Comparing the Effects of False Alarms and Misses on Humans' Trust in (Semi)Autonomous Vehicles

Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T. Esterwood, X. Jessie Yang,
Lionel P. Robert Jr., and Dawn M. Tilbury
University of Michigan
{azevedo,jskumaar,cte,xijyang,lprobert,tilbury}@umich.edu

ABSTRACT

Trust in automated driving systems is crucial for effective driver-(semi)autonomous vehicles interaction. Drivers that do not trust the system appropriately are not able to leverage its benefits. This study presents a mixed design user experiment where participants conducted a non-driving task while traveling in a simulated semi-autonomous vehicle with forward collision alarm and emergency braking functions. Occasionally, the system missed obstacles or provided false alarms. We varied these system error types as well as road shapes, and measured the effects of these variations on trust development. Results reveal that misses are more harmful to trust development than false alarms, and that these effects are strengthened by operation on risky roads. Our findings provide additional insight into the development of trust in automated driving systems, and are useful for the design of such technologies.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Automated driving systems; Trust; Human-robot teaming; Driving simulation

ACM Reference Format:
Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T. Esterwood, X. Jessie Yang, and Lionel P. Robert Jr., and Dawn M. Tilbury. 2020. Comparing the Effects of False Alarms and Misses on Humans' Trust in (Semi)Autonomous Vehicles. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20 Companion), March 23–26, 2020, Cambridge, United Kingdom.*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3371382.3378371

1 INTRODUCTION

Trust in automation is a fundamental factor for achieving the acceptance and use of advanced robotic technologies [7, 9, 14]. In the context of automated driving systems (ADSs), trust-related issues can jeopardize driver-vehicle interaction effectiveness and result in inefficient use of ADSs. For example, when undertrusting the ADS, a driver might not be able to fully leverage the safety and productivity benefits provided by that system. When overtrusting the ADS, however, drivers might not be attentive enough to the road hazards that the automation is not able to address or avoid.

This work focuses on understanding and modeling how *trust in the ADS* (TiA) develops in the interactions between drivers and (semi)autonomous vehicles with SAE level 3 ADSs and the execution of secondary tasks by the driver that demand visual attention. We investigate the effects of distinct ADS error types—i.e., false alarms and misses—and of risk factors perceived by drivers on TiA. These insights are important for the development of TiA control techniques, which will, ultimately, be helpful to avoid trust issues and improve the collaboration between drivers and (semi)autonomous vehicles.

2 BACKGROUND AND RELATED WORK

Trust in automation is considered a factor that directly influences a supervisor's intervention behavior [16]. Lee and See [8] define trust as *the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability.*

Researchers have investigated the influence of *false alarms*—when systems diagnose a risky condition that is in fact non-existent—and *misses*—when systems are not able to diagnose an existent risky condition to the user—on operators' *trusting behaviors* when they interact with automated systems. After being exposed to false alarms, operators were more prone to delay their response to automation alerts or even ignore them, reducing their *compliance* [2, 17]. On the other hand, after being exposed to misses, operators tended to allocate more attention to monitor the system and take over control without being asked, reducing their reliance. Compliance, reliance, and monitoring (i.e., vigilance) [15] are the most relevant behaviors in the driver-vehicle interaction context. These are the trusting behaviors that should be perceived and processed by smart ADSs that aspire to estimate and manipulate TiA.

3 HYPOTHESES DEVELOPMENT

Consider a driver operating a vehicle with the aid of an ADS and executing a concurrent non-driving related task (NDRT). The ADS is able to drive the vehicle if the road is free, as well as to eventually warn the driver about obstacles (i.e., stopped vehicles) on the road. However, the ADS might not be working perfectly, and there might be occasions when false alarms and misses occur. A miss-prone ADS could easily lead to crashes or more serious accidents and thus misses could be perceived as more harmful than false alarms. Further, when environmental characteristics are manipulated by varying the road type (i.e., straight vs. curvy), we expect drivers to perceive the situational risk differences and have low TiA when risk is high. In summary, we propose hypotheses H1, and H2:

HRI '20 Companion, Mar 23–26, 2020, Cambridge, United Kingdom.

ACM ISBN 978-1-4503-7057-8/20/03.
https://doi.org/10.1145/3371382.3378371

H1: *Both misses and false alarms have a negative effect on TiA. Misses have a stronger negative effect on TiA than false alarms, independent of the environmental conditions.*

H2: *In curvy roads, drivers perceive a high situational risk. Also, this perceived risk is negatively correlated with TiA.*

4 METHODOLOGY

The study employed a 4×2 mixed design, given by 4 ADS error type conditions—control (no ADS error), false alarms (4 false alarms in 12 events), misses (4 misses in 12 events), and combined (2 false alarms and 2 misses in 12 events)—as well as 2 road shape conditions—straight or curvy. We used a driving simulation implemented with the *Autonomous Navigation Virtual Environment Laboratory* (ANVEL) simulator [3]. In the driving task, participants operated a simulated vehicle equipped with an ADS that provided it self-driving capabilities (i.e., Automatic Lane Keeping, Cruise Control and Collision Avoidance systems) with the ability to give/take control to/from the ADS. With the ADS activated, participants were requested to execute a visual search non-driving related task (NDRT) implemented with PEBL [10]. A total of 80 participants, aged 18-51 years ($\mu_{AGE} = 25.0$, $\sigma_{AGE} = 5.7$), were recruited. Each subject experienced both road conditions and one of the ADS error type conditions. Measured variables included participants' subjective responses including trust [11], risk [13], and workload perceptions through surveys, behavioral responses and NDRT performance, as well as vehicle dynamics data. NDRT performance consisted of the total number of points obtained by the participants in each trial minus penalties for each time they did not take control on time and the emergency brakes were activated. Figure 1 shows the setup and the tasks performed by the participants.

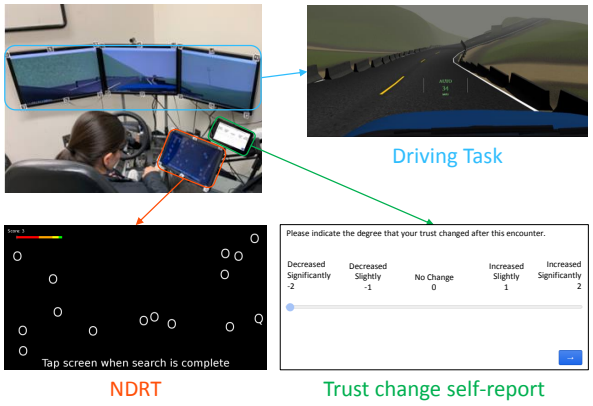

Figure 1: Experimental design, driving task, non-driving related task and trust self-report. The trust change self-report question popped up after every event within the trials.

5 RESULTS

We ran an ANOVA to verify H1 and found significant differences in final (i.e., post-trial) TiA between the four trial error type conditions ($F(3, 156) = 23.33$; $p < 10^{-3}$). We also ran a t-test to confirm that TiA is significantly greater for false alarms than for misses ($t(39) = 3.82$, $p < 10^{-3}$). These results confirm that misses have a stronger negative impact on TiA than false alarms (H1).

We used a linear mixed-effects model to investigate the impact of road shapes on post-trial perceived risk. We found that drivers perceived a higher risk when they operated the simulated vehicle in curvy roads ($p < 10^{-3}$). Moreover, we analyzed the correlation of perceived risk and TiA and confirmed that in more difficult driving situations, drivers do not trust the ADS to help them to drive and execute a concurrent NDRT safely ($p < 10^{-3}$).

6 DISCUSSION

We identified that misses have a stronger negative impact on trust than false alarms, in support of hypothesis H1. In our study, we used direct measures of trust, while previous studies have used different metrics [6] or even found different conclusions [1]. In support of H2, we have also described the negative influences of risk perception on TiA, which adds to the effects of false alarms and misses. These findings align with conclusions from existing literature on trust models [5, 12], and extend the results through the manipulation of road shapes.

There are limitations for this study. In general, people tend to act similarly in real and simulated environments [4]. However, due to the risks involved in driving, we acknowledge that participants might not have felt as vulnerable as they would if this study had been conducted in a real vehicle. Another restriction is that our NDRT is a very specific visual task. Other types of NDRTs could demand drivers attention for longer periods of time, and this could induce a different effect on trust, risk perception, and performance.

This work tries to identify the aspects of TiA development that could be included in computational models for TiA and that are useful for the development of new ADS functions, such as adapting the ADS's behavior to driver's behaviors. In future efforts, these models can be utilized for the design of frameworks for trust manipulation and control. These frameworks should have the goal of optimizing driver-(semi)autonomous vehicles team performances, mainly by avoiding trust-related issues. Future contributions may extend our analysis to characterize TiA short-term dynamics (i.e.: observing TiA at every event). Moreover, trusting behaviors (such as gaze movements) could be integrated in the analyses and correlated with subjects' self-reports of trust.

7 CONCLUSION

We present a user study where participants operated a simulated self-driving car while conducting a NDRT and reporting their level of trust in the system. Our results reveal that when drivers interact with ADSs and use these systems to execute a non-driving related task concurrently, misses are more harmful to trust development than false alarms. Moreover, the inclusion of risk from the operational environment also undermines trust development. While more accurate trust models are still required, our findings are useful for the design of driver-(semi)autonomous vehicles interactive systems.

ACKNOWLEDGMENTS

This research project is partially supported by the Automotive Research Center at the University of Michigan, through the U.S. Army CCDC/GVSC. We greatly appreciate the guidance of Victor Paul (GVSC) with the study design.
DISTRIBUTION A. Approved for public release; distribution unlimited. OPSEC #:3195

REFERENCES

[1] Eric T Chancey, James P Bliss, Yusuke Yamani, and Holly AH Handley. 2017. Trust and the compliance–reliance paradigm: the effects of risk, error bias, and reliability on trust and dependence. *Human factors* 59, 3 (2017), 333–345.

[2] Stephen R. Dixon, Christopher D. Wickens, and Jason S. McCarley. 2007. On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors* 49, 4 (aug 2007), 564–572. https://doi.org/10.1518/001872007X215656

[3] Phillip J Durst, Christopher Goodin, Chris Cummins, Burhman Gates, Burney Mckinley, Taylor George, Mitchell M Rohde, Matthew A Toschlog, and Justin Crawford. 2012. A real-time, interactive simulation environment for unmanned ground vehicles: The autonomous navigation virtual environment laboratory (ANVEL). In *2012 Fifth International Conference on Information and Computing Science*. IEEE, Shanghai, China, 7–10.

[4] Arsalan Heydarian, Joao P Carneiro, David Gerber, Burcin Becerik-Gerber, Timothy Hayes, and Wendy Wood. 2015. Immersive virtual environments versus physical built environments: A benchmarking study for building design and user-built environment explorations. *Automation in Construction* 54 (2015), 116–126.

[5] Y-TC Hung, Alan R Dennis, and Lionel Robert. 2004. Trust in virtual teams: Towards an integrative model of trust formation. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*. IEEE, Honolulu, HI, 11–pp.

[6] Jason D Johnson, Julian Sanchez, Arthur D Fisk, and Wendy A Rogers. 2004. Type of automation failure: The effects on trust and reliance in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48. SAGE, Los Angeles, CA, 2163–2167.

[7] Moritz Körber, Eva Baseler, and Klaus Bengler. 2018. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics* 39, 1 (2018), 9–21. https://doi.org/10.1016/j.apergo.2017.07.006

[8] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (jan 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[9] Jae-Gil Lee, Jihyang Gu, and Dong-Hee Shin. 2015. Trust In Unmanned Driving System. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. ACM, Portland, Oregon, USA., 7–8.

[10] Shane T Mueller and Brian J Piper. 2014. The psychology experiment building language (PEBL) and PEBL test battery. *Journal of neuroscience methods* 222 (2014), 250–259.

[11] Bonnie M. Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (mar 1996), 429–460. https://doi.org/10.1080/00140139608964474

[12] Luke Petersen, Huajing Zhao, Dawn Tilbury, X Jessie Yang, Lionel Robert, et al. 2018. The influence of risk on driver's trust in semi-autonomous driving. In *In Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium (GVSETS 2018)*. NDIA, Novi, MI, 1–7.

[13] Lionel P Robert, Alan R Denis, and Yu-Ting Caisy Hung. 2009. Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *Journal of Management Information Systems* 26, 2 (2009), 241–279.

[14] Paul Robinette, Michael Novitzky, Caileigh Fitzgerald, Michael R Benjamin, and Henrik Schmidt. 2019. Exploring Human-Robot Trust During Teaming in a Real-World Testbed. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Daegu, Korea., 592–593.

[15] Julian Sanchez, Wendy A Rogers, Arthur D Fisk, and Ericka Rovira. 2014. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science* 15, 2 (2014), 134–160.

[16] T.B. Sheridan, T. Vámos, and S. Aida. 1983. Adapting automation to man, culture and society. *Automatica* 19, 6 (nov 1983), 605–612. https://doi.org/10.1016/0005-1098(83)90024-9

[17] C Wickens, S Dixon, J Goh, and B Hammer. 2005. Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis (Tech Rep. No. AHFD-05-02). *Urbana-Champaign, IL: Univ. of Illinois* 21, 3 (2005), 3–12.