi-Autonomous Vehicle Operation*, in *Autonomous Ground Systems (AGS) Technical Session of the 2017 Ground Vehicle Systems Engineering and Technology Symposium*. 2017: Novi, MI.

[5] Michelson, S., *Human Centered Teaming of Autonomous Battlefield Robotics*, in *Ground Vehicle Systems Engineering and Technology Symposium*. 2018: Novi, MI.

[6] Parasuraman, R., M. Barnes, K. Cosenzo, and S. Mulgund, *Adaptive automation for human-robot teaming in future command and control systems*. 2007: Army Research Lab - Aberdeen Proving Ground. Human Research and Engineering Directorate.

[7] Shively, R.J., et al., *Why human-autonomy teaming?* in *International conference on applied human factors and ergonomics*. 2017: Springer.

[8] Chancey, E.T., et al., *Effects of Alarm System Error Bias and Reliability on Performance Measures in a Multitasking Environment: Are False Alarms Really Worse than Misses?* Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2017. **61**(1): p. 1621-1625.

[9] Bliss, J.P., R.D. Gilson, and J.E. Deaton, *Human probability matching behaviour in response to alarms of varying reliability.* Ergonomics, 1995. **38**(11): p. 2300-2312.

[10] Green, D.M. and J.A. Swets, *Signal detection theory and psychophysics*. Vol. 1. 1966: Wiley New York.

[11] Swets, J.A., *Measuring the accuracy of diagnostic systems.* Science, 1988. **240**(4857): p. 1285-1293.

[12] Wickens, T.D., *Elementary signal detection theory*. 2002, Oxford; New York: Oxford University Press.

[13] Meyer, J., *Effects of warning validity and proximity on responses to warnings.* Human factors, 2001. **43**(4): p. 563-572.

[14] Rice, S., *Examining single- and multiple-process theories of trust in automation.* The journal of general psychology, 2009. **136**(3): p. 303-322.

[15] Dixon, S.R. and C.D. Wickens, *Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation*

dependence in high workload. Human factors, 2006. **48**(3): p. 474-486.

[16] Dixon, S.R., C.D. Wickens, and J.S. McCarley, *On the independence of compliance and reliance: Are automation false alarms worse than misses?* Human factors, 2007. **49**(4): p. 564-572.

[17] Lee, J.D. and K.A. See, *Trust in automation: designing for appropriate reliance.* Human factors, 2004. **46**(1): p. 50-80.

[18] Mayer, R.C., J.H. Davis, and F.D. Schoorman, *An integrative model of organizational trust.* Academy of management review, 1995. **20**(3): p. 709-734.

[19] Wilde, G.J., *Risk homeostasis theory: An overview.* Injury prevention, 1998. **4**(2): p. 89-91.

[20] Pavlou, P.A., *Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model.* International journal of electronic commerce, 2003. **7**(3): p. 101-134.

[21] Durst, P.J., et al. *A real-time, interactive simulation environment for unmanned ground vehicles: The autonomous navigation virtual environment laboratory (ANVEL)*. in *Information and Computing Science (ICIS), 2012 Fifth International Conference on*. 2012. IEEE.

[22] Mueller, S.T. and B.J. Piper, *The psychology experiment building language (PEBL) and PEBL test battery.* J neurosci methods, 2014. **222**: p. 250-9.

[23] Hergeth, S., et al., *Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving.* Human factors, 2016. **58**(3): p. 509-519.

[24] Petersen, L., Robert, L., Yang, X., and Tilbury, D., *Situational Awareness, Driver's Trust in Automated Driving Systems and Secondary Task Performance*, SAE Intl. J CAV 2(2):2019, https://doi.org/10.4271/12-02-02-0009.

[25] Dixon, S.R., C.D. Wickens, and J.S. McCarley, *On the independence of compliance and reliance: Are automation false alarms worse than misses?* Human factors, 2007. **49**(4): p. 564-572.

[26] Sanchez, J., W.A. Rogers, A.D. Fisk, and E. Rovira, *Understanding reliance on automation: effects of error type, error distribution, age and experience,* Theoretical issues in ergonomics science, 2014. **15**(2): p. 134-160.

[27] Petersen, L., H. Zhao, D. Tilbury, X.J. Yang, and L. Robert, *The Influence of Risk on Driver's Trust in Semi-Autonomous Driving.* Modeling & Simulation, Testing and Validation (MSTV) Technical Session August 7-9, 2018: Novi, MI.

[28] Wickens, C., S. Dixon, J. Goh, and B. Hammer, *Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis.* 2005. University of Illinois at Urbana Champaign.

[29] Metcalfe, J. S., Marathe, A. R., Haynes, B., Paul, V. J., Gremillion, G. M., Drnec, K., ... & Nothwang, W. D. (2017, May). Building a framework to manage trust in automation. In Micro-and Nanotechnology Sensors, Systems, and Applications IX (Vol. 10194, p. 101941U). International Society for Optics and Photonics.

[30] You, S. and Robert, L. P. (2019). Trusting Robots in Teams: Examining the Impacts of Trusting Robots on Team Performance and Satisfaction. Proceedings of the 52th Hawaii International Conference on System Sciences, Jan 8-11, Maui, HI, Forthcoming

[31] Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert Jr, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. Transportation Research Part C: Emerging Technologies, 104, 428-442.