Evaluating the Reliability of Tesla Model 3 Driver Assist Functions

October 1st, 2020

Benjamin Bauchwitz
M.L. Cummings
Duke University

THE UNIVERSITY OF TENNESSEE KNOXVILLE   Berkeley UNIVERSITY OF CALIFORNIA   FAU FLORIDA ATLANTIC UNIVERSITY   Duke UNIVERSITY   THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

www.roadsafety.unc.edu

U.S. DOT Disclaimer

Acknowledgement of Sponsorship

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. CSCRS-R{X}	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle: Evaluating Reliability of Tesla Model 3 Driver Assist Functions		**5. Report Date** October 1st, 2020	
		6. Performing Organization Code	
7. Author(s) Benjamin Bauchwitz & M.L Cummings		**8. Performing Organization Report No.**	
9. Performing Organization Name and Address Humans and Autonomy Laboratory 130 North Building Duke University Durham, NC 27708		**10. Work Unit No.**	
		11. Contract or Grant No. Collaborative Sciences Center for Road Safety (Grant #: 69A3551747113)	
12. Sponsoring Agency Name and Address Collaborative Sciences Center for Road Safety 730 Martin Luther King Jr. Blvd., Suite 300 Chapel Hill, NC 27599		**13. Type of Report and Period Covered** Annual Report (SEP 2019- SEP 2020)	
		14. Sponsoring Agency Code	

15. Supplementary Notes
Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract
The number of vehicles with advanced driver-assist systems (ADAS) that can simultaneously perform automated steering and acceleration is increasing, with the expectation that these capabilities will be available in low-cost models as well as luxury cars. These systems require human drivers to be alert and available at all times in case they need to take over. Thus, the interface between the human and the machine is of critical importance in ADAS-equipped vehicles. Formal testing and guidance for development of such ADAS systems is limited. The goal of this effort was to assess between- and within-ADAS-equipped vehicle variation in four scenarios involving the interface between a human driver and an ADAS system. These scenarios were: (1) Assessing driver-monitoring system performance during automated highway driving; (2) Alerting a distracted driver of unexpected road patterns during automated driving; (3) Assisting a distracted driver in response to an inadvertent lane departure; and (4) Initiating driver handover to a distracted driver when the vehicle can no longer confidently operate. Given that Teslas can face a range of potentially demanding environments, a Tesla Model 3S was the test platform. The results suggest that the performance of the computer vision systems was extremely variable, and this variation was likely responsible for some, but not all, of the delays in alerting a driver whose hands were not on the steering wheel. Performance was not consistent for a single car, in that a car would perform the best in the most challenging driving scenarios (navigating extreme curves while the driver ignored takeover requests), but performed the worst on seemingly simpler scenarios like detecting a lane departure. These results suggest that practitioners need to develop a richer set of tests that capture within- and between-car variability, and they also indicate that software upgrades through over-the-air updates may induce latent issues that lead to safety issues.

17. Key Words Autonomous vehicle, highly automated vehicle, certification, testing, Tesla, variability		18. Distribution Statement		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified		**21. No. of Pages** 42	**22. Price**

Form DOT F 1700.7 (8-72)   Reproduction of completed page authorized

Contents

Introduction

At least 10% of new cars sold in the United States include features enabling partial automation (1), defined as *Level II Autonomy* in SAE standard J3016 (2). Such vehicles include an advanced driver-assist system (ADAS) which can simultaneously perform automated steering and acceleration. However, these systems require human drivers to be alert and available at all times in case they need to take over. The interface between the human and the machine is of critical importance in ADAS-equipped vehicles as changes in human attention and behavior with high levels of automation make the handover regime particularly dangerous (3–5). Therefore, Level II autonomous vehicles typically employ some type of driver monitoring system for assessing and alerting drivers during these types of events. Currently, there are no US regulations addressing how driver monitoring should be implemented or the performance standards such systems should meet.

Formal testing of ADAS systems in general is limited. The National Highway Traffic Safety Administration's (NHTSA) New Car Assessment Program (NCAP) does not cover driver assist features, limiting its assessment to collision and rollover survival (6). The Insurance Institute for Highway Safety (IIHS) covers some ADAS features such as pedestrian detection and automated emergency braking (AEB) but does not address driver monitoring (7). The European NCAP has similar scope but has announced that it will begin assessing driver monitoring in the 2022 revision of its protocols (8). The Korean Ministry of Land, Infrastructure, and Transport (MOLIT) is the only major authority regulating driver monitoring and as of June 2020, it provides specific guidance regarding how the driver monitoring system should handle engagement and disengagement of automated features (9).

Even in the few cases where guidance is provided on how driver monitoring should be evaluated, certain aspects of testing are still highly ambiguous. There is little to no specification as to how individual differences in vehicles or variations in operating environment should be considered in the test procedure. None of the NHTSA, IIHS, NCAP, or MOLIT test protocols address how vehicles should be sampled to ensure the test results are robust to differences in vehicle trim, configuration, or wear-and-tear. Likewise, these test processes either take place in sterile laboratory settings or make assumptions about the generalizability of the results to different atmospheric conditions or road environments. Such tests may underestimate the effects that subtle changes in atmospheric conditions have on vehicle performance.

The goal of this research was to assess between- and within-ADAS-equipped vehicle variation in four key scenarios involving the interface between a human driver and an ADAS system: (1) assessing driver-monitoring system performance during automated highway driving; (2) alerting a distracted driver of unexpected road patterns during automated driving; (3) assisting a distracted driver in response to an inadvertent lane departure; and (4) initiating driver handover to a distracted driver when the vehicle can no longer confidently operate. In addition, where variation was present, the goal was to evaluate the impact of the under-studied environmental parameters. Given that Teslas have ADAS systems that can be used on interstates, divided highways, and urban and rural roads and thus can face a range of potentially demanding environments, a Tesla Model 3S was used as the test platform.

Background

To develop a testing protocol, a model is needed to better understand the major elements of the human-vehicle system and how they interact. To this end, the model in Figure *1* includes a driver model, a vehicle model, and a model of the environment which includes not only atmospheric issues such as brightness and weather but also the road environment including, for example, lane markings, possible obstacles, and road characteristics. These model elements are further described in the following sections.
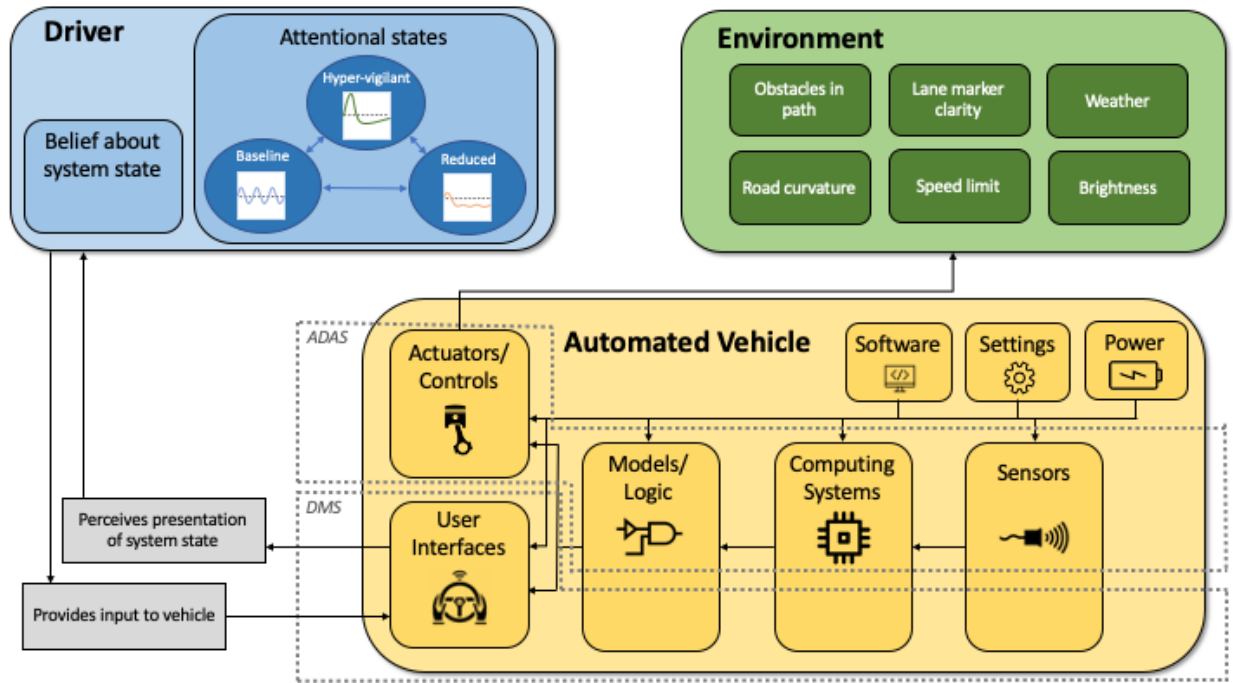
Figure 1: Influence model of autonomous system configuration, driver beliefs about the system and environment, and driver attention level

In this model, operating an ADAS-equipped vehicle driver attention includes three states: baseline, reduced, and hypervigilant (Figure *2*). Baseline attention is characterized by relatively small fluctuations within a modest range around the baseline average. For example, drivers pay attention to the road but occasionally change the radio station or read a billboard on the highway. Reduced attention is characterized by a prolonged period of consistently below-baseline attention levels. This is the state that subsumes the large volume of distracted driving research.

Hypervigilance is characterized by a brief surge in attention far above baseline levels, followed by a regression back to baseline and eventual dip below baseline. Such a state occurs, for example, when a distracted person almost runs off the road and then pays close attention for some period of time. For this effort, a driver is always assumed to be in the reduced attentional state, but more research is needed in how cycles between reduced and hypervigilant attention affect longer-term focused attention.

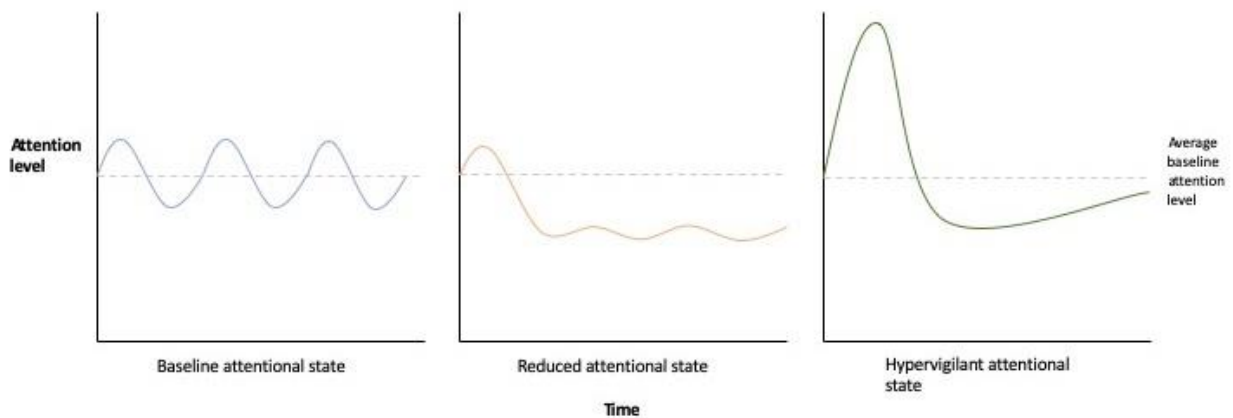

Figure 2: Model of attention levels over time: baseline, reduced, and hypervigilant

In Figure *1*, the vehicle model includes the software, the computational systems, sensors, and actuators. This study focuses on the automated driving system composed of the key functional components. Tesla Model 3S (and indeed all Teslas) have the option of including Autopilot which is made up of several

components including Autosteer, Traffic-Aware Cruise Control, Navigate on Autopilot, Self-park, a torque-monitoring driver detection system, forward, side, and rear facing visible light cameras, sonar, radar, ultrasonic sensors, Nvidia's processing computer, the neural network system for signal processing and decision-making, the console, and the auditory alert system.

The automated driving system also has a "presentation" which is the way it appears to the driver of the car. The presentation is a direct consequence of the system state which is a function of the particular configuration of the components, i.e. which state each of the components is in, as well as the inputs that each component receives. The presentation to the driver is thus a function of the set of its components which in turn are each a function of the set of their inputs. For instance, a functioning Tesla driver detection system has a different presentation when a driver's hands are detected on the wheel versus when they are not, and this detected state then can have consequences in regard to the vehicle performance. For example, if a driver's hands are not detected on the wheel after three successive attempts to alert the driver to take over, the car will come to a stop.

The presentation space is thus the set of all possible presentations that can arise from the various configurations of the system components. The driver has a belief about the underlying configuration of the system components and the inputs driving their behavior, and this belief is based on the presentation. However, the relationship is ambiguous as certain presentations may convey their underlying component configurations less saliently than others. Mode confusion occurs when the driver's belief about the configuration of the system is inconsistent with the true configuration (10). Tesla is no stranger to such problems and has had significant issues with drivers misunderstanding the release of their "Full Self Driving" option which is only an enhanced version of Autopilot (11).

Drivers who think their cars are more capable than they are may be more susceptible to increased states of distractions and thus at higher risk of crashes. This susceptibility is highlighted by several fatal crashes of distracted Tesla drivers with Autopilot engaged (12,13). Past research has characterized a variety of ways in which autonomous systems can influence an operator's attentional levels. When an operator's task load is low while supervising an autonomous system, he can experience increased boredom which results in reduced alertness and vigilance (ability to respond to external stimuli) (14). One study reported that once attention levels are reduced, they do not necessarily rebound to baseline levels once an autonomous system disengages and attention may stay at reduced levels (4). Furthermore, when an operator has initially sustained very high attention, she may experience psychological fatigue, which is a feeling of prolonged exhaustion and reduced capacity for work. This fatigue is not relieved by rest and, so under these circumstances, relying on the autonomous system for primary control may not return attention levels to baseline (14).

Given that ADAS-equipped cars have led to increased reduced attentional states, characterizing how reliable the driver monitoring and alerting system was between different Tesla Model 3S, as well as within these vehicles, is critical for understanding possible variation in such systems. In addition, this effort sought to determine if there are additional vehicle and environmental factors as illustrated in Figure 1 that could have an impact on such outcomes. The next section outlines the experimental methods used to address four basic areas of inquiry: (1) assessing driver-monitoring system performance during automated highway driving; (2) alerting a distracted driver of unexpected road patterns during automated driving; (3) assisting a distracted driver in response to an inadvertent lane departure; and (4) initiating driver handover to a distracted driver when the vehicle can no longer confidently operate. In addition, where variation was present, other potentially important test parameters were identified.

Methods

Three 2018 Tesla Model 3Ses from the Triangle metropolitan area of North Carolina were randomly selected for study using a car sharing service over a period of two weeks during March 2020. All tests were conducted during daylight, between 12:00pm and 5:00pm, under similar environmental conditions (Appendix A: Environmental Conditions by Test Day). The software versions and Autopilot settings used are detailed in Appendix B: Vehicle Software Configurations by Test Day. The same person drove the vehicle for all tests.

Prior to each trial, the vehicle was placed in park with the driver exiting and using the key card to lock and deactivate the vehicle before entering the car to begin a test.

Tests were either performed on a public highway or at the North Carolina Center for Automotive Research (NCCAR), a closed test track facility. For each vehicle, the highway tests were performed on one day while the track tests were performed on a second day with the order of these two test days randomized for each car. The NCCAR test track is a two-mile long, 40-foot-wide paved loop with a mix of straightaways and curves of a widely varying range of angles. Some tests involved the use of painted lane markings, which included lanes 13 feet wide marked with 10 foot long by 6 inch wide white lane markings and 30 feet of longitudinal distance between each marking (15).

Four different experiments were performed, one assessing each of four key driver monitor behaviors: (1) assessing driver-monitoring system performance during automated highway driving, labeled Highway (HW); (2) alerting a distracted driver to unexpected road patterns during automated driving, such as construction lane shifts, during automated driving, labeled Lane Shift (LS); (3) assisting a distracted driver in response to an inadvertent lane departure, labeled Lane Departure (LD); and (4) initiating driver handover to a distracted driver when the vehicle can no longer confidently operate, labeled CRV since this test was performed on a S-curve road. The LS, LD, and CRV tests were all conducted at the NCCAR track and the order trials was randomized for each vehicle. These tests are described in more detail below.

Highway Test

The goal of the first experiment was to determine if a significant within- and between-vehicle difference existed in the type and timing of feedback presented to a driver when the vehicle sensed driver inattention during highway driving. Per the stated design specifications, the vehicle should request that the driver put their hands on the wheel approximately once every 25 seconds, as described in official documentation (12).

The highway test was conducted on two 5.2 mile sections of Interstate 540 between Cary and Apex, NC. The two routes were mirror images of one another, with a posted speed limit of 70 mph. All highway tests occurring before 4:00pm to minimize the influence of rush hour traffic. Each vehicle experienced 10 repetitions of each test, 5 alternating in each direction.

The southern route began at the entrance to the westward I-540 lanes at the junction with state route 55 and concluded at exit 59. The northern route began at the entrance to the eastward I-540 lanes beginning at the junction with state route 64 and concluded at exit 66 (Figure 3). In the southbound route, upon passing under the McCrimmon Parkway underpass after entering the highway, the car was placed in Autopilot at 70 mph and data collection was initiated.

Because this section of road has between 3 and 5 lanes at various points, the car was driven in Autopilot in the third lane from the left, which allowed it to be driven without the need to change lanes.
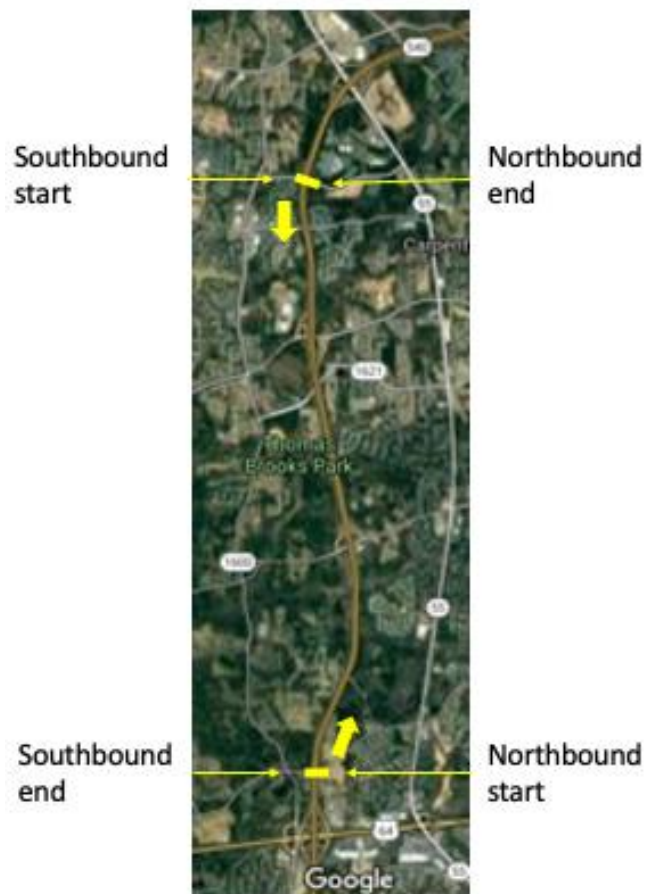

Figure 3: Route used for highway testing, which included two equivalent 5.2 mile sections of I-540: The southbound route between exits 66 and 59, and the corresponding northbound route.

Once the vehicle was in Autopilot, the driver did not interact with the controls other than to provide the minimum steering wheel input necessary to respond to any alerts for the driver to apply force to the steering wheel.

The alert consists of a message on the car's 15 inch, 1920 x 1080 pixel display mounted on the center of the dash that says "Apply slight turning force to steering wheel" and is accompanied by a quick pair of beeps. Tesla vehicles recognize a driver has taken control through a torque monitoring system on the steering wheel that measures how forcefully the steering wheel has been rotated in an attempt to infer whether the driver has deliberately manipulated it.

The required force was applied immediately upon presentation of the alert and was continued until the alert disappeared. Then the driver took his hands off the steering wheel again and waited until the next alert, with this sequence continuing for each 5.2 mi section. The test was concluded after 5.2 mi (passing under the Jenks Road underpass near the conclusion of the route), at which time Autopilot was disengaged. The same protocol was used for the northbound route, with data collection and autopilot initiated after passing under the Jenks Road underpass and concluded after passing under the McCrimmon Parkway underpass (Figure 3).

Given the posted speed limit, the car was expected to take approximately 4.5 minutes to complete the route. With the permitted hands-free interval of about half a minute, up to 8 cycles of hands-free driving followed by a vehicle request for steering input could have occurred in each trial. Because the driver only responded to alerts requesting steering input, the car was not maneuvered around other traffic. In a few instances, the Tesla slowed behind other vehicles traveling at slower speeds. In these cases, the Tesla was allowed to travel at sub-70 mph speeds until the other vehicle changed lanes. The driver only took control in response to safety issues including (1) changing lanes due to a police lane closure, (2) steering to avoid workers on the roadway, and (3) taking over to mitigate unsafe behavior by the vehicle.

Lane Shift Test

The goal of the second experiment was to determine if a significant within- and between-vehicle difference existed in a vehicle's ability to avoid obstacles while encountering an unexpected road pattern with a distracted driver. Also investigated was the type and timing of feedback presented to the driver upon encountering this anomaly. Given that the Tesla Autopilot is not designed to be operated in construction sites or other areas with similarly confusing road markings or obstacles, the hypothesis was that all vehicles would present a driver takeover alert immediately upon encountering the lane shift and would steer to avoid obstacles.

For this test, the vehicle began at the position marked 'start' in Figure 4. The car was driven manually along a 515 foot curved section of track and accelerated to 25 mph. At the conclusion of the curve, there was a 330 foot section of straight track marked with three highway-style lanes, with the car aligned with the rightmost lane. Immediately upon passing a cone at the beginning of this straightaway, the car was placed in Autopilot with the speed fixed at 25 mph. After 200 feet, a solid yellow line marked a lane shift in which the right-hand lane merged into the central lane. The original dashed white lines were also visible. In the final 40 foot section of the straightaway, an angled barricade of 7 orange traffic cones blocked the rightmost lane. If the car failed to follow the lane shift, it would collide with the cones, although the driver, only simulating a distracted driver, took evasive steering if a collision was imminent.
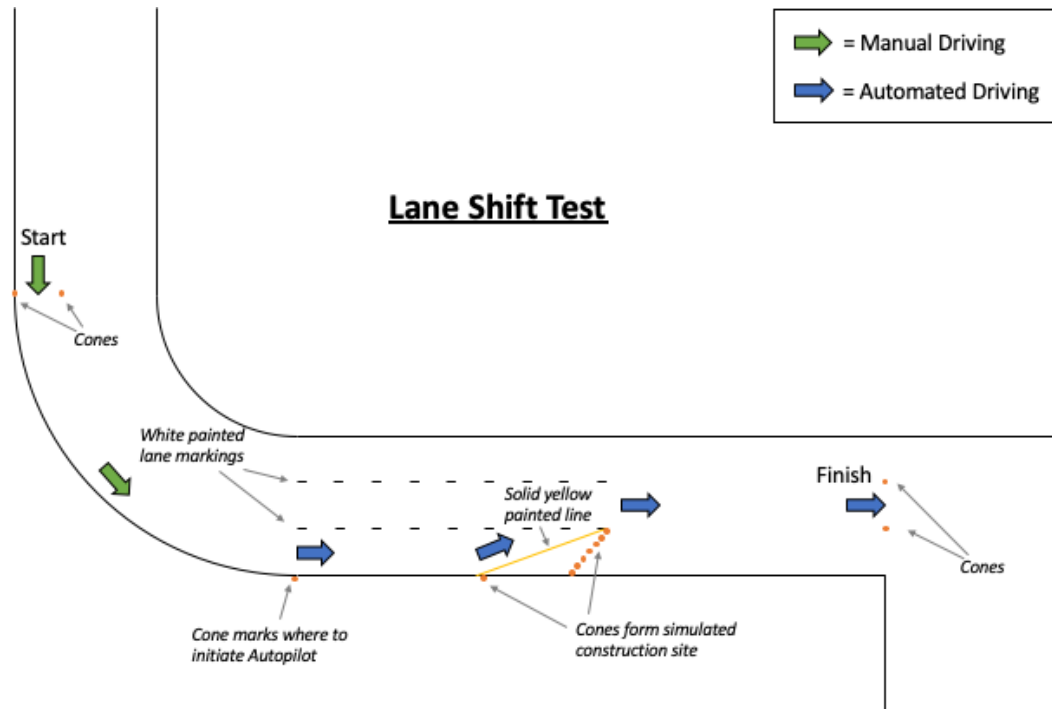
Figure 4: Setup of test track for Lane Shift (LS) tests. The cars started between two traffic cones at the position marked 'start'. There was a 330-foot section of straightaway painted to depict three highway style lanes. The last 130-foot segment included a solid yellow painted line indicating a lane shift that forced the outermost lane to merge into the central lane. The last 40-foot segment included an angled barricade of 7 traffic cones blocking further travel in the outermost lane.

Lane Departure Test

The goal of the third experiment was to determine if a significant within- and between-vehicle difference existed in the type and timing of feedback provided to a distracted driver when the vehicle drifted out of its lane at a very low level of automation; in this case, automated cruise control. The vehicles were configured to provide emergency lane departure assistance, meaning the car should have provide evasive automated steering to prevent the vehicle from exiting the lane. Therefore, the hypothesis was that all vehicles would provide alerting and emergency assistive steering as the cars drifted out of their lanes relatively close to the road's edge.

For this test, each vehicle began between two traffic cones at the position marked 'start' in Figure 5. Starting from an inactive, parked state, the vehicle was driven toward the painted lane lines in the inner most lane. After accelerating to 35 miles per hour, the car was immediately placed in Adaptive Cruise Control to fix the speed, and Autopilot was not initiated. Upon passing the cone marking the beginning of the painted section of track, the driver "nudged" the steering wheel 3-5° so that the front of the car was aimed just to the left of a second cone on the right outer edge of the track 130 feet away. The car was allowed to move in that direction with no steering input until either the car left the lane or the lane-keep assist feature activated, steering the car back into the lane. The trial was concluded as soon as the vehicle passed the final set of white lane markings.
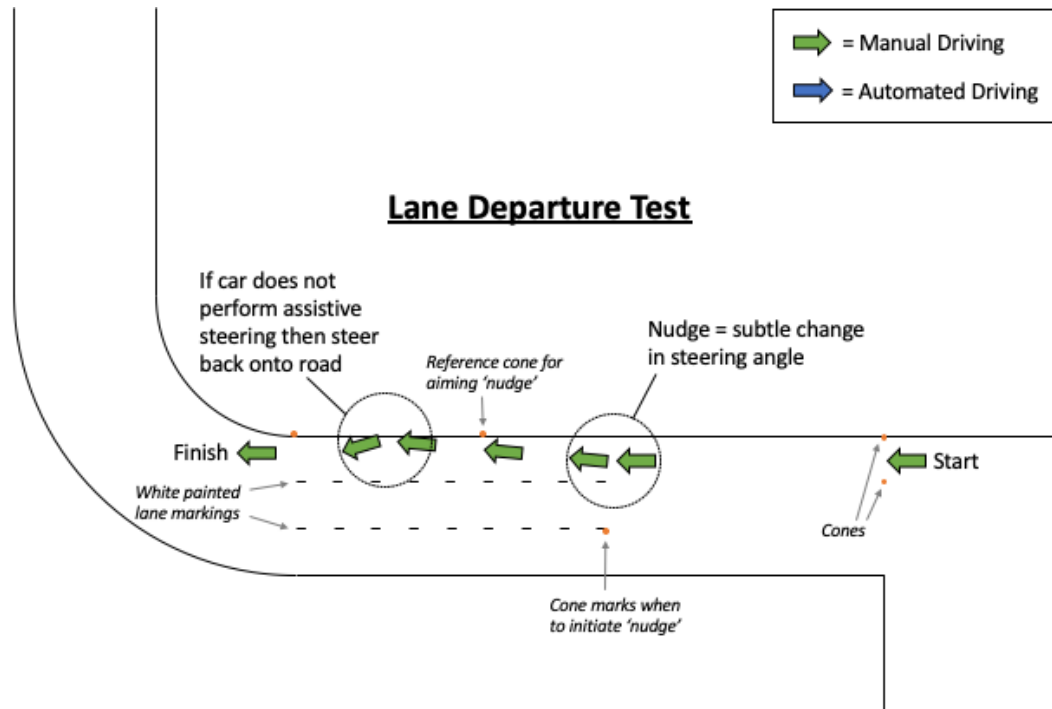
Lane Departure Test

Figure 5: Setup of test track for LD tests. The cars started between two traffic cones at the position marked 'start'. There was a 330-foot section of straightaway painted to depict three highway style lanes. A cone at the beginning of this section marked the point where the driver was supposed to apply a slight nudge to the steering wheel. This nudge was meant to aim the car just left of a cone 130 feet further down on the right side of the track.

Curve Test

The goal of the fourth and final experiment was to determine if there is significant within- and between-vehicle variability in the type and timing of feedback presented to a distracted driver when Autopilot can no longer adequately track lane markings. Given that computer vision is based on probabilistic reasoning with a potential large performance range, the hypothesis was that there would be significant variation in the timing and duration of events in the driver handover alerting sequence both between vehicles as well as for successive tests for the same vehicle.

A demanding driving scenario was created on the NCCAR test track in which the Teslas encountered the disappearance of lane markings and extreme curves. Prior to the data collection, the 330-foot section of straightaway was marked with highway-style white lane lines (10 feet long by 6 inches wide with 30 feet of longitudinal spacing between) to form three lanes 13 feet wide. The straightaway lane markings ended 250 feet after the start position, followed by two sharp curves, approximately 120 and 190 degrees. (Figure 6). This setup was intended to degrade the vehicle's model of the roadway to force initiation of the driver takeover sequence.

When Tesla Autopilot can no longer confidently track lane markings on the road, it requests that the driver take control of the car. As discussed in the highway test, requests in the Tesla Model 3 are displayed on a touch screen centered between the driver and passenger seat, shown in Figure 6. This display also displays all information about the car, including speed, rpm, and map.

If the driver does not grasp the steering wheel after the first request (Figure 6a), after approximately 10 seconds, the top left of the Autopilot display on the console flashes blue with a progressively increasing intensity (Figure 6b), culminating in two pairs of beeps. If the second request for driver control is ignored, after approximately 5 seconds the small icon and text are replaced with a large red icon and red highlighted text (Figure 6c), accompanied by three pairs of beeps. The hazard lights then activate, and the car will slow to a stop. The combined total alert duration from the first to third alert is designed to last 15 seconds (12).

a. First Alert   b. Second Alert   c. Third Alert

Figure 6: Progression of alerts in the takeover sequence. Left: The first alert includes a black bubble with an icon and the message "Apply slight turning force to steering wheel." Center: The second alert has the same icon and text but introduces a flashing blue light at the top of the screen. Right: The third alert displays a large red icon with hands on the steering wheel and includes a red bubble with the message "Autosteer unavailable for the rest of the drive."

To test whether and when these alerts would be triggered by the loss of lane markings, the vehicle began between a pair of traffic cones at the position marked "start" in Figure 7. Starting from its inactive, parked state, the vehicle was manually driven towards the painted lane lines in the inner most lane. After accelerating to 35 mph, the car was immediately placed in Adaptive Cruise Control to fix the speed. After passing the cones at the end of the fourth lane marking, Autopilot was activated, and the confederate driver never responded to any takeover alerts. The car was allowed to drive autonomously until the system reached the third alert without driver response, which should result in Autopilot shutting down the vehicle.

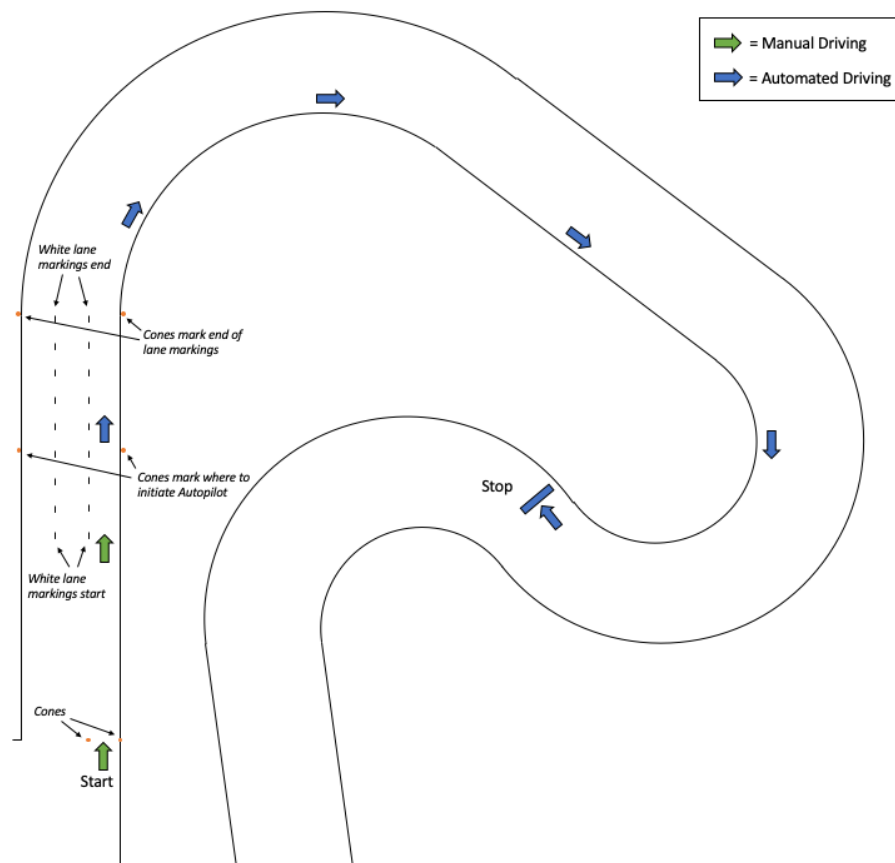**Figure 7: Test track setup for CRV tests. The cars started between two traffic cones at the position marked 'start'. There was a 330-foot section of straightaway painted to depict three highway-style lanes. A pair of cones on the outer edges of the track 130 feet past the beginning of the lane markings indicated the point at which Autopilot was engaged. Beyond this point, the driver allowed the car to**

drive autonomously without intervention until the car came to a stop on its own. There were no painted lane markings beyond the end of the straightaway and the car encountered two sharp curves with angles of approximately 120 and 190 degrees.

Data Collection

Video Data

Video data were collected using three GoPro Hero 7 Black cameras synchronized with SyncBac Pro devices, mounted at fixed positions in the vehicle interior. These cameras obtained views of the roadway, the driver, and the center console (Figure 8). The console-facing camera was intended to provide exact timing of when various alerts were presented on the center console. The time-synchronized data identified events of interest from the other camera views (i.e., actions taken by the driver or views of the road as seen from the forward-facing camera).

The road-facing camera was placed on the dashboard on a flat adhesive mount so that it was centered laterally with the front edge of the mount set back two inches from the front curved lip of the dashboard. The driver-facing camera was adhered to the dashboard with a curved mount facing directly backwards, perpendicular to the edge of the dashboard. The center of the mount was 20 inches from the driver-side edge of the dashboard and the front of the mount was aligned with an angled crease on the dashboard surface.

The console-facing camera was attached to the sunroof with a suction-mount and six-inch extender arm. The camera was positioned so the center of the suction mount was over the "T" logo on the sunroof with the rear edge of the mount flush against the edge of the sunroof. The camera was angled downward so that the entire console was visible and centered. All cameras were set to 1440 pixels per inch resolution, 25 frames per second, wide field of view, and automatic stabilization, with protune off. The next section details the results from the tests and data collected.
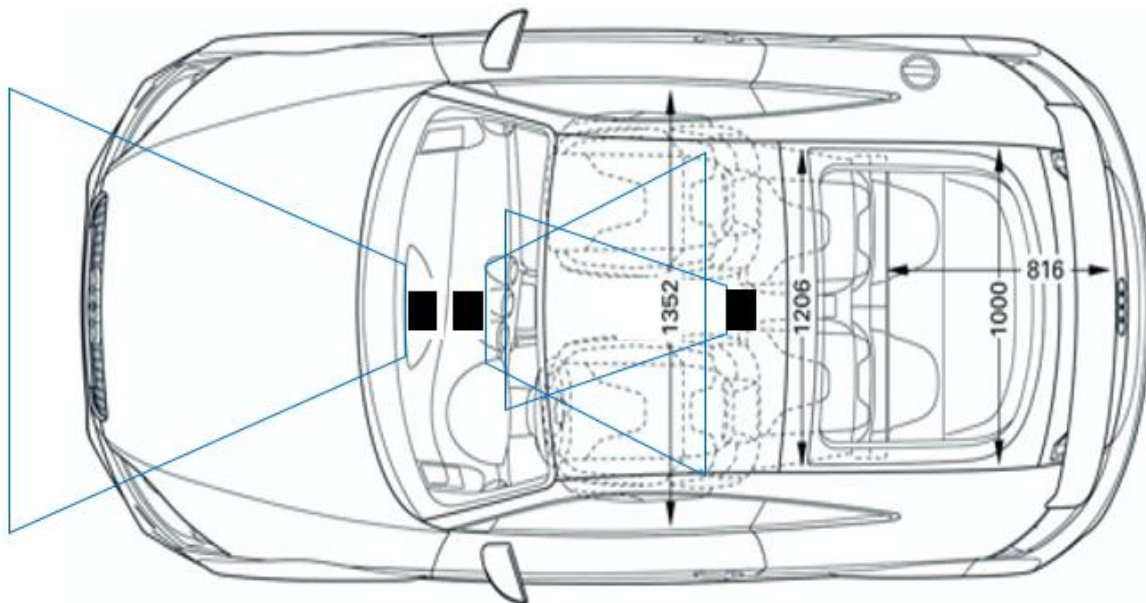

Figure 8: Data collection camera positions

Brightness

Given that brightness could influence camera perception, for those tests with directional changes (the lane shift, lane departure and curve tests), a measure of environmental brightness was developed. Information about scene brightness was extracted from video frames by converting the videos from RGB to

YCrCb color space using the Python OpenCV image processing toolkit (16). The brightness emanating from a particular region in a visual scene is defined as that region's *luminance* and is measured in *candela per square meter*. It can be obtained from a weighted sum of the red, green, and blue light in that region (17).

Since humans have limited capability in perceiving differences in light intensity, commercial cameras generally do not store data on the full spectrum of captured light. Instead, they compress data on the intensity of captured light by several orders of magnitude through a process called *gamma compression*, where some information is lost (18). The "Y" channel in a YCrCb image encodes the image's *Luma*, which is as close of an approximation of true luminance as can be obtained from gamma compressed digital imagery (19). Luma is a unitless index taking a value between 0 and 255. Other color models that encode light intensity, such as HSV, HSI, or HSL, do not match the perceptual qualities of brightness as well for certain colors (20), making YCrCb superior for extracting brightness from images.

The analysis of brightness was restricted to only the top 380 rows of pixels in the images from the forward-facing roadway camera which corresponded to regions of the scene that only contained the sky. This was done to reduce the influence of sudden changes of color, such as variations in the treeline or the appearance of distant objects. A number of studies have shown strong results in image processing tasks using luminance-based color models, including for tasks related to perceiving pixels in the sky (21–25).

Results

Videos for specific test metrics were manually scored to identify the first frame at which events of interest occurred and SyncBac timecode annotations were used to link the corresponding frames taken from different cameras. All alpha values for the statistical tests are .05 unless otherwise noted.

Highway Test

The goal of this test was to determine how consistent and timely the cars were in notifying drivers that their hands were no longer on the wheel. An alert cycle was defined as the following: (1) a period of hands-free automated driving; (2) the presentation of an alert requesting that the driver apply light force to the steering wheel; (3) a driver response; and (4) the disappearance of the alert, removal of the driver's hands from the wheel, and beginning of the next cycle (i.e., a return to automated driving). The driver response to alerts was a two-handed continuous "wiggle" of the steering wheel, deflecting it approximately 5 degrees in each direction, for as long as necessary to make the alert disappear. The driver continuously monitored the alert console so as to respond as quickly as possible when an alert appeared (Appendix C.1: Highway (HW) Test, Table *6*). Over the course of the 5.2 mi course, a typical run would include 7-8 such cycles.

Based on the observed data, there were three possible outcomes for each event cycle: *success*, *shutoff*, or *failure*. A cycle was a *success* if, after the driver responded to the alert, the alert disappeared and the car returned to automated driving, which is what it is supposed to do. A cycle concluded in a *shutoff* if, after the driver responded to the alert, the car did not return to automated driving and instead ceded control to the driver. This handover in control was associated with an auditory alert consisting of two chimes. A cycle concluded in *failure* if at any point during the cycle the car failed to operate safely while in Autopilot, such as if the vehicle veered off the road and struck a rumble strip.

Car	Total	Success	Shutoff	Failure
Car 1	62	61	1	0
Car 2	23	15	1	7
Car 3	64	61	3	0

Table 1: Counts of event cycle outcomes in the HW test.

Table 1 and Figure *9* summarize the counts of the event cycle outcomes observed for each car. Car 2 had a higher number of event cycles ending in failure, which resulted in fewer observed total event cycles. If the

driver was forced to takeover, Autopilot was not reengaged during the remainder of the 5.2 mi route for safety reasons. As a result, trials with a "shutoff" or "failure" event occurring early in a test trial led to fewer observed event cycles than trials in which the car drove the entire route on Autopilot. Frequencies of the outcomes were assessed using a chi-squared independence test, and the distributions were determined to be significantly different across cars (x^2 = 52.703, p < 0.0001).
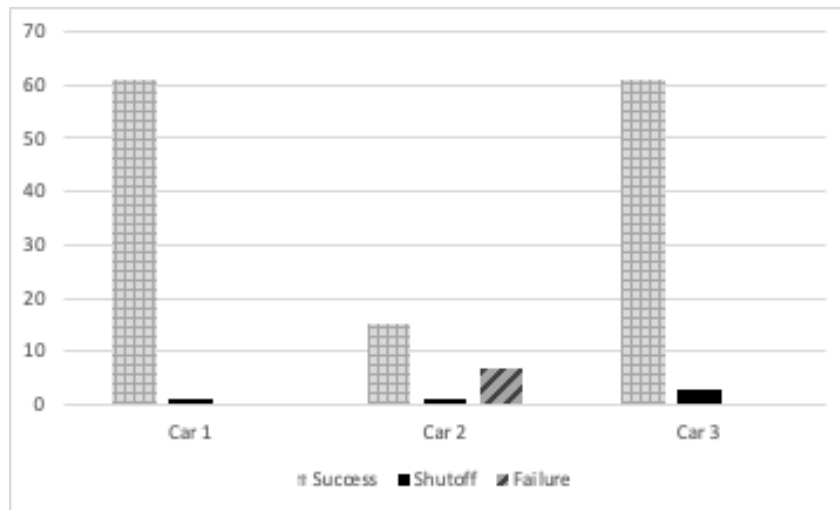
Figure 9: Counts of event cycle outcomes in the highway test.

Next, variability in the duration of hands-free driving during each event cycle was assessed. This interval is defined as the time between when the driver's hands left the wheel to when the next alert appeared on the vehicle's console. According to Tesla documentation, this interval is designed to decrease linearly with increasing speed (12), with a maximum duration of 60 seconds at 25 mph and a minimum of 10 seconds at 90 mph. Therefore, at 70 mph, the expected duration of hands-free driving between alerts is 25.38 seconds. A mean duration of this interval of just over 30 seconds occurred for all three cars (Appendix C.1: Highway (HW) Test, Table *7*). Several durations of longer than 40 seconds were observed during cycles when the Tesla traveled slower than 70 mph due to slower-moving lead vehicles.
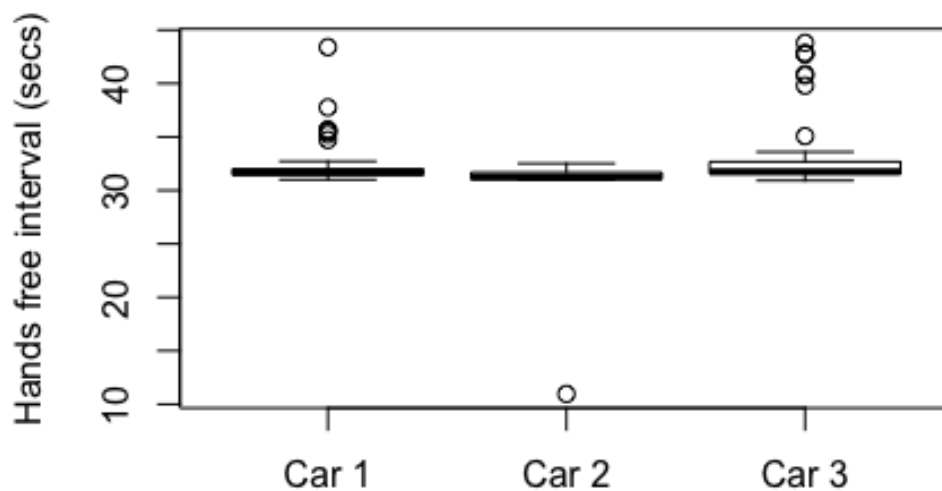
Figure 10: Distribution of event cycle hands-free driving intervals for each car.

To determine if there was any statistical difference in the duration of hands-free driving intervals between cars, controlling for possible speed changes, an analysis of covariance test was conducted with average speed as the covariate. Speed was estimated by averaging the displayed speed at the beginning and end of the alerting interval. This analysis was significant for both speed ($F_{(1,139)} = 260.25$, p < 0.0001) and car ($F_{(2,139)} = 5.58$, p = 0.0047; see Appendix D.1: Highway (HW) Test, Table *17*). A Tukey-post hoc test

with Bonferroni-adjusted significance level of 0.0167 revealed that there was a significant difference in the main effect between car 3 and car 2 (p = 0.003). Additionally, a contrast between car 3 and the pooled data of cars 1 and 2 revealed a significant difference in the mean interval of hands-free driving (p = 0.016).

A regression model of speed vs. the hands-free driving interval revealed that, at least for the speed ranges observed in testing, for every .93 mph decrease, there is a 1s increase in the hands-free driving interval. By comparison, Tesla documentation indicates that by design, this time interval should increase by 1s for every 1.3 mph decrease in speed (12). See Appendix D.1: Highway (HW) Test, Table *18* for additional details.

The times from when hands first touched the steering wheel to the times when the alert disappeared from the console (Appendix C.1: Highway (HW) Test, Table *8*) were also analyzed. This is important since a driver may become overly focused on clearing the alert, and so this represents another possible source of distraction. The ANCOVA model did not detect a significant effect for either car ($F_{(2,138)}$ = 2.164, p = 0.1188) or speed ($F_{(1,138)}$ = 3.094, p = 0.0808; Appendix D.1: Highway (HW) Test, Table *19*).

Lane Shift Test

The goal of this test was to determine within- and between-vehicle variability in the vehicle's behavior when encountering an unexpected road pattern, in this case a construction site. The vehicles autonomously navigated a simulated construction site including a painted lane shift and a barricade of traffic cones (Figure *4*). Whether vehicles presented a driver takeover alert was assessed, as well as what point in the trial such an alert occurred and whether the vehicle successfully maneuvered to avoid hitting the traffic cones.

In terms of maneuvering to avoid obstacles, Cars 1 and 3 avoided all cones on all 10 of their trials, while Car 2 failed to maneuver away from the cones on all 10 of its trials (Figure 11). Differences in the counts of each observation for each car were analyzed using a chi-squared independence test, which detected a significant difference between cars (x^2 =130.0170, p < 0.0001).
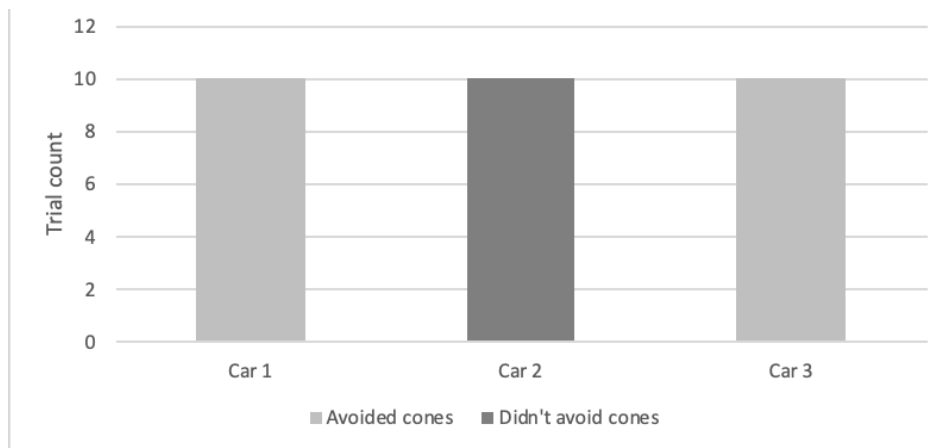
Figure 11: Counts of trials avoiding vs. not avoiding all cones for each car in the LS test.

Separate from the cars' ability to avoid the traffic cones, variability existed for each car in terms of whether an alarm was presented upon nearing the cones. While it was not clear whether Cars 1 and 3 guided on the cones or the yellow line, whether an alert was generated indicates if a car detected the obstacles. Cars 1, 2, and 3 had 6, 3, and 7 trials in which an alarm was presented respectively (Figure 12). Differences in the counts of each observation for each car were analyzed using a chi-squared independence test, which did not detect a significant difference between cars (x^2 = 3.4821, p = 0.1753). Overall, the driver was not alerted in 47% of trials. If Car 2 is disregarded, this rate is 35%.
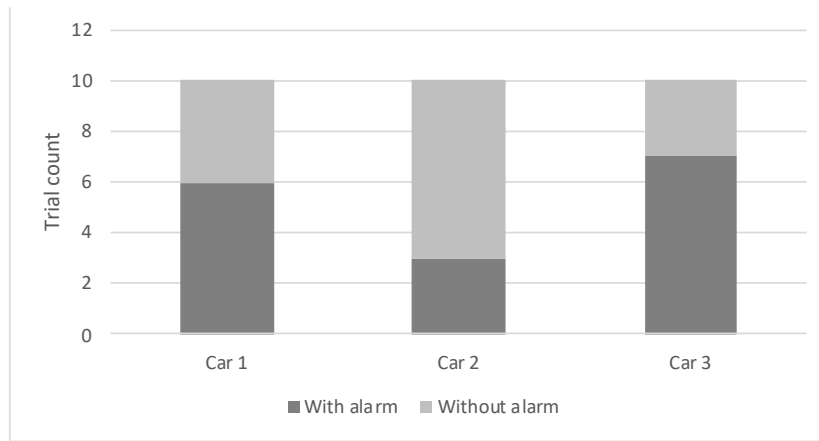
Figure 12: Counts of trials with vs. without an alarm for each car in the LS test.

Whether atmospheric brightness contributed to differences in whether cars presented an alert was examined. The luma for each car was computed as outlined in the Methods section at the moment the car passed the first roadside cone indicating the beginning of the lane shift. An analysis of covariance model was developed with the luma as the dependent variable and alert presence/absence as the main factor. Sun angle (azimuth), based on the date and time each trial was conducted, was included as a covariate because it has been cited as a potential source of problems for Autopilot (26).

With alert as an independent variable, car as a block and sun azimuth as a covariate, an ANCOVA found that luma was significantly predicted by both car ($F_{(2,25)} = 34.163$, $p < 0.0001$) and sun azimuth are ($F_{(1,25)} = 5.064$, $p = 0.0335$; Appendix D.2: Lane Shift (LS) Test, Table *20*). These results indicate that each car experienced different brightness levels, which again is likely related to sun angle, but that the presence of an alert was not a likely contributor to sensed brightness levels.

Data from the forward-facing camera was used to estimate the location at which the alarm was sounded (Appendix E: Views of Road at Time of Alert in Lane Shift Test) by computing the area of traffic cone visible, to the nearest 10% of a cone (Appendix C.2: Lane Shift (LS) Test, Table *9*). This metric is robust because cones were placed at fixed locations that did not vary across trials. Using this analysis, a one-way ANOVA detected a significant difference in the quantity of cones visible between cars ($F_{(2,13)} = 25.52$, $p < 0.0001$; Appendix D.2: Lane Shift (LS) Test, Table *21*). A Tukey Kramer post hoc test with Bonferroni adjusted significance threshold of 0.0167 revealed a significant difference between car 3 and car 2 ($p = 0.0055$) as well as between car 3 and car 1 ($p < 0.0001$). This means that while there was no statistical difference between the cars for the number of alerts generated, there was a statistical difference in where the cars generated alerts, illustrated by Figure 13.
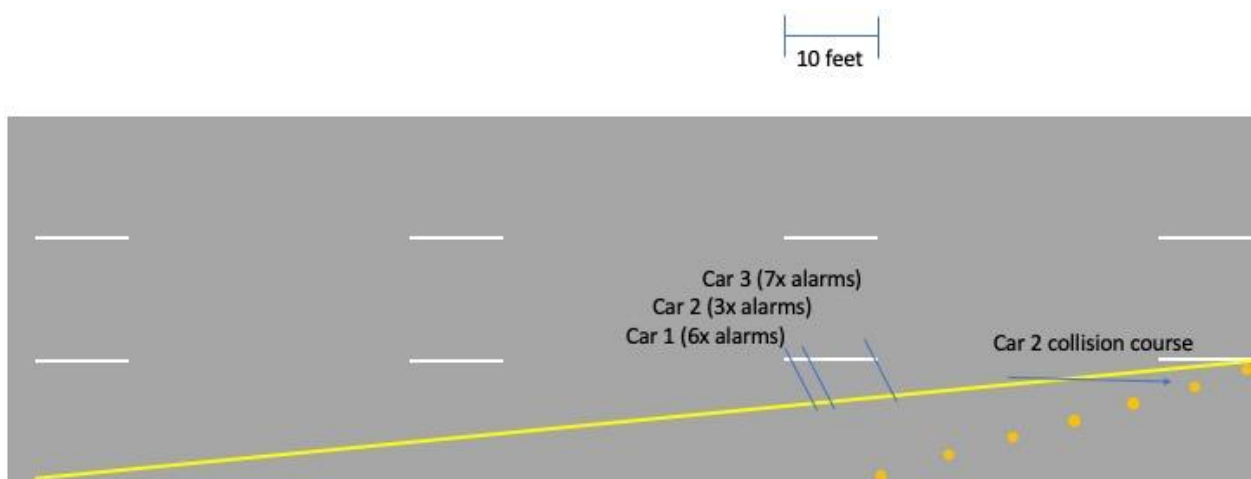
Figure 13: Approximate average location of alarm and collision events for each car in LS test.

Figure *13* also depicts Car 2's approximate trajectory toward the traffic cones, as this was the only vehicle that failed to avoid all of the obstacles during the test. However, even though it failed to steer the car away during any trial, it alerted the driver in 30% of the trials. When an alarm occurred, each car was internally consistent in where it presented the alarm, but cars did not present alarms at the same locations as one other. Car 3 progressed furthest through the construction site before presenting an alarm, approximately 10 ft beyond where Cars 1 and 2 sounded their alarms.

Lane Departure Test

The goal of this test was to evaluate between- and within-car variation in the timing of emergency assistive steering application if the car drifted towards the edge of the road during automated cruise control driving. This was simulated by having cars drive along a straightaway on the test track and at a fixed point. The driver provided a slight nudge to the steering wheel to aim the car towards an area on the outer edge of the road (Figure *5*). Because all cars were configured to provide emergency assistive steering, emergency assistive steering should have engaged in all trials. Two trials were discarded because the driver's nudge did not result in a trajectory that took the car outside the lane.

There were three different outcomes across trials: emergency assistive steering in conjunction with an alarm, an alarm only, and neither alarm nor steering. Counts of these three outcomes per car are shown in Figure *14* as well as Appendix C.3: Lane Departure (LD) Test, Table *11*. While cars were not consistent in terms of their individual performances, a chi-squared independence test did not reveal a significant difference between the distribution of counts between cars (x^2 = 3.4375, p = 0.4874). Overall, in 50% of trials, no alert or assist was provided, meaning that if the driver had been truly distracted, half the trials would likely have resulted in a crash.

To determine the consistency of the driver's angular input to nudge the car on a road departure trajectory, the angle of wheel rotation was estimated from the forward-facing cameras by computing the degree of rotation of the cross bar on the steering wheel from the point at the beginning of the driver's nudge to the point of maximum deflection. Video frames were manually extracted for the beginning and peak deflection of the nudge for each trial, annotating the pixel locations of the upper right and left corners of the crossbar, and computing the rotation of that line between the two timepoints. Mean peak angle of rotation was approximately 4 degrees for each car (Figure *15*; Appendix C.3: Lane Departure (LD) Test, Table *10*).
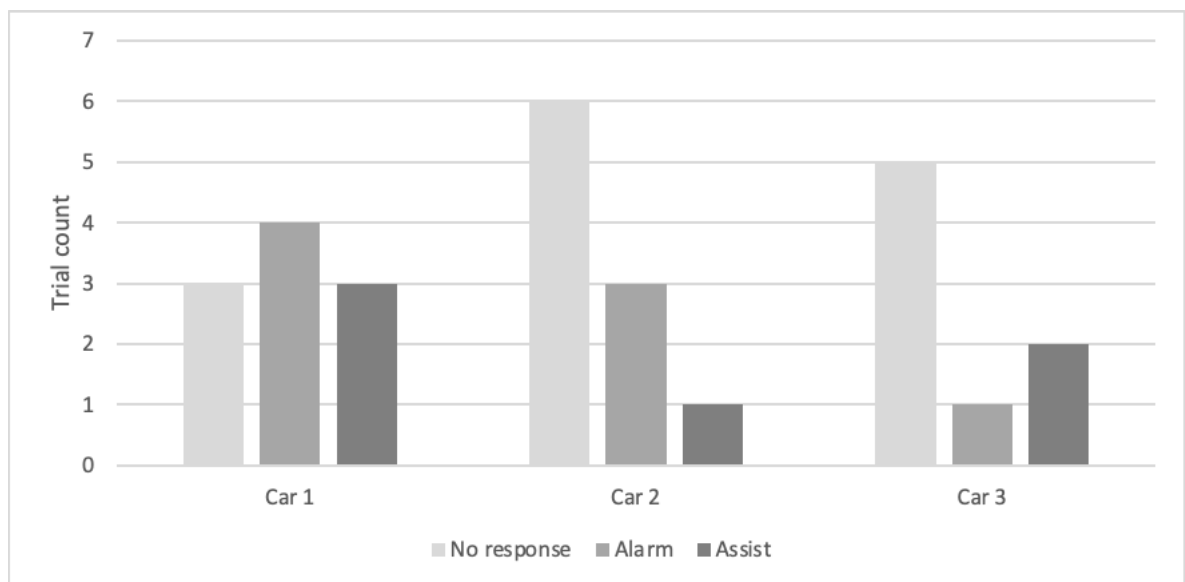
Figure 14: Trial outcomes by car in the LD test.

A blocked ANOVA was used to assess whether systematic differences in how the steering nudge was applied contributed to different trial outcomes. With the angle of nudge as the dependent variable, there was

no significant effect for either car ($F_{(2,23)} = 0.242$, $p = 0.7870$) or the trial outcome ($F_{(2,23)} = 0.122$, $p = 0.8860$), indicating that variation in wheel rotation was not systematically different between cars and not obviously correlated with particular trial outcomes (Figure *15*; Appendix D.3: Lane Departure (LD) Test, Table *22*).
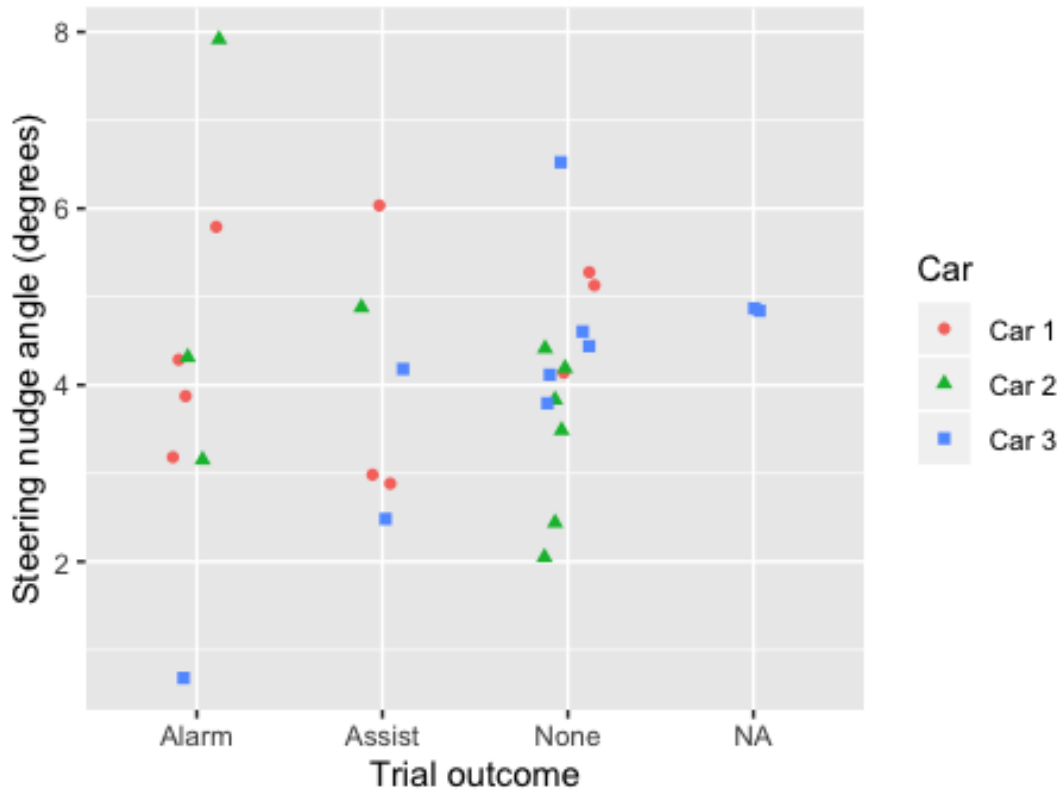
Figure 15: Trial outcomes by car and degree of wheel rotation during "nudge" phase in lane departure test.

To assess the role of lighting immediately prior to the event of interest (presence and type of feedback upon exiting the roadway), sky luminosity was measured at the moment the car reached its peak lateral position (i.e., when it had traveled furthest to the edge of the road just before corrective steering was applied, regardless of whether that steering was initiated by the driver or by the ADAS system). Using sun azimuth as a covariate, an analysis of covariance was conducted to determine whether individual vehicle or type of feedback was associated with luminosity, (Appendix D.3: Lane Departure (LD) Test, Table *25*).

Type of trial outcome was not associated with a significant difference in luminosity ($F_{(2,22)} = 0.037$, $p = 0.964$), though car was ($F_{(2,22)} = 6.695$, $p = 0.005$). Additionally, the covariate sun azimuth was not significant ($F_{(1,22)} = 0.065$, $p = 0.801$), indicating that different luminosity was observed through the front windshield of each vehicle, but sun angle in these tests was not a factor. It should be noted that the lane departure tests were conducted 180 degrees longitudinally from the lane shift test, where sun angle was a factor. Lane departure tests were conducted with the sun generally facing the car (which was headed west), while the lane shift tests generally had the sun shining from the rear of the car (which was headed east).

Curve Test

Because the curve test was designed to simulate a challenging driving environment in which Autopilot may not be able to maintain control after the loss of lane markings, a driver takeover alert should be presented in all cases. The key metrics assessed were (1) the time duration between the first alert and the second alert; (2) the time duration between the second alert and the final alert; (3) the distance traveled before the alarm sequence was initiated; and (4) whether brightness played a role in these alerts. It is worth

noting how well the cars performed in general in terms of controllability. Despite the challenging course and the driver's intentional ignoring of the takeover requests, Autopilot successfully maintained control through the test durations and brought the cars to a safe stop in all 30 trials without human input.

Variation in the time period between the first and second alert was minimal and corresponded with the 10s specification provided by Tesla (12), with a mean of 10s and a standard deviation of approximately 0.1 seconds across the entire dataset (Figure *16*a). A repeated measures ANOVA did not reveal any significant differences between cars or within cars across the different trials ($F_{(1,25)} = 0.001$, $p = 0.973$; Appendix D.4: Curve (CRV) Test, Table *26*).

Variation in the duration of the second alert stage was also small, although car #2 skipped this second alerting stage (Figure *16*b). Regardless of this outlier, all three vehicles had a median value of 5.0 seconds, again matching the 5s Tesla specification. A repeated measures ANOVA did not reveal any significant differences between cars or within cars across the different trials ($F_{(1,26)} = 0.721$, $p = 0.404$; Appendix D.4: Curve (CRV) Test, Table *27*).

Variation in the overall duration of the alert sequence was also small (Figure *16*c). Again, the data for all three cars closely matched the Tesla specification with a median duration of 15s across the dataset. A repeated measures ANOVA did not reveal any significant differences between cars or within cars across the different trials ($F_{(1,26)} = 0.772$, $p = 0.388$; Appendix D.4: Curve (CRV) Test, Table *28*).

a. Time duration between 1st and 2nd alerts   **b. Time duration between 2nd and 3rd alerts**   **c. Time duration of entire alert sequence**
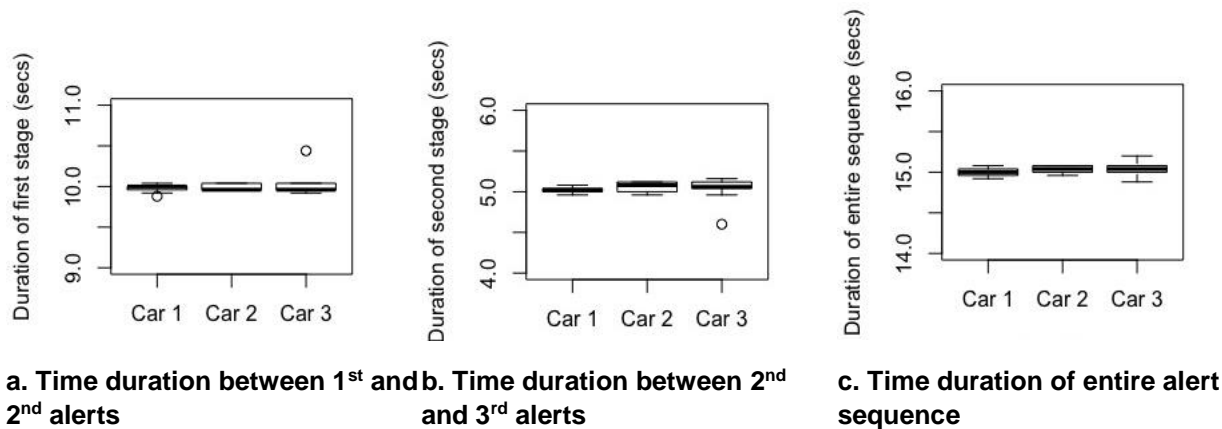
Figure 16: Duration of alert stages across cars. Left: Duration of interval between first and second alert. Center: Duration of interval between second and third (final) alert. Right: Duration of total sequence.

While no meaningful variation in the timing of the takeover alert sequence was detected once initiated, significant variation was detected in where the first alert was initiated. The first alert location was estimated by comparing the time the first alert occurred with the time the vehicle first entered the curved portion of the track (marked by a pair of traffic cones that were easily identifiable in the frames of the video data collected, Figure *7*). Approximate distance was computed from this time interval by multiplying the vehicle's reported speed (fixed at 35 mph for the duration of each test) over the time interval traveled with the position confirmed through a manual inspection of the video.

Figure *17* shows the approximate position of each car when the alert sequence was first initiated for every trial. Each trial is represented with a car-specific mark, where car #1 is depicted by a red X, car #2 a yellow circle, and car #3 a blue plus sign. The mark illustrates the approximate longitudinal position along the track, with lateral variation included solely to improve readability. It is important to remember that all 30 trials began in the same place and at the same speed, with Autopilot engaged.

It is clear from Figure *17* that there were three distinct areas where the first alert occurred, with separation between cluster centers. An iterative K-means analyses using 1-10 clusters and assessed with the Gap statistic yielded K = 3 statistical clusters (**Figure 18**). The first cluster (Zone 1) was centered at 121 feet after the beginning of the curve, with a range of 43 to 172 feet; the second cluster (Zone 2) was centered at

637 feet with a range of 609 to 768 feet; and the third cluster (Zone 3) was centered at 1248 feet with a range of 1240 to 1255 feet. The distribution of these distances for each car are shown in Figure *19*a. Clusters were not equally represented across cars; Car #1 and Car #3 experienced alert initiation points in Zones 2 and 3, while Car #2 experienced alert initiation points in Zones 1 and 3 (Table 2; Figure *19*b). A chi-squared independence test revealed a significant difference in the counts between cars (x^2 = 22.4, p = 0.0002).

Figure 17: Approximate location of the first alert for each trial in the Curve test.

Table 2: Counts of alert location cluster by car in the Curve test.

Alert location	Car #1	Car #2	Car #3
Zone 1	0	5	0
Zone 2	9	0	6
Zone 3	1	5	4

Figure 18: Optimal number of clusters in data for distance from curve to alert sequence initiation based on K-means clustering with the Gap statistic in CRV test.

a. **Distance traveled by each car from first curve until first takeover alert is presented**

b. **Number of alerts at each zone by car**

Figure 19: Position on the track of first alert in the takeover sequence (left) and counts of where the first alarm was initiated for each car in CRV test (right).

Given that there were clear differences in where the cars experienced the first alert due to each car's loss of lane marker tracking, any differences between the brightness levels perceived by the cars in Zone 1 vs all other events were investigated. The luma was computed for each car as outlined in the Methods section for the 150-foot segment of track immediately preceding Zone 1, a distance roughly 3s prior to the alert. Appendix F illustrates the brightness at each of the initial alert locations. An analysis of covariance model was developed with the luma as the dependent variable and alert zone as the main factor. Sun angle (azimuth), based on the date and time each trial was conducted, was included as a covariate because it has been cited as a potential source of problems for Autopilot (26).

The results were marginally significant ($F(1,27) = 3.87$, $p = .06$), meaning that the trials with alerts in Zone 1 experienced different brightness levels than those without alerts. The sun angle covariate was marginally significant ($F(1,27) = 3.39$, $p = .08$). Car 2, the only car to experience the initial alert in Zone 1, had a mean luma of 178 and the rest were at a mean of 163. The mean sun angle for car #2 in Zone 1 was 189 degrees and 212 degrees for the no alert group in Zone 1. Thus, one possible influencing factor is not only how much brightness is experienced when lane lines are lost but also the sun angle.

A backwards logistic regression (LR) model was used to predict whether the alert occurred in Zones 1 and 2 as a function of sun angle and brightness in the 3s prior to each alert. In this model, luma was also

marginally significant (p=.06, B = -.176) and sun angle was not (model accuracy = 85%). Because LR models produce regression coefficients for each feature that are log odds, taking the exponential of the coefficient weights estimates the expected change in the log odds of the target variable per unit increase in the corresponding predictor variable, holding the other predictor variables constant. This means that as luma increased by one unit there was a 12% increase in the likelihood the car would experience a Zone 1 initial alert.

Discussion

The goal of this study was to examine whether there were significant between- and within-vehicles driver-alerting differences in three randomly selected Tesla vehicles of the same model and year (Model 3S, 2018). Table *3* summarizes the general levels of consistency of each vehicle platform across the four driving tests. As will be discussed in detail, the bulk of tests yielded dramatic inconsistencies both within a single vehicle as well as across the vehicles.

Between-vehicle differences were observed across numerous metrics. Cars 1 and 3 generally performed similarly, but not always. Overall behavior of Car 3 tended to appear less "cautious" than Car 1. For example, compared to Car 1, Car 3 tended to drive in autopilot for longer durations in the sharp curve scenario before forcing the driver to take control. It was also less likely to provide lane departure alerts on the lane departure test, and when it did, it was less likely to supply emergency assistive steering in conjunction with the alert. Car 3 also traveled further into the simulated construction site before presenting an alert to the driver.

Despite the performance differences between Cars 1 and 3, they were overall more similar than they were different and varied primarily in subtle aspects of their execution of the different driving tasks. Conversely, the behavior of Car 2 was substantially different from both the other cars. During track testing, Car 2's behavior was erratic on multiple tasks. Despite having an overall flawless record of successfully braking on the sharp curve task, it had the highest degree of within-vehicle variability in distance traveled before presenting a takeover alert, as it accounted for the 5 shortest distances as well as the longest 3 distances. Similarly, on the construction test, Car 2 failed to maneuver around the obstacles in any trial, and in spite of this, was also least likely to initiate a takeover alert.

Car 2's behavior during highway testing was very unpredictable. The car vigorously pinballed from side-to-side in the lane almost immediately upon autopilot engagement and routinely hit the rumble strips, triggering an end to the test, which is why there were far fewer total observations for Car 2. Curiously, the car's performance seemed to improve marginally over progressive trials. This behavior was also observed for Car 2 outside the formal test context; when driving to the track, the test driver struggled to use autopilot consistently on the highway. Additionally, during both automated and driver-operated control, on multiple occasions the vehicle produced abrupt lane departure alarms.

Other abnormalities were noted as well, such as Car 2 inexplicably skipping an alert cycle during one trial of the curve test, as well as presenting a hands-on-wheel alert after only 11 seconds while driving at 70 mph, with the typical alert occurring at 32s. Because the most recent over-the-air update had occurred the evening before testing began, the owner had not used autopilot while the car was using the most recent version of the software. However, the owner reported that while using prior software versions, he had experienced similar issues during the first 1-2 hours after supercharging and that they gradually improved over time. The vehicle was supercharged immediately prior to the highway tests as well as approximately 1 hour prior to the track tests, which may explain some of the odd behaviors seen in testing. Future research efforts should look at the interaction between charging and vehicle control.

In addition to the significant between-vehicles differences that were present, within-vehicles differences were observed on multiple metrics. The only metrics that were generally consistent were (1) the interval of hands-free driving prior to an alert in the highway task, (2) the intervals between takeover alerts in the curve task, and (3) the location of takeover alert (when they occurred) in the construction task. In the construction task, while vehicles were internally consistent in where they presented an alert, they were not consistent in whether they presented an alert.

Test	Metric	Car 1	Car 2	Car 3
Highway	Alert interval	C2	CI	I
	Time to clear	CA	CA	CA
	Unsafe behavior	2%	35%	5%
Lane Shift	Sounding of alert	I	I	I
	Location sounded	CW	CW	CW
	Unsafe behavior	0%	100%	0%
Lane Departure	Alert sounded	I	I	I
	Steering Assist	I	I	I
	Unsafe behavior	30%	60%	63%
Curve	1st Alert	CW	I	I
	2nd Alert	CA	CA	CA
	3rd Alert	CA	CA	CA
	Unsafe behavior	0%	0%	0%

CA = Consistent All, C(1,2,3) = Consistent with Car 1, 2, or 3, CW = Consistent Within a single car, I = Inconsistent, percentages represent proportion of trials the behavior was exhibited

Table 3: Overall vehicle consistencies across tests

In the lane departure test, arguably the most interesting source of within-vehicle variation was the type of feedback presented when the vehicle approached the edge of the lane and roadway edge. Although the cars were configured to provide emergency assistive steering in all trials, they did so in only 21% of cases. However, they did present an alarm without providing assistive steering in another 30% of cases, indicating that the vehicle had at least acknowledged the imminent lane departure but could not provide steering.

Perhaps most concerning were the unsafe behaviors exhibited in all the tests but the curve test. In this effort, unsafe behaviors are defined as behaviors (or lack of alerting) that would have likely led to an adverse event given a distracted driver. For the highway tests, Car 2 was the most unsafe car primarily because the driver had to manually take over due to unsafe Autopilot behavior with no warnings. While Cars 1 and 3 were less unsafe, they did experience high risk events when Autopilot unexpectedly disengaged, which may not be observed by a distracted driver. The Lane Departure tests exposed many unsafe behaviors in that Car 3 had the highest number of unsafe lane departure incidents where no alert was sounded, nor was any steering assistance provided. Car 1 did not alert the driver to 3 out of 10 imminent road departures in the Lane Departure tests, so its performance was also highly variable and occasionally unsafe.

While there were no unsafe events for the Curve test by the previous definition, there was significant increased risk for an adverse event as 30% of the cars traveled ¼ of mile without a lane marking and with no alert to a driver with no hands on the wheel. At 35mph, this translates into 26s. Similar to the highway tests where cars traveled ~30s with no hands on the wheel, the question must be asked as to whether it is safe or advisable to let a car that requires driver attention to travel on a highway or on an extremely curvy road with the driver not paying attention and with no hands on the wheel? Finally, in the curve test while Car 1 was somewhat consistent in where it initiated the takeover alert sequence, Car 3 showed modest variation and Car 2 showed extreme variation.

Another issue with potential significant impact is that even where vehicles performed consistently with themselves and with each other, they did not behave consistently with stated design specifications. For

instance, according to Tesla documentation, during automated driving on 70 mph highways, hands-on-wheel alerts should be presented every 25 seconds (12), but this actual interval was 32 seconds, a 25% discrepancy. Similarly, Tesla notes that during automated driving if there is a disappearance in lane markings, a takeover alert should be presented within 400ms. However, not only did none of the cars present an alert with 400ms during the Curve test, in 30% of trials the vehicles drove autonomously for nearly 30 seconds on extreme curves that lacked even a single lane marking. The potential for safety-critical events under these circumstances is enormous.

This research also highlights the possibility that environmental factors could influence camera vision systems in potentially surprising ways. In the curve test there was some evidence that brightness and possibly sun angle could be influencing factors. The lane shift tests indicated a strong association between brightness and different cars, whereas this relationship was not true for the lane departure tests, held 180 degrees opposite. Sunlight has been shown to cause anomalies in the perception systems used in these and other autonomous vehicles (27–30), so it is possible that sun angle could contribute to a camera system's degradation. While these results suggest that there could be an important connection between brightness, sun angle and the triggering of camera vision-based alerts, more work is needed to further investigate these findings.

Limitations

The analysis of brightness-based contributors to vehicle performance variation is limited by the choice of camera used to collect the data. For this study, the camera was set to automatically adjust its exposure; this provided a consistent level of image contrast, enabling us to detect console alerts under a variety of light conditions, but also masks some of the variation in natural light present in the environment. While variations in luma of up to approximately 20% were detected across images, this may not have represented the true variation in light across the different scenes. More work is needed with light meters for better accuracy. In addition, the vulnerability of computer-vision systems engaging in lane tracking and obstacle detection to different sun angles deserves further scrutiny.

Small sample size was also a limitation of this study. One major difference across the three vehicles that could account for some of the observed variation was that despite their identical model and year, the cars had different software versions at various points in the 11-day testing time period. Car 3 completed all testing using software *v10.2 (2020.4.1 4a4ad401858f)*. Car 1 completed the track tests with this same software version but completed the highway tests with software *v10.2 (2020.8.1 ae1963092ff8)*. Car 2 completed all tests with a third software version, *v10.2 (2020.12 4fbcc4b942a8)*.

In addition to the different software versions, Car 2 included a full-self driving chip and All Wheel Drive while Cars 1 and 3 just had standard autopilot and a single motorized axle. Although the full-self driving chip was present on Car 2, the associated full-self driving visualization was disabled to make the car's driver monitoring and alerting system as consistent as possible with the other vehicles. As a result of these hardware and software variations, some driving configuration options differed between cars and it was not possible to operate them in exactly the same settings. For example, Car 1 was set to only allow "chill" acceleration mode, while Car 2 did not have this option and was instead operated in "sport," and where Car 3 used the factory default acceleration mode "standard". None of the tests theoretically should have been impacted by the acceleration mode, but since the logic of the cars' decisions is a black box, this cannot be certain.

Overall, the small sample size makes it difficult to distinguish individual vehicle differences from differences arising from the unique software configurations present in each vehicle. However, the presence of these significant differences is itself noteworthy, regardless of the root cause. Modern vehicle certification frameworks do not consider variation across individual vehicles in a class, so significant between-vehicles differences are not currently accounted for regardless of their source.

Conclusions

The goal of this research was to assess between- and within-ADAS-equipped vehicle variation in four key scenarios involving the interface between a human driver and an ADAS system. To this end, three Tesla Model 3 vehicles displayed significant between- and within-vehicle variation on several metrics related to driver monitoring, alerting, and safe operation of the underlying autonomy.

These results suggest that the performance of the computer vision systems was extremely variable, and this variation was likely responsible for some, but not all, of the delays in alerting a driver whose hands were not on the steering wheel. Ironically, in some trials the cars seemed to perform the best in the most challenging driving scenarios (navigating extreme curves while the driver ignored takeover requests), but performed worse on seemingly simpler scenarios like detecting a lane departure or responding to obstacles in the roadway.

This finding highlights a common misconception that what humans perceive to be hard in driving may not necessarily be what an autonomous system finds difficult. It may be that the extreme road angles in the curve test were detected more easily as opposed to the road edges in a much more gradual drift in the lane departure test. Another possibility is that Tesla engineers spend more effort on the more difficult problems and spend less time on seemingly easy problems. Whatever the reason for such variable and often unsafe behaviors, these results indicate that more testing is needed for these vehicles before such technology can operate without humans in direct control. It also suggests that driver monitoring systems need to have both a high degree of certainty as well as rapid response times.

These results should be interpreted considering the discrepancies in the software/hardware configurations of the vehicles, which present a confound for assessing the nature of performance variation. Despite the very similar configurations of Cars 1 and 3, they completed the tests using different versions of software. Car 2 possessed the purported "full self-driving chip," so in theory should have the most advanced Autopilot system, but this car objectively performed the worst.

Such results also indicate that the concept of over-the-air updates needs to be revisited when safety-critical functionalities may be changed. While agile software engineering techniques may be suitable for smartphones and other similar devices, these techniques likely cause significant problems in safety-critical systems. Unfortunately, these processes have never been formally studied or evaluated by a regulatory body. Indeed, these results highlight the need for more scrutiny of the cars and software embedded in them, as well as the certification processes, or lack thereof, that allow these cars on the road.

Acknowledgements

References

1. Canalys. 10% of new cars in the US sold with level 2 autonomy driving features [Internet]. Canalysis; 2019 Sep [cited 2020 Jul 29]. Available from: https://www.canalys.com/newsroom/canalys-level-2-autonomy-vehicles-US-Q2-2019?time=1596044908

2. SAE International. SAE International Standard J3016 [Internet]. SAE International; 2014. Available from: https://cdn.oemoffhighway.com/files/base/acbm/ooh/document/2016/03/automated_driving.pdf

3. Favarò FM, Seewald P, Scholtes M, Eurich S. Quality of control takeover following disengagements in semi-automated vehicles. Transportation Research Part F: Traffic Psychology and Behaviour. 2019 Jul 1;64:196–212.

4.   Miller EE, Boyle LN. Adaptations in attention allocation: Implications for takeover in an automated vehicle. Transportation Research Part F: Traffic Psychology and Behaviour [Internet]. 2019 Oct [cited 2019 Nov 3];66(0). Available from: https://trid.trb.org/View/1650717

5.   Kim HJ, Yang JH. Takeover Requests in Simulated Partially Autonomous Vehicles Considering Human Factors. IEEE Transactions on Human-Machine Systems. 2017 Oct;47(5):735–40.

6.   NHTSA. Test Procedures [Internet]. NHTSA. 2018 [cited 2020 Jul 19]. Available from: https://www.nhtsa.gov/vehicle-manufacturers/test-procedures

7.   IIHS. Test protocols and technical information [Internet]. IIHS-HLDI crash testing and highway safety. [cited 2020 Jul 19]. Available from: https://www.iihs.org/ratings/about-our-tests/test-protocols-and-technical-information

8.   Euro NCAP. Safety Assist | Euro NCAP [Internet]. [cited 2020 Jul 19]. Available from: https://www.euroncap.com:443/en/for-engineers/protocols/safety-assist/

9.   MOLIT. MOLIT Ministry of Land, Infrastructure and Transport [Internet]. [cited 2020 Jul 19]. Available from: http://www.molit.go.kr/english/USR/BORD0201/m_28286/DTL.jsp?id=eng0301&cate=&mode=view&idx=2905&key=&search=&search_regdate_s=&search_regdate_e=&order=&desc=asc&srch_prc_stts=&item_num=0&search_dept_id=&search_dept_nm=&srch_usr_nm=N&srch_usr_titl=N&srch_usr_ctnt=N&srch_mng_nm=N&old_dept_nm=&search_gbn=&search_section=&source=&search1=&lcmspage=1

10. Bredereke J, Lankenau A. A Rigorous View of Mode Confusion. In: Anderson S, Felici M, Bologna S, editors. Computer Safety, Reliability and Security. Berlin, Heidelberg: Springer; 2002. p. 19–31. (Lecture Notes in Computer Science).

11. Hawkins AJ. No, Elon, the Navigate on Autopilot feature is not "full self-driving" [Internet]. The Verge. 2019 [cited 2020 Sep 29]. Available from: https://www.theverge.com/2019/1/30/18204427/tesla-autopilot-elon-musk-full-self-driving-confusion

12. NTSB. Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator Mountain View, California March 23, 2018. National Transportation Safety Board; 2020 Feb. Report No.: NTSB/HAR-20/01 PB2020-100112.

13. NTSB. Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida May 7, 2016. National Transportation Safety Board; 2017 Sep. Report No.: NTSB/HAR-17/02.

14. Mkrtchyan AA. Modeling operator performance in low task load supervisory domains [Internet] [Thesis]. Massachusetts Institute of Technology; 2011 [cited 2019 Nov 20]. Available from: https://dspace.mit.edu/handle/1721.1/67190

15. US Department of Transportation. Manual on Uniform Traffic Control Devices, Part 3: Markings. US Department of Transportation; 2000.

16. OpenCV Dev Team. Miscellaneous Image Transformations — OpenCV 2.4.13.7 documentation [Internet]. OpenCV 2.4.13.7 documentation. [cited 2020 Jul 27]. Available from: https://docs.opencv.org/2.4/modules/imgproc/doc/miscellaneous_transformations.html#cvtcolor

17. Dï B, Mï H, Gross H. Handbook of Optical Systems, Volume 5: Metrology of Optical Components and Systems. John Wiley & Sons; 2012. 1005 p.

18. Hunt RWG. The Reproduction of Colour. John Wiley & Sons; 2005. 727 p.

19. Poynton C. Luminance, luma, and the migration to DTV [Internet]. Lecture presented at: 32nd SMPTE Advanced Motion Imaging Conference; 1998 Feb 6 [cited 2020 Jul 27]; Toronto, CN. Available from: http://poynton.ca/papers/SMPTE_98_YYZ_Luma/index.html

20. Brewer C. Color use guidelines for data representation. In: Proceedings of the Section on Statistical Graphics, American Statistical Association. Alexandria, VA; 1999.

21. Cao XY, Liu HF. A Skin Detection Algorithm Based on Bayes Decision in the YCbCr Color Space. Applied Mechanics and Materials; Zurich. 2011 Oct;121–126:672.

22. Patil N, Yadahalli RM, Pujari J. Comparison between HSV and YCbCr Color Model Color-Texture based Classification of the Food Grains. Int J Comput Appl. 2011 Jan 1;34.

23. Kolkur S, Kalbande D, Shimpi P, Bapat C, Jatakia J. Human Skin Detection Using RGB, HSV and YCbCr Color Models. Proceedings of the International Conference on Communication and Signal Processing 2016 (ICCASP 2016) [Internet]. 2017 [cited 2020 Jul 27]; Available from: http://arxiv.org/abs/1708.02694

24. Haider K, Sattar Q, Ali A. An Efficient Approach for Sky Detection. 2013;10(4):6.

25. Shen Y, Wang Q. Sky Region Detection in a Single Image for Autonomous Ground Robot Navigation. International Journal of Advanced Robotic Systems. 2013 Oct 1;10(10):362.

26. Stewart J. Tesla's "Navigate on Autopilot" Changes Lanes—With the Human's Help | WIRED. Wired [Internet]. 2018 Nov [cited 2020 Aug 1]; Available from: https://www.wired.com/story/tesla-navigate-on-autopilot/

27. Tesla. Autopilot and Full Self-Driving Capability [Internet]. Support-Autopilot. 2019 [cited 2019 Oct 23]. Available from: https://www.tesla.com/support/autopilot

28. Tesla. Model S Owner's Manual [Internet]. Tesla; 2019. Available from: https://www.tesla.com/sites/default/files/model_s_owners_manual_north_america_en_us.pdf

29. General Motors. 2019 Cadillac CT6 Owner's Manual [Internet]. Cadillac; 2019. Available from: https://www.cadillac.com/content/dam/cadillac/na/us/english/index/ownership/technology/supercruise/pdfs/2019-cad-ct6-owners-manual.pdf

30. Flodström K, Strömberg E. Vulnerable Road User Detection System for City Buses [Internet]. 2011 [cited 2020 Jul 29]. Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:ltu:diva-57154

Appendix

Appendix A: Environmental Conditions by Test Day

Test	Start time (hour)	Temperature (F)	Wind speed (mph)	Wind origin	Precipitation	Visibility (miles)	Pressure (in)	Humidity (%)	Reported sky cover
Highway 1	1400	67	20	SW	None	10	30	42	Cloudy
Highway 2	1300	71	6	N	None	10	29.6	49	Cloudy
Track 1	1600	63	6	N	None	10	30.3	45	Fair
Track 2	1300	63	11	SW	None	10	30.2	69	Mostly Cloudy
Highway 3	1200	59	8	E	None	10	30	65	Cloudy
Track 3	1300	84	18	SW	None	10	30.1	47	Fair

Table 4: Atmospheric conditions by test day

Appendix B: Vehicle Software Configurations by Test Day

Vehicles were standardized with respect to the driver customization preferences to the greatest extent possible; because vehicles were rented from private owners, certain settings were subject to security restrictions and could not be manipulated. While the entire set of vehicle customizations is not described here, the autopilot configuration was set as follows:

- Cruise follow distance: 4
- Autosteer: On
- Navigate on Autopilot: Off
- Full-self-driving visualization: Off
- Summon: On
- Speed limit warning: Display
- Speed limit mode: Relative
- Speed limit offset: 0
- Forward collision warning: Medium
- Lane departure warning: Assist
- Emergency lane departure avoidance: On
- Blind spot collision warning chime: On
- Automatic emergency braking: On
- Obstacle-aware acceleration: On

Because Tesla transmits over-the-air updates to vehicles, vehicle software was updated over the course of the test battery, and vehicles did not necessarily use the same software version as each other or across different days of testing. Table *5* shows the Autopilot and Navigation software versions installed on the vehicle during each test.

Vehicle	Test type	Software	Navigation data
Highway 1	Highway	v10.2 (2020.4.1 4a4ad401858f)	NA-2019.20-10487
Highway 2	Track	v10.2 (2020.4.1 4a4ad401858f)	NA-2019.20-10487
Track 1	Highway	v10.2 (2020.4.1 4a4ad401858f)	NA-2019.20-10487
Track 2	Track	v10.2 (2020.12 4fbcc4b942a8)	NA-2019.20-10487
Highway 3	Highway	v10.2 (2020.12 4fbcc4b942a8)	NA-2019.20-10487
Track 3	Track	v10.2 (2020.8.1 ae1963092ff8)	NA-2019.20-10487

Table 5: Software configurations for each test

Appendix C: Descriptive Statistics by Test

Appendix C.1: Highway (HW) Test

Car	Count	Mean	Min	Max	Median	SD
Car 1	62	0.92	0.64	3.44	0.84	0.349
Car 2	16	1.28	0.72	3.64	0.96	0.837
Car 3	64	0.91	0.60	2.04	0.84	0.287

Table 6: Descriptive statistics for time of driver response (seconds) to alerts in HW test.

Car	Count	Mean	Min	Max	Median	SD
Car 1	62	32.3	31.0	43.4	31.7	2.00
Car 2	17	30.2	11.0	32.5	31.2	4.97
Car 3	64	33.0	30.9	43.8	31.8	3.26

Table 7: Descriptive statistics for duration in seconds of interval of automated driving between alerts in HW test.

Car	Count	Mean	Min	Max	Median	SD
Car 1	62	1.24	0.40	5.88	0.64	1.22
Car 2	16	0.69	0.44	1.72	0.60	0.30
Car 3	64	1.38	0.40	10.40	0.58	1.66

Table 8: Descriptive statistics for duration (seconds) of hand contact required to remove an alert during event cycles in HW test.

Appendix C.2: Lane Shift (LS) Test

Car	Count	Mean	Min	Max	Median	SD
Car 1	6	6.6	6.3	6.9	6.6	0.237
Car 2	3	6.3	5.8	6.6	6.5	0.436
Car 3	7	5.67	5.5	5.8	5.7	0.111

Table 9: Number of cones visible at time alarm occurred for trials where an alarm was presented in LS test.

Appendix C.3: Lane Departure (LD) Test

Car	Count	Mean	Min	Max	Median	SD
Car 1	10	4.36	2.88	6.03	4.21	1.16
Car 2	10	4.07	2.05	7.92	4.01	1.62
Car 3	8	4.05	0.68	6.52	4.31	1.55

Table 10: Descriptive statistics for angle of wheel rotation used to initiate nudge in LD test.

Car	Count	Mean	Min
Car 1	3	3	4
Car 2	6	3	1
Car 3	5	1	2

Table 11: Count of trial outcomes by car in the LD test.

Appendix C.4: Curve (CRV) Test

Car	Count	Mean	Min	Max	Median	SD
Car 1	10	9.98	9.88	10.00	10.00	0.051
Car 2	9	9.99	9.96	10.00	9.96	0.040
Car 3	10	10.00	9.92	10.40	9.96	0.153

Table 12: Descriptive statistics for time from first alarm to second alarm in CRV test.

Car	Count	Mean	Min	Max	Median	SD
Car 1	10	5.02	4.96	5.08	5.02	0.034
Car 2	10	5.54	4.96	9.84	5.08	1.510
Car 3	10	5.03	4.60	5.16	5.06	0.162

Table 13: Descriptive statistics for time from second alarm to third alarm in CRV test.

Car	Count	Mean	Min	Max	Median	SD
Car 1	10	15.00	14.90	15.10	15.00	0.053
Car 2	10	14.50	9.84	15.10	15.00	1.650
Car 3	10	15.00	14.90	15.20	15.00	0.084

Table 14: Descriptive statistics for duration of overall alarm sequence in CRV test.

Car	Count	Mean	Min	Max	Median	SD
Car 1	10	736.9	628.3	1252.5	680.7	185.7
Car 2	10	686.0	43.1	1254.6	710.5	596.5
Car 3	10	895.7	609.8	1246.4	692.0	300.6

Table 15: Descriptive statistics for distance from first curve traveled by each car at moment of alert.

Car	Count	Mean	Min	Max	Median	SD
Cluster 1 (earliest)	5	121.1	43.1	172.5	117.0	53.0
Car 2 (middle)	15	673.2	609.8	767.9	673.5	38.8
Car 3 (latest)	10	1248.2	1240.2	1254.6	1248.4	4.6

Table 16: Descriptive statistics for location of each cluster measured as distance from beginning of first curve.

Appendix D: Statistical Models

Appendix D.1: Highway (HW) Test

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Speed	1	897.6	897.6	260.25	< 0.0001
Car	2	38.5	19.2	5.58	0.0047
Residuals	139	479.4	3.4		

Table 17: ANCOVA table for analysis of covariance in duration of automated driving interval during event cycles in HW test.

	Estimate	Standard error	T-value	P-value
Intercept (Car 1)	96.51	4.14	23.28	< 0.0001
Car 2	-1.57	0.51	-3.09	0.0024
Car 3	0.07	0.33	0.22	0.8248
Speed	-0.92	0.06	-15.52	< 0.0001

Table 18: Linear regression for effect of speed and car on duration of automated driving interval during event cycles in HW test.

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Speed	1	5.79	5.79	3.09	0.0808

Car	2	8.10	4.05	2.16	0.1188
Residuals	138	258.29	1.87		

Table 19: ANCOVA table for analysis of covariance in duration of seconds of hand contact required to remove an alert during events in HW test.

Appendix D.2: Lane Shift (LS) Test

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Car	2	3466	1733.1	34.163	< 0.0001
Alert	1	33	32.7	0.644	0.43
Sun azimuth	1	257	256.9	5.064	0.0335
Residuals	25	1268	50.7		

Table 20: ANCOVA comparing luma values observed for trials with or without an alert blocked by car, with sun azimuth as a covariate.

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Cones visible	2	2.88	1.44	25.52	< 0.0001
Residuals	13	0.7343	0.0565		

Table 21: ANOVA table for analysis of variance in quantity of cones visible at moment of alarm in LS test.

Appendix D.3: Lane Departure (LD) Test

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Car	2	1.16	0.58	0.242	0.787
Trial outcome	2	0.59	0.29	0.122	0.886
Residuals	23	55.22	2.40		

Table 22: ANOVA table for analysis of variation in wheel angle rotation in LD test.

	Estimate	Standard error	Z-value	P-value
Intercept (Car 1)	8.97	18.77	0.478	0.633
Car 2	-1.22	0.97	-1.252	0.211
Car 3	-1.20	1.17	-1.028	0.304
Luma	-0.07	0.16	-0.403	0.687

Wheel angle	-0.10	0.29	-0.342	0.732

Table 23: Logistic regression table for effect of car, wheel angle rotation, and luminosity on whether trial had any sort of feedback in LD test.

	Estimate	Standard error	Z-value	P-value
Intercept (Car 1)	-108.52	65.80	-1.649	0.099
Luma	0.94	0.58	1.632	0.103
Wheel angle	-0.39	0.71	-0.551	0.582
Car 2	-0.89	2.12	-0.421	0.674
Car 3	-0.46	2.33	-0.197	0.844

Table 24: Logistic regression table for effect of car, wheel angle rotation, and luminosity on whether trial had assistive steering given that it had an alarm in LD test.

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Trial outcome	2	10.2	5.1	0.037	0.964
Car	2	1838.4	919.2	6.695	0.005
Sun azimuth	1	8.9	8.9	0.065	0.801
Residuals	22	3020.4	137.3		

Table 25: ANCOVA table for effect of car and trial outcome in predicting observed luminosity with sun azimuth as a covariate.

Appendix D.4: Curve (CRV) Test

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Between Car:					
Trial	1	0.001	0.001	0.138	0.774
Residuals	1	0.008	0.008		
Within Car:					
Trial	1	0.000	0.000	0.001	0.973
Residuals	25	0.248	0.010		

Table 26: Repeated measures ANOVA for duration of interval between first and second takeover alert.

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Between Car:					

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Residuals	2	1.748	0.874		
Within Car:					
Trial	1	0.563	0.563	0.721	0.404
Residuals	2	29,287	0.781		

Table 27: Repeated measures ANOVA for duration of interval between second and third takeover alert.

	Degrees of freedom	Sum of squares	Mean square	F-value	P-value
Between Car:					
Residuals	2	1.678	0.839		
Within Car:					
Trial	1	0.706	0.706	0.772	0.388
Residuals	26	23.775	0.914		

Table 28: Repeated measures ANOVA for total duration of takeover alert sequence.

Appendix E: Views of Road at Time of Alert in Lane Shift Test

Figure *20*, Figure *21*, and Figure *22* show the camera viewpoints at the point of alert for Car 1, Car 2, and Car 3 respectively. Note that images are only present for those trials in which an alert occurred, since the alert did not occur in all trials. For Car 1, all seven cones are visible in all trials; their apparent positions range from being somewhat ahead of the car to aligned with the car's front. In all trials, several feet of the second-to-last white lane marking are visible, though the visible length also varies across trials. For Car 2, the distribution of stopping points is similar to Car 1, though there is more apparent variability in the car's position, as can be seen by comparing the visible length of the white line in the upper left image to that in the upper right. For Car 3, the seventh cone is never visible and the white line is not visible in all but one trial.

Trial 1

Trial 2

Figure 20: Test track location at point of alarm for Car 1 in LS test.
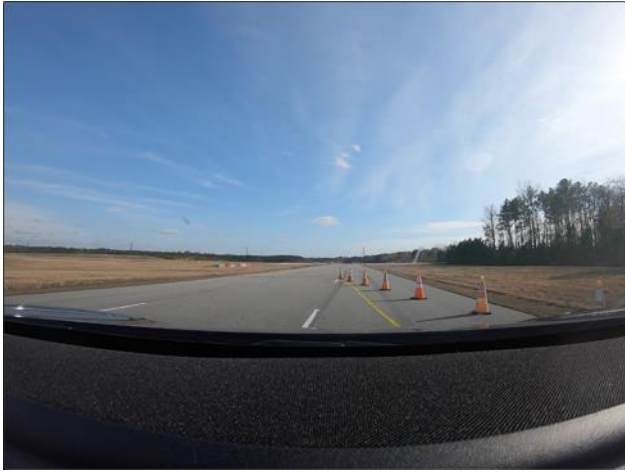
Trial 7	

Figure 21: Test track location at point of alarm for Car 2 in LS test.

Trial 1	Trial 3
Trial 4	Trial 7

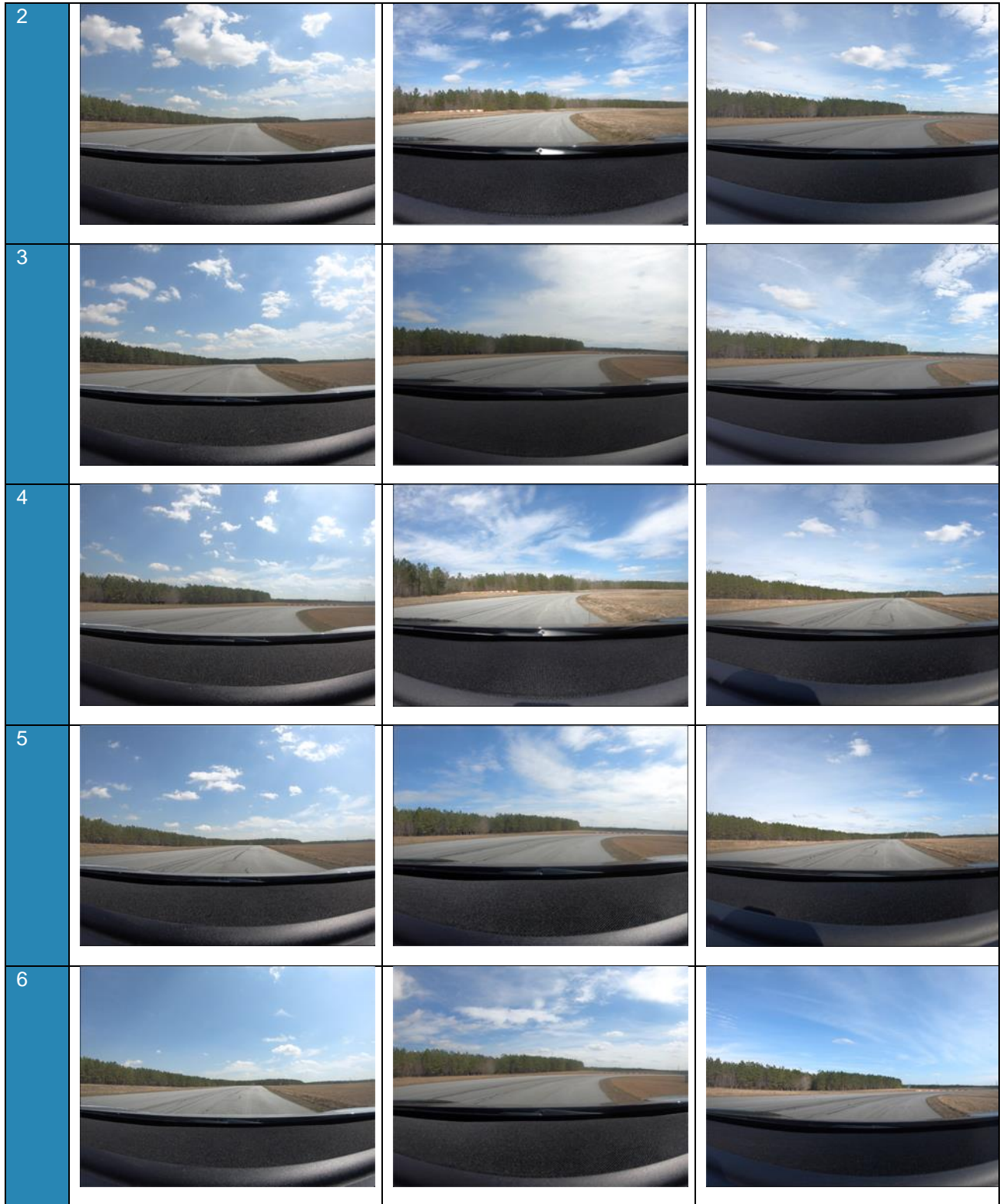| Trial 8	Trial 9
Trial 10	

Figure 22: Test track location at point of alarm for Car 3 in LS test.

Appendix F: Road View at Time of Alert in Curve Test
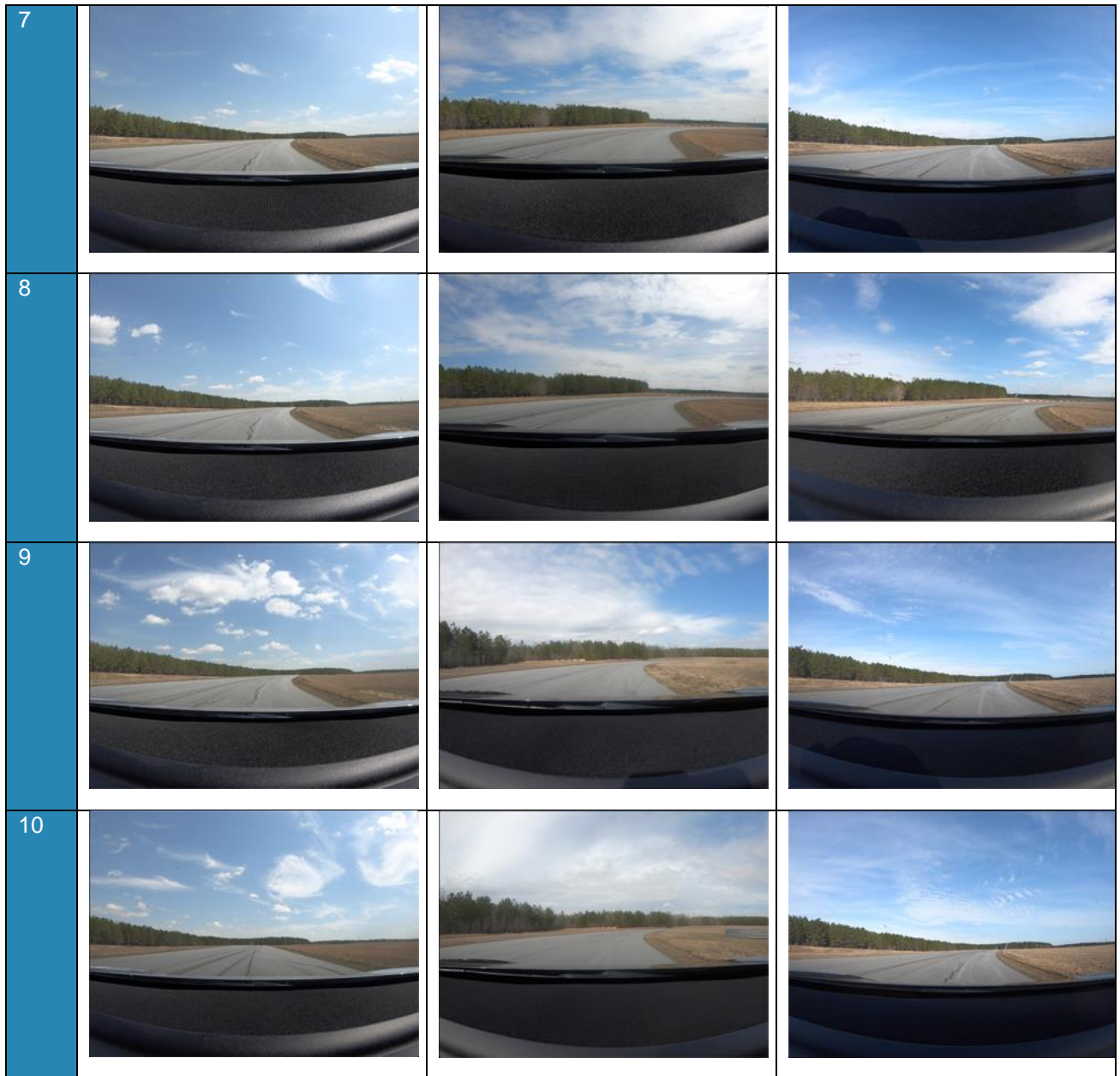
Trial	Car 1	Car 2	Car 3
1			

2			
3			
4			
5			
6			

Figure 23: View of roadway at point of alarm initiation in each trial for each car in CRV test.

Collaborative Sciences Center for
ROAD SAFETY

730 Martin Luther King Jr. Blvd.
Suite 300
Chapel Hill, NC 27599-3430
info@roadsafety.unc.edu

www.roadsafety.unc.edu