Assessing the Effects of Failure Alerts on Transitions of Control from Autonomous Driving Systems

Ernestine Fu, David Hyde, Srinath Sibi, Mishel Johns, Martin Fischer, David Sirkin

***Abstract*— Autonomous vehicle systems and their users need to collaborate to navigate the driving environment, particularly during an unstructured transition of control from automation, when the system releases control and expects the human to immediately assume driving responsibility. We investigated how such transitions affect the user's trust in the system and subsequent performance. In a full-vehicle driving simulator, participants encountered two system failures: the first varied in severity (mild or severe failure), and the second required a transition of control that was either detected and alerted (loud failure) or not (silent failure). We observed (i) significant changes in user trust in the system over time and between events, and (ii) that the first failure's severity level did not affect user performance in the subsequent failure; rather, the system's detection and alert both times was sufficient to successfully complete the transition of control.**

I. Introduction

As autonomous driving systems continue in their technical advances, the role of the operator is evolving—the human driver and the autonomous system need to collaborate with one another to safely navigate the driving environment [1]. This needed teamwork is most pronounced during the transition of control from automation, when the autonomous system hands control of the vehicle over to the human driver, often because the system has difficulty interpreting or navigating the driving situation, or because there is a failure in the autonomous system itself. Current systems alert the driver to the potential need for a transition and then release control of the vehicle, expecting the driver to quickly take responsibility of the driving task [2]. In these instances, effective design mechanisms are necessary to assist drivers in safely handling transitions of control.

Designers of these automation systems need to consider any number of complicated, yet likely, scenarios when specifying system actions and responses. For example, what happens when an autonomous system issues an alert during a system failure that potentially requires a transition, but then is able to successfully resume navigational control? How do such actions affect the user's trust when a subsequent transition is needed in a critical, potentially fatal, failure event—particularly when the system is unable to detect its own failure and provide an alert? And would the lack of an alert drastically reduce driver performance, given that it was issued earlier?

To address these questions, we developed a simulated driving experience (see Figure 1) that included two instances of failures, with experimental conditions that altered whether the system issued transition of control alerts or not. All participants encountered a first navigational failure, with an accompanying beeping alert warning of an expected mandatory transition of control. This first failure varied in severity, either a mild or a severe loss of lane tracking with accompanying swerving, depending on the participant's experimental condition; in both cases, the car was able to successfully navigate back onto its original path, so that the participants could continue the simulated drive. In the second failure, the vehicle either provided the same beeping alert, or it did not, in which case the car was unaware of its system failure, and thus did not know that the driver was needed to assume control. This allowed for four experimental conditions that we compared between.

Figure 1. Full-vehicle driving simulator with programmed failure events

Our results show that when an automation system alerts drivers of a failure and the potential but unnecessary need for a transition of control, then, when the system does not provide an alert during a critical failure, drivers are less likely to navigate that critical transition of control successfully. The severity level of a first failure does not seem to affect performance in a later failure event. We also observed significant changes in user trust in the system over time and between events.

II. Background

While we focus on autonomous vehicles, the work presented here builds on several broader areas of study related to human-system interaction. A fundamental question implicit in our study is whether and how humans use knowledge of past experiences with systems to inform their decision-making during future, and in particular, novel experiences with those systems. The simulation we programmed is based on capabilities of transition of control systems, along with

Ernestine Fu, Srinath Sibi, Martin Fischer, and David Sirkin are with the Department of Mechanical Engineering and Department and Civil and Environmental Engineering at Stanford University, Stanford, CA 94043. (Phone: 415-793-5525; email: ernestinefu@stanford.edu).

David Hyde is with the Department of Mathematics at UCLA, Los Angeles, CA 90095. (email: dabh@math.ucla.edu).

Mishel Johns was formerly with the Department of Mechanical Engineering, Stanford University. He is now employed by Ford Motor Company USA. (email: mishel@stanford.edu).

research on types of autonomous system failures, and trust and risk in automation.

A. Levels of Automation and Need for Transitions of Control

Autonomous vehicles can be classified based on the amount of oversight required of the human driver when it comes to intervention and attentiveness. The taxonomy used most broadly in the field of automated driving research is the Society of Automotive Engineers (SAE) International's standard J3016, which defines six levels of automation, from Level 0 "no automation" to Level 5 "full automation" [3].

Level 3, "conditional automation," vehicles allow the system to monitor the driving environment and manage most aspects of driving. These vehicles are able to make decisions themselves; for example, they can sense a slower moving vehicle in front of them before deciding to switch lanes and overtake the other vehicle [4]. While mostly autonomous, such vehicles still face instances when they cannot handle a driving situation. The human driver is therefore responsible for remaining alert and being able to appropriately respond to a request to intervene [5].

B. Transitions of Control from Autonomous Driving Systems

As noted above, the role of the human driver is evolving as autonomous driving systems develop and advance. The human driver is no longer the sole operator of a vehicle; rather, the human driver and the autonomous system share that responsibility—they need to act as a team. This needed teamwork is most pronounced during transitions of control *from* automation *to* manual operation, when the autonomous system releases control of the vehicle, forcing the driver to quickly make sense of the driving situation and take appropriate action. There are several taxonomies and frameworks that categorize transitions of control [2, 6, 7, 8]. Within these frameworks, common factors that characterize transitions of control include the initiator and receiver of the transition (e.g., the driver, the autonomous system, or a remote actor), the optionality of the transition (e.g., whether a transition is merely suggested or whether it is forced), and the timing of the transition (e.g., whether the driver or system has any warning before a transition is initiated).

Transitions of control can occur in a takeover scenario [9] when the autonomous vehicle encounters a challenging situation and must defer to the human's expertise. It can also occur in order to increase driver attention or reduce other out-of-the-loop problems [6]. In near-term autonomous vehicle systems, which lack full autonomy and are not infallible, transitions of control will likely occur due to limitations or failures of the autonomous system. Such failures of the autonomous system can occur as either driver-initiated discretionary transitions or automation-initiated mandatory transitions.

Driver-initiated discretionary transitions occur when the human driver voluntarily seeks to take control of the vehicle, often because they believe they can navigate the driving environment better than the system [10]. Automation-initiated mandatory transitions occur when the system immediately hands control of the vehicle over to the human driver, often because the system has difficulty understanding the driving situation or because there is a system failure [11]. While both driver-initiated discretionary transitions and automation-initiated mandatory transitions occur in autonomous vehicles, we focus on the latter, which is an important area for study due to its safety-critical nature—human drivers are expected to quickly adapt to the driving task, gain situation awareness and successfully handle a potentially unanticipated driving situation on short notice [12, 13].

Prior studies on transitions of control have examined the amount of driver time required to regain situational awareness and safely respond to takeover from automation [14], the effects of distraction on driver performance in transitions of control [15], and the design of auditory, visual, and tactile takeover requests [16].

C. Types of Autonomous System Failures

Failures in autonomous systems can have varying levels of severity. A characterization of the severity of a system failure is provided by the controllability framework of Hobley et al., which defines severity as the "*qualitative assessment of the controllability of the safety of the situation (after a failure)*" [17, 18, 19]. Controllability may be quantized into levels such as Nuisance, Distracting, Debilitating, Difficult to Control, and Uncontrollable [18, 20]. For example, if a bright and reflective piece of debris enters the roadway, an autonomous vehicle's cameras could become unable to properly analyze and understand the scene, leading to a crash. In another case, vegetation could obscure a stop sign until a vehicle is near an intersection; this could result in an autonomous vehicle braking sharply, which could be a nuisance but would not necessarily lead to an accident [21]. In our study, we presented users with either mild failures in which a system behaves poorly and slightly deviates from the center of its driving lane, or severe failures in which the system's performance could lead to an accident such as a swerving into oncoming traffic.

In addition to severity, autonomous system failures are also characterized by the mechanism and extent to which the system alerts the user. For instance, one may imagine that a severe failure with no alert would create an overall worse situation than a severe failure that provided a blatant warning well in advance.

Loud failures, or structured transitions of control, occur when drivers are given a certain period of time to resume control while the automation system continues to drive. The autonomous system can prepare the driver for the transition of control since it knows it is about to fail or cannot handle a situation. Loud failures are more likely to occur for future systems, once fully autonomous vehicles are approved for widespread consumer use on roads [22].

On the other hand, silent failures, also called unstructured transitions, occur when no alert is presented to the human user before the transition of control. Silent failures occur because the vehicle is unaware that its autonomous system has failed, and thus represent a potentially much worse type of failure, as human drivers receive no stimuli from the system that alert them of the impending failure [9, 23, 24, 25]. Prior research has considered the effects of varying sensitivity levels in obstacle classification and interaction on driver performance during a silent system failure [26]. The dangers of silent failures have been heavily publicized in the media, e.g., in a

fatal accident in March 2018 in Arizona involving a pedestrian and an autonomous system being tested by Uber [27].

The failure events in our study were based on the premise that the vehicle encountered faded lane markings and an added set of pylons, resulting in the vehicle deviating from the center of its driving lane. We chose this design based on prior driving simulator studies that have also implemented this premise for system failures requiring takeover [15, 28].

D. Trust, Risk and Mental Models in Automation Systems

Trust, risk, and understanding of an autonomous system all contribute to a user's mental model of that system [29]. Trust has been characterized as a user's willingness to be "*vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor*," though there is inherent risk that the user may be failed by an imperfect actor or system [30]. Additionally, risk homeostasis theory [31, 32] posits that people change their behavior when their assessment of risk changes in order to maintain a desired level of risk. Moreover, a user's understanding and trust of a system may change as they experience interactions with the system in varying scenarios. This leads to the notion of a user's calibrated trust of a system, which refers to the adjusted level of trust a user arrives at after sufficient experience with the system [33].

When using any system, a person creates a mental representation, or mental model, of the system's capabilities and how to operate it. Such models inform human interactions with the system—whether digital or physical. A reliable, usable mental model of the autonomous system is essential for users to successfully learn and appropriately trust the system [34].

Similarly, the more similar a new event is to previous experiences a user has had with a system, the better equipped they will be to deal with this new interaction. These ideas have been investigated in a simulation study of aircraft pilot training, where more unexpected and varied training scenarios made pilots better equipped to handle unexpected situations in flight [35]. Another flight simulator study found that decreasing system reliability during training resulted in users more actively and successfully detecting automation failures [36]. Even thought experiments about novel scenarios, with no digital or physical simulation, are reported to improve pilots' perceived preparedness for unexpected interactions with systems [37]. Our present study relates to these works by exposing participants to different conditions with an autonomous vehicle and then evaluating their performance and behavior in a novel scenario.

As a proxy for understanding one dimension of users' mental models of autonomous systems, various instruments have been proposed to evaluate users' trust of those systems. Trust can be measured as self-report measures through a pre- and post-driving experience questionnaire (here, we used Jian, Bisantz, and Drury's instrument [38]), and also through ongoing measures of trust throughout the drive.

III. Study Goals and Methods

We hypothesize that if the autonomous vehicle system alerts a user (potential operator) of a failure and the potential need for a transition of control, that user will be less prepared to then navigate a similar failure in the future; in particular, if the system does not provide an alert (silent failure) compared to if the alert is issued again (loud failure). We also hypothesize that the severity of the first failure affects the user's performance in a second failure; the more severe the first failure, the less likely the user (now an operator) will be able to navigate the second failure.

To investigate these two hypotheses, we developed an experiment, based around a simulated driving course, that exposed participants to two failure events: the first failure varied in severity (with mild or extreme vehicle swerving to regain the original lane position), and the second required a transition of control that either included an alert (a loud failure, suggesting prior detection) or not (a silent failure, suggesting that the system was unaware of the failure). We assessed participants' levels of trust across the drive, and their performance during the second failure.

A. Participants

Participants between the age of 18 to 60 years old (M = 30.15 years, Mdn = 25.50 years, SD = 11.39 years) were recruited using online postings, emails and flyers. They were compensated for their time with an Amazon gift card. Participant driving experience ranged from 1 to 40 years (M = 13.46 years, Mdn = 7.25 years, SD = 11.54 years). Participants reported driving between 0 and 7 days per week (M = 3.44 days, Mdn = 3.55 days, SD = 2.54 days).

B. Driving Simulator

We used an immersive driving simulator consisting of a modified vehicle and visual display system (see Figure 1). A full Toyota Avalon compartment was adapted to provide study participants with a realistic interface, with the modifications stimulating on-road driving movement through haptic feedback in the steering wheel, seat and pedals. The visual display system included a 270-degree view screen that surrounds the car, with a rear projection that enabled the rear-view mirror to work. LCD panels were installed to act as side view mirrors. Speakers simulated road noise and provided audio alerts to study participants. Two GoPro cameras were installed inside the vehicle's cabin in order to monitor and record drivers' behavior during the study. We programmed the car's behavior and created the audio and visual components of the simulation with Realtime Technologies' SimCreator software.

To disengage automation in the simulated vehicle, participants were instructed to either initiate the brake pads, or turn the steering wheel a minimum of 15 degrees.

C. Experimental Design

TABLE I. Experimental Conditions

	Loud	**Silent**
Mild	N = 12	N = 12
Severe	N = 12	N = 12

We designed two types of failure in the first event (mild or severe), and then two types of failure in the second event (silent or loud). This is represented in a 2x2 matrix in Table 1. Participants (N=48) were randomly assigned to one of four

experimental conditions based on the combination of failures they encountered: Mild then Loud Failure, Severe then Loud Failure, Mild then Silent Failure, Severe then Silent Failure.

Both failure events occurred along similar right-hand curves in the road. In the first failure event, participants encountered faded lane markings and an added set of orange pylons (see Figure 2).

Figure 2. Front and aerial view of road with faded lane markings and orange pylons, as indicated by the white arrows

In the mild failure, unable to identify the lane's markings, the car wavered back and forth from the center of its driving lane, although mostly still keeping within its driving lane (see Figure 3). In the severe failure, the car exhibited the same swerving pattern; however, instead of remaining within its lane, it briefly swerved into the oncoming traffic lane before returning to its correct lane (see Figure 4). In both conditions, the system was able to successfully navigate and resume driving as another vehicle appeared in the opposite lane and triggered the system to resume driving in its intended lane. Both failures also included an alert for a potential need to transition control, although the system was then able to correct itself.

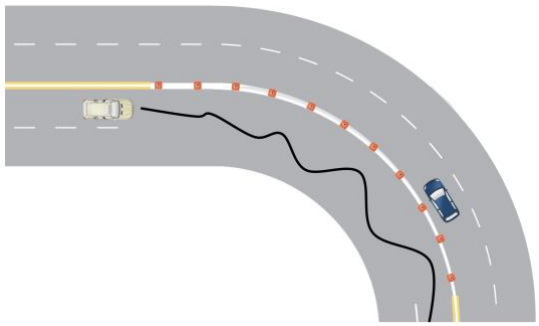

Figure 3. Illustration of mild failure (not actual path), where vehicle wavered in its driving lane

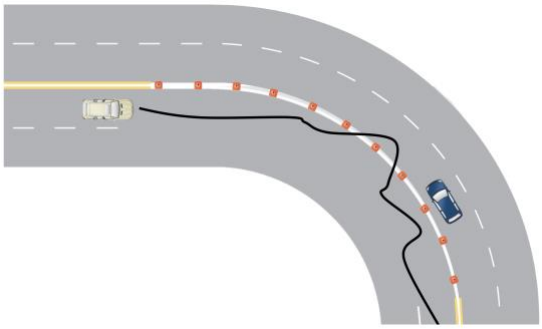

Figure 4. Illustration of severe failure (not actual path), where vehicle swerved into oncoming traffic's lane

In the second failure event, participants experienced a similar situation with faded lane markings, only this time, the vehicle exhibited either a silent or loud failure. In the silent failure, the transition of control was immediately passed on to the driver without any alert. This simulated the vehicle being unaware of the automation system failure. In the loud failure, the system issued an alert before the vehicle disengaged, providing the driver with a certain period of time to resume control while the automation system continued to drive. In both instances, the system was unable to correctly identify the faded lane markings and because there was no vehicle or object in the opposing lane, it incorrectly drove through the opposing traffic lane and subsequently into a grass area (see Figure 5). We measured automation as starting to fail once the vehicle began to enter the wrong driving lane.

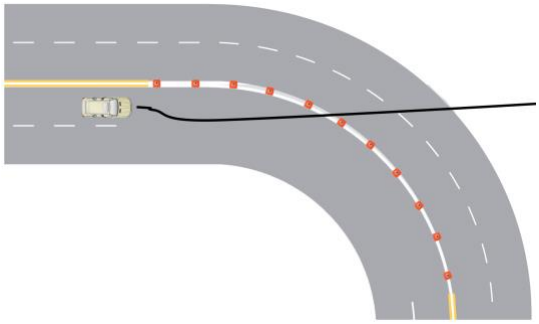

Figure 5. Illustration of second failure event (not actual path), where vehicle drove off the road and into the grass; system either detected the failure and issued an alert, or did not

D. Procedure

Participants sat in the driver's seat of a Level 3 autonomous vehicle (able to successfully navigate most roads and paths, but still requiring driver supervision). They were instructed that they would be traveling to the airport, and the course consisted of 5 driving segments. In some segments, participants drove the car manually, and in others, the automation system was in control.

In the first driving segment, participants manually drove the vehicle for 4 minutes in order to familiarize themselves with the simulation environment, including various road types such as curves and intersections, as well as signs such as speed limits and stop signs. Participants were also asked to practice engaging and disengaging automation. At the end of the segment, participants were asked to engage automation to allow for automated driving. During the rest of the drive, they were still able to disengage automation any time, particularly if they felt concerned about an imminent threat or accident.

In the second driving segment, participants monitored the vehicle as it successfully navigated various driving events for 10 minutes. These events included the system obeying speed limits, slowing down and stopping for pedestrians crossing the road, as well as stopping and proceeding through a stop sign and traffic lights (see Figure 6). Overall, the intention of this segment was to allow participants to observe the system correctly navigating events and not only encountering failures, thereby reflecting existing industry system design.

Figure 6. Front and aerial view of road as vehicle successfully slows down and brakes at a red traffic light

In the third driving segment, participants experienced either the mild or severe failure, before the system resumed navigation of the vehicle. Following the failure event, in the fourth driving segment, the vehicle once again successfully navigated various straight roads, curves, intersections, road signs and pedestrian crossings. Similar to the second segment, this segment lasted 10 minutes.

In the final driving segment, the vehicle either encountered a silent or loud failure. In both cases, participants were required to disengage automation and take over control of the vehicle. Participant performance in this critical event served as the crux of our data analysis.

Throughout the drive, there were 12 instances where participants were asked to evaluate and record their trust in the autonomous system using a dial we installed on the vehicle's dashboard, approximately at shoulder height, in between the instrument cluster and the center stack. When receiving the audio prompt "Trust Meter Indication Requested," participants were responsible for adjusting the dial right or left, and the changes were recorded. There were five options for participants to rate the intensity of their feeling of trust in the autonomous system: 1 for "I do not trust the autonomous system"; 2 for "I somewhat distrust the autonomous system"; 3 for "I have no opinion about the autonomous system"; 4 for "I somewhat trust the autonomous system"; and 5 for "I trust the autonomous system." As participants turned the dial, a horizontal bar indicator on the instrument cluster reflected their trust level selected from 1 to 5 (see Figure 7).

Figure 7. Trust meter indicator on instrument cluster, reflecting participants' self-reported trust in system

IV. Results

To understand participant performance and reliance on the system, we focused our analysis on user behavior in the critical event, which occurred at the second failure and required a transition of control. Successfully navigating the event meant that the user was able to correct the car's behavior as the failure started to occur. Participant performance varied based on the first type of failure they encountered. We also analyzed the vehicle's maximum lane offset and time to automation off, as well as participants' self-reported measures of trust in the autonomous system.

Data on participant performance in the critical event was measured in the 8-second window following the event trigger. This duration was chosen to account for how much time it took for the system to incorrectly cross into the oncoming traffic lane and drive into the grass, if the participant did not first intervene in the system failure. In the case where no participant reaction occurred, data was instead collected in a 30-second window. This duration was chosen to account for how much time the system continued its incorrect path before making a complete stop far into the grass.

Participant data was grouped into four categories, corresponding to each of the possible combinations of behavior during the first and second failure events.

A. Participant Performance in Critical Event

The number of participants who navigated the critical event successfully was highest in the Mild, Loud condition (7). Fewer participants navigated the critical event in the Severe, Loud condition (5). Mild, Silent condition had very few participants who successfully navigated the critical event (2), as was the case for the Severe, Silent condition (1) (see Figure 8).

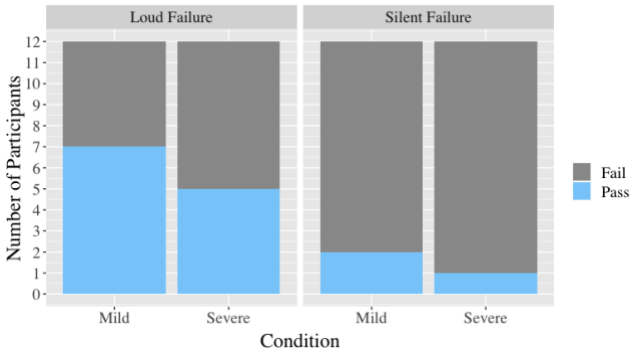

Figure 8. Performance of success or failure in navigating the critical event and its required transition of control

We considered the effect of the failure type on whether participants passed or failed the critical event. We performed binomial logistic regression to assess the correlation between the alert performance (loud or silent) and the participants' success rate, grouping together the mild and severe failure condition data. This model yielded a statistically significant difference between the presence or lack of an alert, with a p-value of 0.008.

B. Maximum Lane Offset in Critical Event

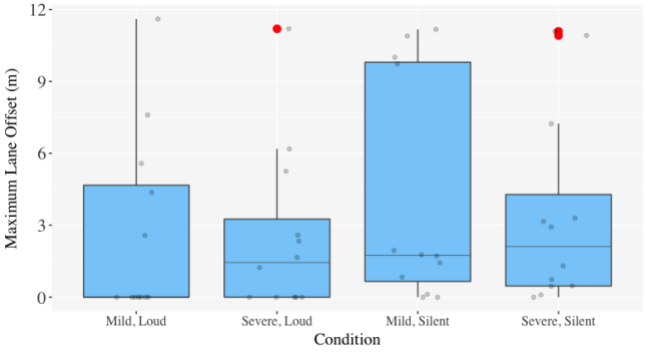

Figure 9. Offset of vehicle from its driving lane, representing one measure of participant's ability to navigate the required transition of control

We measured the absolute value of the maximum lane offset during the critical event, which signifies how far the vehicle strayed from its correct driving lane (see Figure 9). Lower offset values meant that participants successfully corrected the system's failed driving behavior. The Mild, Loud group had a mean and standard deviation of (2.64, 3.88) meters. For the same statistics, the Severe, Loud group scored

(2.54, 3.44), the Mild, Silent group reported (4.14, 4.73), and the Severe, Silent group yielded (3.47, 4.07). Two-way ANOVA testing was performed to assess the significance of the differences between these conditions. While the test did not suggest significant differences (Alert $F(1,44)=1.075$, $p=.30$; Severity $F(1,44)=.108$, $p=.74$; Alert:Severity $F(1,44)=.057$, $p=.81$), we conjecture that this is primarily limited by the small sample size of our study, and we highlight the noticeably smaller mean and standard deviation for the Severe, Loud group versus the Mild, Silent group.

C. Time to Automation Off in Critical Event

We measured the time it took users to disengage the vehicle's automation during the critical event (see Figure 10). Two participant data points were removed given their null values; we did not have their Time to Automation Off metric given these drivers' lack of takeover reaction. The mean and standard deviation of each group, in seconds, were found to be (4.41, 2.77) for the Mild, Loud group, (5.65, 4.66) for the Severe, Loud group, (7.87, 4.6) for the Mild, Silent group, and (5.62, 1.19) for the Severe, Silent group. p-values obtained from two-way ANOVA testing were again relatively weak (Alert $F(1,42)=3.673$, $p=.06$; Severity $F(1,42)=.032$, $p=.86$; Alert:Severity $F(1,42)=.722$, $p=.40$), though we again surmise that the difference in means between the conditions is large enough to warrant repeating the study with a larger participant pool. In particular, we observed a large mean in the Mild, Silent condition.

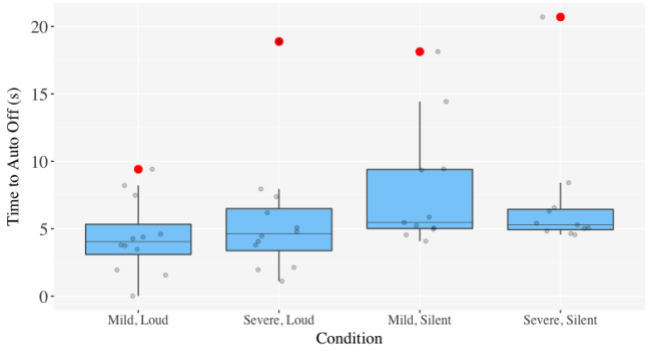

Figure 10. Amount of time participant took before disengaging automation during the critical event

D. Trust in System

We measured participants' trust in the autonomous system both during the drive, as well as before and after the drive. Results were averaged over the participants in each group, and the trust measurements are shown in Figure 11.

Trust measured at different events throughout driving course There are several noteworthy observations in this dataset on trust:

- All groups begin with approximately the same mean level of trust, suggesting a lack of initial bias in the composition of the groups.
- A sharp drop in mean trust is observed from before the first failure event (Location 6) and after (Location 7). The reduction in trust is comparable for all groups.
- Trust recovers between the two failure events but appears to reach a lower trending average than the average level of trust before the first failure event. This average decrease is suggestive of users incorporating the notion of calibrated trust into their mental models of the system.
- A similar drop in trust occurs from before the second failure event (Location 11) and after (Location 12). The reduction in trust is similar to the effect of the first failure event, even for the loud failure groups and despite users' knowledge from the first failure that the autonomous system can fail.

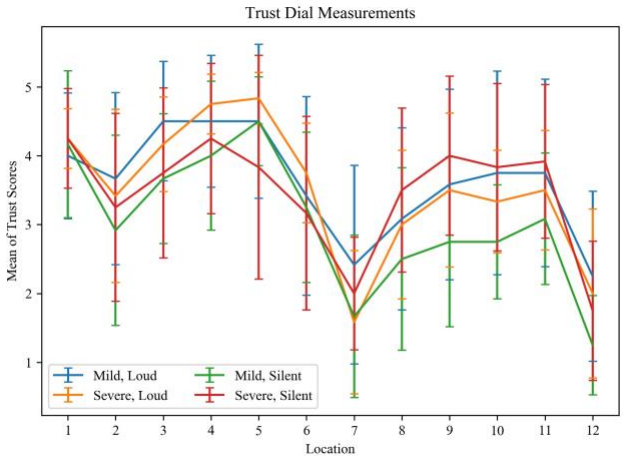

Figure 11. Trust measured at different events throughout driving course

We also measured participants' trust in the autonomous system both prior to and after the driving experience. Using an empirical trust evaluation methodology by Jian et al [38], participants completed a trust questionnaire at the beginning and end of the overall study. Repeated-measures ANOVA testing found insignificant changes in trust between the pre- and post-study assessments (see Figure 12).

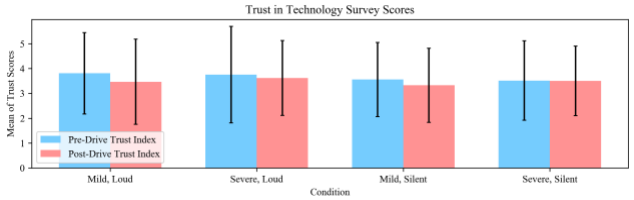

Figure 12. Trust measured pre- and post-drive

V. Discussion and Conclusions

When comparing the effects of a loud or silent failure in the critical event, we observed a significant relationship between having an audible alert and succeeding in the second failure event. We expect this may have occurred because the system issued an alert during the first failure event, causing users to believe that the system recognized the failure and would inform them of future failures. Therefore, when the second failure occurred, users may have complacently supposed the system would always issue an alert, especially if it were a critical failure that required a transition of control.

In regards to the participants who were unsuccessful in navigating the critical event despite receiving an audible alert, this may have occurred because, although the system encountered a failure and provided an alert for a possible transition of control, it also managed to ultimately resume control, thereby causing users to view the alert as unnecessary.

Such user behavior speaks to the risks of providing excessive alerts, particularly in scenarios where the car is able to ultimately resume driving control. Regardless, overall, engineers need to design how the system detects failures and alerts users—despite the vehicle's action or inaction in response to those failure.

The severity of failure that participants received in the first failure event did not affect their performance in the critical event. This was surprising to us, because we hypothesized that a more severe failure would have significant effects on increasing the user's vigilance, and thereby performance, in the critical event. We assumed that using a vehicle that crossed into the opposite driving lane and encountered a near-collision with an oncoming vehicle would be a startling scenario and result in increased driver caution, compared to when the vehicle simply wavered drastically in its own lane. The results suggest that when designing a system, engineers need to consider that any failure, despite the severity, affects driver performance in a subsequent failure event.

User trust in the system varied over time and between events. We observed that while users started with the same level of trust in the autonomous system, those in all condition groups experienced a sharp reduction in trust after the first failure event, despite the severity of the failure. Consistent with our analysis of participant performance, the severity of the first failure did not affect subsequent trust and performance. While trust levels after the first failure event never returned to its initial higher levels, we noticed that participants gradually built or rebuilt trust in the system as the vehicle successfully navigated events such as traffic signs and obstacles in the road.

In fact, during the second failure event, we were surprised that several participants were unable to navigate the critical event's required transition of control because they simply did not take action—their maximum lane offset numbers were high because they deviated from the vehicle's correct path, as were the time-to-automation-off numbers because they waited a significant amount of time before manually taking control of the driving. The time-to-automation-off data, coupled with the video data we collected, indicate that some participants trusted the system enough that even after the vehicle plowed through the opposing driving lane and through the grass, they waited up to 30 seconds before disengaging automation and then manually operating the vehicle back into the correct driving lane. The significant amount of time that these participants took to manually correct the vehicle and complete the transition of control is a surprising indicator that they expected the system to be able to self-correct.

In conclusion, our study provides a preliminary analysis of the effect of failures and transitions of control on driver trust and performance. Revisiting our two hypotheses, the findings support our first expectation that vehicle system failure alerts can cause drivers to be less prepared to handle future silent failures relative to loud failures (due to a decline in vigilance and over-reliance on the warning system). The findings do not support our second expectation that a more severe first failure would lead to decreased performance in a second failure.

VI. Limitations and Future Work

Simulator studies are limited by the skewed perception of danger and urgency, and one shortcoming is the potential difference in driver trust and performance when compared to that during on-road studies. We hope to modify the study design in the future to allow for experiments on physical roads, while maintaining the ethical integrity required for understanding failures and autonomous vehicle systems.

While we observed that participant trust levels did not return to their original levels during the second trust-building course segment, a longer study with more tests of successful system performance may show more nuanced changes in trust and provide designers with suggestions on how quickly and completely trust can be rebuilt.

In future studies, we are also interested in pursuing more nuanced gradations and variants of alerts. In this work, we considered only audio alerts, however, different modalities can be used simultaneously or concurrently, and using multiple alerting modalities may be more effective than just one [39]. Most warnings today are communicated through auditory or visual modalities. However, the driving environment is important when considering which warning system should be used. If visibility on the windshield where the notification will be displayed is limited, an auditory warning could be preferred. And if there is background noise, a visual warning including words or symbols with varying visual design such as color could be preferred [40]. Tactile warning systems include varying pressure, texture and temperature. Vibrotactile presentations, which are coded vibrations, are frequently used in warnings today and can be particularly effective for announcing binary events, such as a state change in automated systems. However, their effectiveness may be compromised during low-frequency whole body vibration, such as when the system is driving through rough terrain [41]. Variants of alerts are possible, and we are interested in pursuing these more advanced and subtle distinctions in future work.

Acknowledgment

This study was conducted under Stanford University IRB Protocol 30016. We thank our fellow researchers at Stanford's Center for Design Research for their advice in the development of the study.

References

[1] J. M. Anderson, N. Kalra, K. D. Stanley, P. Sorensen, C. Samaras, and T. A. Oluwatola, *Autonomous Vehicle Technology: A Guide for Policymakers*. Santa Monica, CA: RAND Corporation, 2016.

[2] Z. Lu, and J. C. F. De Winter, "A review and framework of control authority transitions in automated driving," *Procedia Manufacturing*, vol. 3, pp. 2510–2517, 2015.

[3] SAE. Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems. SAE Standard J3016_201401, Jan. 2014.

[4] V. Milanés, D. F. Llorca, J. Villagrá, et al., "Intelligent automatic overtaking system using vision for vehicle detection," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3362–3373, 2012.

[5] "Trucks: Hitachi, Wenco Complete Successful Autonomous Haulage Tests," *Canadian Mining Journal*, Dec. 2017.

[6] Z. Lu, R. Happee, C. D. D. Cabrall, M. Kyriakidis, and J. C. F. De Winter, "Human factors of transitions in automated driving: A general framework and literature survey," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 43, pp. 183–198, Nov. 2016.

[7] R. Mccall, F. Mcgee, A. Mirnig, et al., “A taxonomy of autonomous vehicle handover situations,” Transportation Research Part A: Policy and Practice, vol. 124, pp. 507–522, June 2019.

[8] A. G. Mirnig, M. Gärtner, A. Laminger, et al., “Control transition interfaces in semiautonomous vehicles: A categorization framework and literature analysis,” in *Proc. 9th Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI’17)*, 2017, pp. 209-220.

[9] C. Gold, M. Körber, D. Lechner, and K. Bengler, “Taking over control from highly automated vehicles in complex traffic situations: The role of traffic density,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 58, no. 4, pp. 642–652, 2016.

[10] M. A. Goodrich and E. R. Boer, “Multiple mental models, automation strategies, and intelligent vehicle systems,” in *Proc. 199 IEEE/IEEJ/JSAI Int. Conf. on Intelligent Transportation Systems (Cat. No.99TH8383)*, pp. 859–864.

[11] M. Saffarian, J. C. F. De Winter, and R. Happee, “Automated driving: Human-factors issues and design solutions,” in *Proc. Human Factors and Ergonomics Society Annual Meeting*, vol. 56, no. 1, pp. 2296–2300, 2012.

[12] M. R. Endsley, “Measurement of situation awareness in dynamic systems,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, no. 1, pp. 65–84, Mar. 1995.

[13] M. R. Endsley, “Toward a theory of situation awareness in dynamic systems,” in *Situational Awareness*, E. Salas, Ed. London, UK: Routledge, 2017, pp. 9–42.

[14] H. Clark, A. C. Mclaughlin, and J. Feng, “Situational awareness and time to takeover: Exploring an alternative method to measure engagement with high-level automation,” in *Proc. Human Factors and Ergonomics Society Annual Meeting*, vol. 61, no. 1, 2017, pp. 1452–1456.

[15] B. Mok, M. Johns, D. Miller, and W. Ju, “Tunneled in: Drivers with active secondary tasks need more time to transition from automation,” in *Proc. 2017 CHI Conf. on Human Factors in Computing Systems (CHI ‘17)*.

[16] S. Petermeijer, F. Doubek, and J. De Winter, “Driver response times to auditory, visual, and tactile take-over requests: A simulator study with 101 participants,” in *2017 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*.

[17] K. M. Hobley, et al., Framework for Prospective System Safety Analysis Vol. 1 – Preliminary Safety Analysis. *Deliverable 9a*, V2058 PASSPORT project of the Advanced Transport Telematics (ATT/DRIVE II) sector of the Telematics Applications Programme, Third Framework Programme, 1995, pp. 1991-1994.

[18] P. H. Jesty, and K. M. Hobley, “Integrity levels and their application to road transport systems,” in *Proc. 15th Int. Conf. on Computer Safety, Reliability and Security (SafeComp ‘96)*, E. Schoitsch, Ed. London, UK: Springer, 1997, pp. 365–374.

[19] P. H. Jesty, K. M. Hobley, R. Evans, and I. Kendall, “Safety analysis of vehicle-based systems,” in *Proc. 8th Safety-critical Systems Symposium*, 2000, pp. 90-110.

[20] MISRA, *Development Guidelines for Vehicle Based Software*. Nuneaton, UK: MIRA, 1994.

[21] M. Bertozzi, A. Broggi, and A. Fascioli, “Vision-based intelligent vehicles: State of the art and perspectives,” *Robotics and Autonomous Systems*, vol. 32, no. 1, pp. 1–16, 2000.

[22] C. Gold, D. Damböck, L. Lorenz, and K. Bengler,”‘Take over!’ How long does it take to get the driver back into the loop?” in *Proc. Human Factors and Ergonomics Society Annual Meeting*, vol. 57, no. 1, pp. 1938–1942, Sept. 2013.

[23] A. Eriksson and N. A. Stanton, “Takeover time in highly automated vehicles: Noncritical transitions to and from manual control,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 59, no. 4, pp. 689–705, 2017.

[24] C. Gold, M. Körber, C. Hohenberger, D. Lechner, and K. Bengler, “Trust in automation – Before and after the experience of take-over scenarios in a highly automated vehicle,” *Procedia Manufacturing*, vol. 3, pp. 3025–3032, 2015.

[25] V. Melcher, S. Rauh, F. Diederichs, H. Widlroither, and W. Bauer, “Take-over requests for automated driving, *Procedia Manufacturing*, vol. 3, pp. 2867–2873, 2015.

[26] E. Fu, S. Sibi, D. Miller, M. Johns, B. Mok, M. Fischer, and D. Sirkin, “The car that cried wolf: Driver responses to missing, perfectly performing, and oversensitive collision avoidance systems,” in *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1830-1836.

[27] A. Marshall, “Uber’s Self-Driving Car Just Killed Somebody. Now What?” *Wired*, 2018.

[28] B. Mok, M. Johns, S. Yang, and W. Ju, “Actions speak louder: Effects of a transforming steering wheel on post-transition driver performance,” in *2017 IEEE 20th Int. Conf. on Intelligent Transportation Systems (ITSC)*.

[29] N. Staggers, and A. F. Norcio, “Mental models: Concepts for human-computer interaction research,” *International Journal of Man-Machine Studies*, vol. 38, no. 4, pp. 587–605, 1993.

[30] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An integrative model of organizational trust,” *The Academy of Management Review*, vol. 20, no. 3, p. 709, 1995.

[31] N. J. Ward, “Automation of task processes: An example of intelligent transportation systems,” *Human Factors and Ergonomics in Manufacturing*, vol. 10, no. 4, pp. 395–408, Autumn 2000.

[32] G. J. S. Wilde, “Risk homeostasis theory and traffic accidents: Propositions, deductions and discussion of dissension in recent reactions,” *Ergonomics*, vol. 31, no. 4, pp. 441–468, 1988.

[33] J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance, “*Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.

[34] D. E. Kieras and S. Bovair, “The role of a mental model in learning to operate a device,” *Cognitive Science*, vol. 8, no. 3, pp. 255–273, July 1984.

[35] A. Landman, P. Van Oorschot, M. M. (René) Van Paassen, E. L. Groen, A. W. Bronkhorst, and M. Mulder, “Training pilots for unexpected events: A simulator study on the advantage of unpredictable and variable scenarios,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 60, no. 6, pp. 793–805, June 2018.

[36] R. Parasuraman, R. Molloy, and I. L. Singh, “Performance consequences of automation-induced complacency.” *The International Journal of Aviation Psychology*, vol. 3, no. 1, pp. 1–23, 2009.

[37] W. L. Martin, P. S. Murray, and P. R. Bates, “What would you do if....? Improving pilot performance during unexpected events through in-flight scenario discussions,” *Aeronautica*, vol. 1, pp. 8-22, 2011.

[38] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, “Foundations for an empirically determined scale of trust in automated systems,” *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[39] M. S. Wogalter, V. C. Conzola, and T. L. Smith-Jackson, “Research-based guidelines for warning design and evaluation,” *Applied Ergonomics*, vol. 33, no. 3, pp. 219–230, May 2002.

[40] M. S. Wogalter, V. C. Conzola, and T. L. Smith-Jackson, “Research-based guidelines for warning design and evaluation,” *Applied Ergonomics*, vol. 33, no. 3, pp. 219–230, May 2002.

[41] A. E. Sklar and N. B. Sarter, “Good vibrations: Tactile feedback in support of attention allocation and human-automation coordination in event-driven domains,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 41, no. 4, pp. 543–552, Dec, 1999.