Safety Implications of Variability in Autonomous Driving Assist Alerting

Mary L. Cummings, *Senior Member, IEEE*, and Ben Bauchwitz

***Abstract*— Advanced Driving Assist Systems (ADAS) are on the rise in new cars, including versions that embed artificial intelligence in computer vision systems that leverage deep learning algorithms. Because these systems, at the present time, cannot operate in all operational driving domains, they employ some type of driver monitoring system for assessing driver attention, so that drivers can effectively take control if and when an ADAS system can no longer control the car. To determine the reliability of a driver alerting system when linked to autonomy that leverages deep learning, a set of increasingly complex tests were conducted on three Tesla Model 3 vehicles. Tests were conducted on a highway and a closed test track to test road departure and construction zone detection capabilities. Results revealed significant between- and within-vehicle variation on a number of metrics related to driver monitoring, alerting, and safe operation of the underlying autonomy. In some cases, cars performed better than expected but all cars exhibited both inconsistent and unsafe behaviors as well as poor driver alerting. These results highlight that a post-deployment regulatory process is ill-equipped to flag significant issues in vehicles with embedded artificial intelligence.**

***Index Terms*— Autonomous vehicle, deep learning, advanced driving assist, self-driving, driverless, testing, driver monitoring.**

I. Introduction

MORE than 92% of new cars sold in the US include some advanced driving assist feature [1], defined as partial automation or *Level II Autonomy* in the SAE J3016 standard [2]. Such vehicles are equipped with advanced driver-assist systems (ADAS) that include features like automatic emergency braking (AEB), lane departure warning, and blind spot warning. However, some of these cars can perform automated steering and/or acceleration, which many informally call Level II+ (L2+) vehicles. These L2+ systems embed artificial intelligence in the form of machine (aka deep) learning that requires human drivers to be available at all times in case the underlying autonomy fails. While there is a range of definitions of automation versus autonomy in driving domains [3], for the purposes of this paper, an autonomous system is one that leverages probabilistic algorithms in the form of neural networks that make classification estimations of vehicle states.

Several recent high-profile Tesla crashes have highlighted both the brittleness of machine learning-enabled ADAS systems and the debate about if and how much testing such systems should undergo before widespread deployment. In these accidents, the underlying computer vision systems struggled to accurately capture a dynamic world model, and the driver monitoring systems also failed to detect driver disengagements, leading to several fatalities [4]. Because of such problems, the interface between the human and machine is of critical importance in such L2+ vehicles as changes in human attention and behavior with high levels of autonomy make the handover regime particularly dangerous [5]–[7].

L2+ 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 performance standards for the hardware or software in L2+ systems. In addition to the lack of performance standards for known computer vision problems [8], there are no requirements that over-the-air software updates be vetted in any formal way prior to deployment.

The National Highway Traffic Safety Administration's (NHTSA) New Car Assessment Program (NCAP) does not address driver-assist features, limiting its assessment to system functionality like Automated Emergency Braking, Forward Collision Warning, Dynamic Brake Assist, and Lane Departure Warning systems [9]. The Insurance Institute for Highway Safety (IIHS) addresses testing of some ADAS features such as pedestrian detection and automated emergency braking (AEB), but does not address driver monitoring [10]. The Korean Ministry of Land, Infrastructure, and Transport (MOLIT) provides specific guidance regarding how the driver monitoring system should be designed [11].

The European NCAP (aka Euro NCAP) will begin assessing driver monitoring in 2022 [12], and has already been assessing various ADASs on safety backup behaviors (i.e., collision avoidance), as well as Assistance Competence (how well the system/manufacturer explains the system's limits) [13]. Despite increasing interest in the testing of L2+ systems including the driver monitoring component, there is no formal guidance on exactly how to test joint human-autonomous interaction and how to capture variability inherent in systems with embedded AI.

In addition to a lack of formal L2+ ADAS test protocols, there is little-to-no information as to how different software

Manuscript received 25 January 2021; revised 1 May 2021 and 9 July 2021; accepted 17 August 2021. Date of publication 28 December 2021; date of current version 9 August 2022. This work was supported by the U.S. Department of Transportation's University Transportation Center grant through the University of North Carolina's Collaborative Sciences Center for Road Safety. The Associate Editor for this article was M. Brackstone. *(Corresponding author: Mary L. Cummings.)*

The authors are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: m.cummings@duke.edu).

Digital Object Identifier 10.1109/TITS.2021.3109555

versions may affect outcomes, which could be significant given over-the-air updates (OTAs). 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, software configuration or wear-and-tear. For the few L2+ ADAS tests that have been published, only a single vehicle was used with only a handful of trials per test and the full range of system performance was not detailed [14], [15]. Others have conducted naturalistic studies that observe how drivers interact with ADAS systems in various operational domains [16], [17], but such studies cannot leverage statistical inference across a set of controlled trials.

To address this gap, the goal of this effort was to assess L2+ ADAS-equipped vehicle and driver monitoring performance across increasingly complex scenarios including variability within and between cars. These tests were not meant to replicate or validate US NCAP or Euro NCAP testing protocols. These procedures were used as a guide for our testing, which needed to be tailored to our research questions and available resources. This paper first presents background information about the need for principled testing of autonomous vehicles, then the experimental setup and results, and concludes with a discussion of the implications of the findings.

II. Background

The rise of AI-based vehicle technologies, fundamental to both L2+ ADAS and self-driving cars, has also led to increasing debate about safety and if and how regulation should be developed to address these nascent technologies. Proponents of such systems cite possible increases in safety as a benefit, although this is debated [18], as well as increased innovation opportunities [19]. Self-driving technologies are promised to revolutionize accessibility for those who cannot drive and reduce congestion [20], although this is also a hotly debated topic, e.g., [21]. The hopes for such benefits led to the failed federal legislation in the form of the AV START Act, which struggled in part due to the debate about what is needed for adequate testing [22].

While there are many possible benefits, there have also been many problems with L2+ technologies. Studies increasingly show that drivers in general struggle to remain engaged as autonomy increases in driving assist scenarios both in simulation and in real-world settings [23]–[25]. There have been a number of naturalistic studies that indicate people who use L2+ systems spend more time looking in the car when these systems are engaged, and thus not on the road, increasing the risk of a possible accident [26], [27].

As high-profile accidents increase, regulatory bodies and safety advocates are increasing calling for more action. The National Safety Council recently launched the "My car does what?" initiative (mycardoeswhat.org) to ensure that drivers are educated about advanced features on their vehicles. The high-profile Tesla crashes have led to one government agency, the National Transportation Safety Board (NTSB), calling out a regulatory agency (NHTSA) for not doing its job [4]. Advocacy groups have also called for more regulation and oversight, especially after a pedestrian was killed during an Uber self-driving test [28].

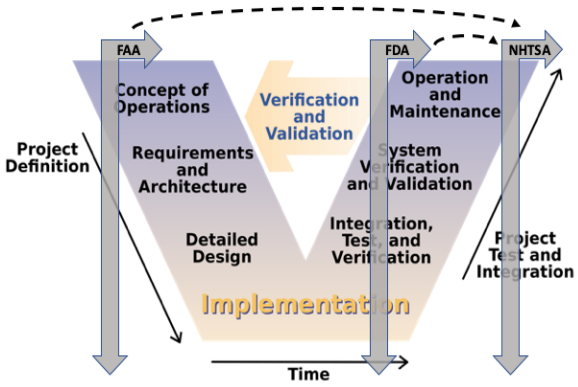

Fig. 1. Regulatory intervention in a systems engineering framework. The dashed lines indicate legal short-cuts based on equivalency.

At the heart of this debate is a lack of data and understanding about how well such L2+ vehicles perform the task of vehicle control as well as the task of maintaining driver engagement. The recent crash of a Tesla that resulted in the death of two people because neither was in the driver's seat highlights the importance of linking vehicle control to effective driver monitoring [29]. Moreover, if autonomous vehicles could safely control themselves in all operational domains, then driver monitoring would not be needed. However, given that self-driving vehicles are in the experimental stage, driver monitoring is still very much a key consideration for autonomous vehicles.

There are no requirements, either at the state or federal levels, for proof of safe operation of L2+ systems, including driver monitoring. While other safety-critical systems like airplanes and medical devices have such requirements with significantly more regulatory oversight, L2+ systems are not regulated because, in theory, these systems are *assistive* and not required for systems to operate. For such assistive systems, NHTSA, the regulatory agency that oversees vehicle safety, will not intervene until it determines that some sort of defect exists that poses an unreasonable safety risk, which then leads to a recall.

Figure 1 illustrates the timing of US regulatory agency involvement for safety-critical systems, against the backdrop of the systems engineering V lifecycle diagram [30], [31]. On the horizontal axis of the V diagram is time, which indicates temporal ordering that can include some overlap for activities. The vertical axis represents stages of the systems engineering process. The left side of the V represents system ideation and design, while the right side generally represents testing and implementation.

The Federal Aviation Administration (FAA) interacts with aviation companies very early in the product development lifecycle, while the Federal Drug Administration (FDA) intervenes only in late-stage testing. NHTSA does not regulate technologies until *after* deployment in the form of recalls, assuming basic design standards are met. The FDA and FAA both have shortcuts in their regulatory processes that are akin to NHTSA's approach, as indicated by the dashed lines in Fig. 1. The short cuts are based on establishing equivalency

between an existing and the new technology. However, such shortcuts have been shown to have disastrous consequences when advanced autonomy is introduced, which contributed to the 2018 and 2019 Boeing 737MAX crashes, as well as incidents with robotic surgical devices [31].

This post-deployment regulatory approach may have been the appropriate approach for ensuring that deterministic technologies provably perform safe and consistently in well-established test environments. However, AI-based L2+ systems with embedded deep learning algorithms that underpin computer vision systems have no such guarantees, and can produce dramatically different outcomes with seemingly the same inputs [32]. Currently, there are no formal methods in controlled settings of actual systems that help engineers or regulators determine the scope of such anomalies in computer vision systems as well as critical vulnerabilities and areas of high risk.

To this end, this paper describes a series of experiments that investigated variability in L2+ driver monitoring alerting in order to reveal potential AI-enabled computer vision and driver monitoring problems within and across multiple vehicles of the same make and model. Such results could inform the debate about whether post-deployment regulatory action is appropriate or whether a more conservative approach is needed.

III. Methods

Three increasingly complex scenarios were presented to three Tesla Model 3s. From lowest to highest complexity, these scenarios were: (1) Assessing driver-monitoring system performance during highway driving; (2) Assessing driver alerting in response to an inadvertent road departure; and (3) Assessing driver alerting of obstacles and a lane shift during autonomous driving in a construction zone. These three tests are labelled the highway, road departure and construction zone tests, respectively. These three tests were designed to test the performance of the driver monitoring systems and the underlying autonomy, not a specific driver's response.

For this effort, complexity is defined by the degree to which the ADAS system must detect anomalous driver and environmental states. In scenario 1, the system only has to detect an anomalous driver state (low complexity), whereas in scenario 3, the system has to detect a distracted driver, an unexpected road shift and obstacles (high complexity). Given that Tesla models have L2+ ADAS systems called Autopilot that can be used on interstates, divided highways and urban and rural roads, and thus can face a range of potentially demanding environments, they were the only test platform available that could perform across the range of defined scenarios.

Three 2018 Tesla Model 3s from the Triangle metropolitan area of North Carolina were randomly tested over a period of two weeks during March 2020. While the three cars were the same make and model, they all had different software packages, which was an uncontrollable confound. In addition, car 2 had the full self-driving version of Autopilot.

All tests were conducted during daylight, between 12:00pm and 5:00pm, under similar environmental conditions. 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.

TABLE I
Testing Summary

Test	Environment	Speed	System Assessed
Highway	Public highway	70 mph	Alerting when driver's hands not detected
Road Departure	Test track	35 mph	Alerting & steering assistance with driver's hands not detected
Construction Zone	Test track	25 mph	Alerting & steering assistance with driver's hands not detected

Tests were either performed on a public highway or at the North Carolina Center for Automotive Research (NCCAR), a closed test track facility as outlined in Table I. 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 [33].

A. Highway Test Experimental Setup

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 Tesla's 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 [4].

This test was conducted on two 5.2-mile sections of Interstate 540 in the Triangle area. 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.

Because these sections had between 3 and 5 lanes at various points, each car was driven in Autopilot in the third lane from the left, which allowed it to be driven without the need to change lanes. 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.

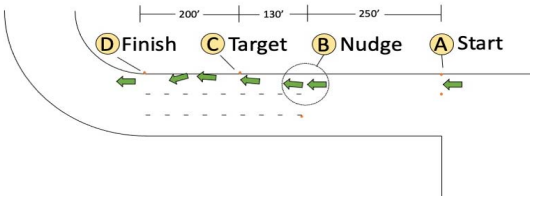

Fig. 2. Road departure test setup.

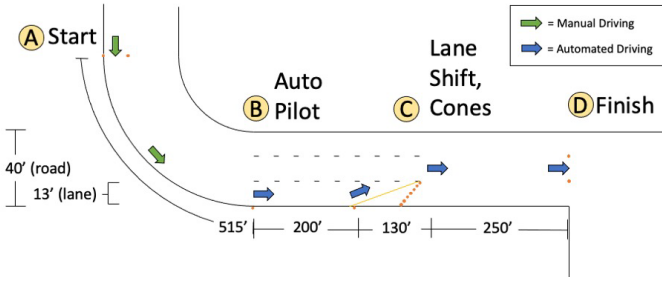

Fig. 3. Construction zone test setup.

The alert consists of a message on the car's 15 in., 1920 × 1080 pixel display center mounted on the dashboard that says "Apply slight turning force to steering wheel" and is accompanied by a quick pair of beeps. Tesla vehicles recognize that 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 at which time Autopilot was disengaged. The same protocol was used for both directions of the driving route.

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.

B. Road Departure Test Experimental Setup

The goal of the second 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 off the road while in the adaptive cruise control mode, but without the automated steering provided when Autopilot is engaged. Tesla advertises that its vehicles are equipped with emergency lane departure avoidance, meaning the car should provide evasive automated steering to prevent the vehicle from exiting the lane or departing the roadway. Therefore, the hypothesis was that all vehicles would provide alerting and emergency assistive steering as the cars drifted off the road's edge.

This test was conducted on a straight section of track at the NCCAR facility. For this test, each vehicle began between two traffic cones at the position marked 'start' in Fig. 2. Starting from an inactive, parked state, the vehicle was 250' driven towards painted lane lines in the inner most lane, as seen in Fig. 2. There was no white line at the road's edge, which was bordered by approximately one foot of low-cut grass, with higher-cut grass beyond that edge. After accelerating to 35 mph, the car was immediately placed in Adaptive Cruise Control. Autopilot was not initiated.

Upon passing the cone marking the beginning of the painted section of track (point B in Fig. 2), 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 (point C in Fig. 2). The car was allowed to move in that direction with no steering input until either the car left the road or the lane-keep assist feature activated, steering the car back onto the road. The trial was concluded as soon as the vehicle passed the final set of white lane markings.

While this test procedure was similar to both the Euro NCAP Emergency Lane Keeping Road Edge Test [34] and NHTSA lane keeping tests [35], due to track limitations and for additional safety precautions, the 45 mph testing speed was reduced to 35 mph. The targeted lateral velocity range was between.4 and.6 m/s, which is commensurate with the upper ends of the test protocols. In addition, the point of this test was to examine the interaction and reliability between the emergency road assist system and the driver alerting system, which does not have an established test protocol.

C. Construction Zone Test Experimental Setup

The goal of the last 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 detecting this environment and would steer to avoid obstacles.

For this test, the vehicle began at the position marked 'start' in Fig. 3. The car was driven manually along a 515-foot curved section of track and accelerated to 25 mph. At the conclusion of the curve (point B in Fig. 3), 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 at 25 mph. After 200 feet, a solid yellow line marked a lane shift in which the right-hand lane merged into

Fig. 4. Video camera setup. Road camera on left, console camera on upper right and driver camera on lower right.

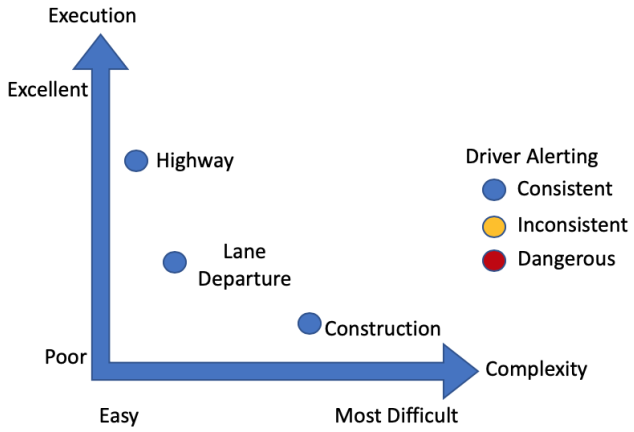

Fig. 5. L2+ technology test hypotheses.

the central lane. Such markings are not uncommon in North Carolina. 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 (point C in Fig. 3). 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.

D. Data Collection

One objective of this effort was to develop a test protocol and data collection system that did not rely on access to proprietary data and could easily be moved between cars. To this end, video data were collected using three GoPro Hero 7 Black cameras synchronized with SyncBac Pro devices and mounted at fixed positions in the vehicle interior. These cameras obtained views of the roadway, the driver, and the center console (Fig. 4).

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). It 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.

The road-facing camera 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 faced 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. All cameras were set to 1440 pixels per inch resolution, 25 frames per second, wide field of view, and automatic stabilization, with protune off.

E. Hypotheses

Figure 5 illustrates our expectation for the outcomes of the three tests as a function of car execution, complexity and driver alerting vis-à-vis the color of the circle. The highway test focused on whether the driver monitoring system worked as advertised in the operational domain for which Autopilot is optimized. Thus, the cars were expected to perform well in this low complexity setting with no directional changes.

The lane departure test was somewhat more complex as the car's vision system had to detect a slight drift that would send the car off the road with no driver response. Because the road edge did not have a white line for the vision system to detect, we expected reduced Autopilot performance, although the car is advertised to be able to detect a road's edge, regardless of the presence of a line.

The construction zone test was the most difficult, requiring a lane shift and obstacle detection, all on Autopilot. Teslas do not have LIDARs (light detection and ranging) and so the vision system is primarily responsible for obstacle detection. However, these vision systems can have difficulties in such settings, which is why Autopilot is not supposed to be used in such settings (but often is). Therefore, we did not expect the cars to perform well on this test.

Lastly, in all three testing scenarios, we expected the cars to successfully and consistently alert the driver about potential hazards, whether that be a hands-off condition, a failure of the Autopilot or an anomalous condition such as the presence of obstacles. The next section details the results from these tests.

IV. Results

All statistical alphas are.05 unless otherwise stated and effect sizes for analyses of variance are reported as eta squared values.

A. The 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: (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).

TABLE II
EVENT CYCLE COUNTS FOR THE HIGHWAY TEST

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

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. Over the course of the 5.2 mi course, a typical run would include 7-8 such cycles. Each car experienced ten runs in a randomized and counterbalanced fashion.

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 a vehicle veering off the road.

Table II summarizes the counts of the event cycle outcomes observed for each car. Frequencies of the outcomes were assessed using a chi-squared independence test, and the distributions were determined to be significantly different across cars ($\chi 2 = 52.703$, $p < 0.0001$). While the occurrences of shutoffs were low (3.4% of total trials), they were the most dangerous as it was not immediately obvious to the driver that Autopilot was no longer engaged.

Car 2 was the only car to experience failure (30% of its trials), which resulted in fewer observed total event cycles. It should be noted that this was the car with the full self-driving package. 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.

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 [4], 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. The mean duration of this interval was just over 30 seconds for all three cars (Car 1 mean (M) = 32.3s, standard deviation (SD) = 2.0s; Car 2 M = 30.2 s, SD = 5.0s; Car 3 M = 33.0s, SD = 3.3s), which was slightly longer than expected due to a few instances of slower-moving lead vehicles.

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 ($F(1,139) = 260.3$, $p < 0.0001$, effect size = 0.64) and car ($F(2,139) = 5.58$, $p = 0.005$. effect size = 0.07). A Tukey-Kramer post-hoc test with Bonferroni-adjusted significance level of 0.02 revealed that there was a significant difference in the main effect between cars 2 and car 3 ($p = 0.003$).

These results mean that, as suspected, the average speed for each car was different for the tests. Controlling for this difference led to a statistical, but small difference in hands-free duration between cars 2 and 3. Given this difference, the average alert duration was slightly longer than that published in the Tesla's owner's manual

The times from when hands first touched the steering wheel to the time when the alert disappeared from the console 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. An ANCOVA model with average speed as a covariate did not detect a significant effect for either car ($F(2,138) = 2.164$, $p = 0.1188$, effect size = 0.02) or speed ($F(1,138) = 3.094$, $p = 0.0808$, effect size = 0.03). Thus, there were no differences across or within cars in terms of how long it took a driver to clear an alert.

B. Lane Departure Test

The goal of this test was to evaluate between- and within-car variation in the application of emergency assistive steering if the car drifted towards the edge of the road during automated cruise control driving. To execute this, each car drove 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 (Fig. 2).

Because all cars were configured to provide emergency assistive steering, emergency assistive steering should have engaged in all trials. This was repeated ten times in a randomized and counterbalanced fashion for each of the three cars. Two trials were discarded because the driver's nudge did not result in a trajectory that took the car outside the lane.

Counts of the three different outcomes across the twenty-eight trials in Fig. 6 include emergency assistive steering in conjunction with an alarm, an alarm only, and neither alarm nor steering. 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 ($\chi 2 = 3.4375$, $p = 0.4874$). 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. The locations of the triggered alerts occurred between points B and C in Fig. 2.

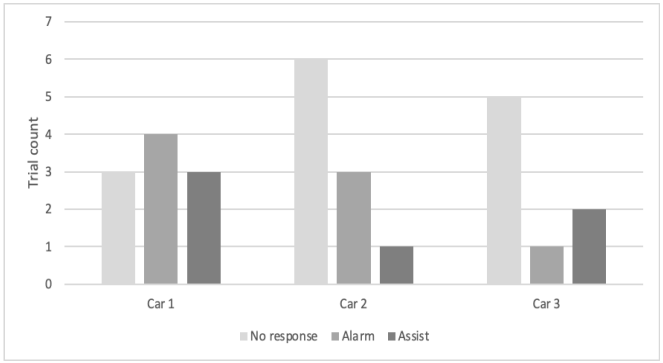

Fig. 6. Road departure test outcomes. The assist category includes both steering and an alarm.

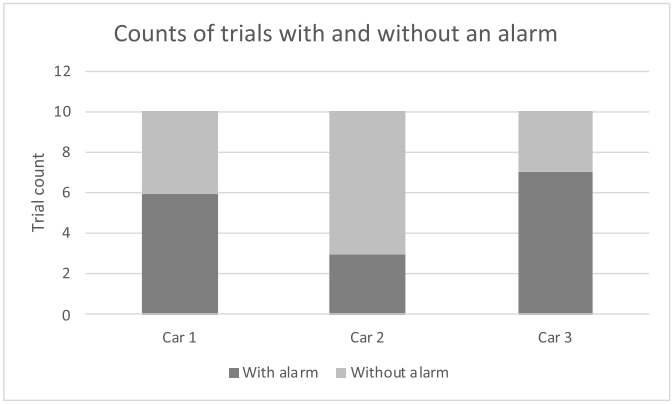

Fig. 7. Construction zone test alerting outcomes.

Their exact occurrence was dependent on the driver's angular input and there were no statistical correlations between where they occurred and either the car or the kind of alert.

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.

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 ($F(2,23) = 0.242$, $p = 0.7870$, effect size =.02) or the trial outcome ($F(2,23) = 0.122$, $p = 0.8860$, effect size =.02), indicating that variation in wheel rotation was not systematically different between cars and not correlated with particular trial outcomes.

Our intended lateral velocity range was between.4-.6 m/s which resulted in an observed mean of.5 m/s, SD =.2 m/s. Using a regression model, there were no statistical correlations between observed lateral velocities and the individual cars or the presence of alerts and assistance.

C. Construction Zone Test

The goal of this test was to determine within- and between-vehicle variability when encountering an unexpected lane shift and obstacles, in this case a simulated construction zone (Fig. 3). Each car drove this course ten times, also randomized and counterbalanced. Whether vehicles presented a driver takeover alert was assessed, as well as at 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, yielding a significant difference chi square between cars ($\chi 2 = 130.02$, $p < 0.0001$).

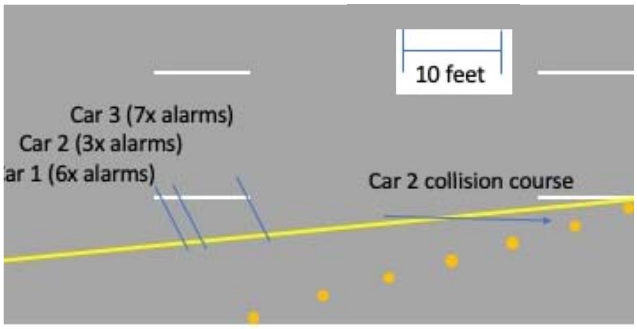

Fig. 8. Location of alerts in the construction zone.

Separate from the 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 if Cars 1 and 3 guided on the cones or the yellow line, whether an alert was generated indicates that a car detected the obstacles. Cars 1, 2, and 3 had 6, 3, and 7 trials in which an alarm was presented, respectively (Fig. 7). Differences in the counts of each observation for each car were analyzed using a chi-squared independence test, with no significant difference between cars ($\chi 2 = 3.4821$, $p = 0.1753$). Overall, the driver was not alerted in 50% of trials where the car encountered the construction zone. If Car 2 is disregarded, this rate is 35%.

Data from the forward-facing camera was used to estimate the location at which the alarm was sounded by computing the area of traffic cone visible, to the nearest 10% of a cone (Fig. 8). 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 ($F(2,13) = 25.52$, $p < 0.0001$, effect size =.75). A Tukey-Kramer post-hoc test with Bonferroni adjusted significance threshold of 0.02 revealed a significant difference between cars 2 and 3 ($p = 0.006$) as well as between cars 1 and 3 ($p < 0.0001$).

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. Figure 8 also depicts Car 2's approximate trajectory toward the traffic cones, as this vehicle failed to avoid all of the obstacles. However, even though it failed to steer the car away during

TABLE III
SUMMARY OF RESULTS ACROSS TESTS

Test	Metric	Car 1	Car 2	Car 3
Highway	Alert interval	C2	C1	I
	Time to clear	CA	CA	CA
	Unsafe behavior	2%	35%	5%
Lane Departure	Alert sounded	I	I	I
	Steering Assist	I	I	I
	Unsafe behavior	30%	60%	63%
Construction Zone	Sounding of alert	I	I	I
	Location sounded	CW	CW	CW
	Unsafe behavior	0%	100%	0%

CA = Consistent All, C(1,2,3) = Consistent with Car 1, 2, or 3, CW = Consistent Within a single car, I = Inconsistent

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 another. Car 3 progressed furthest through the construction site before presenting an alarm, approximately 10 ft beyond where Cars 1 and 2 sounded their alarms.

V. DISCUSSION

The goal of this study was to examine whether there were significant between- and within-vehicles driver-alerting differences in three randomly-selected 2018 Model 3 Tesla vehicles. Table III summarizes the general levels of consistency of each vehicle platform across the three driving tests. As will be discussed in detail, the bulk of tests yielded dramatic inconsistencies both within a single vehicle as well as across all three 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. Car 3 was 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. 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. On the construction test, Car 2 failed to detect and maneuver around the obstacles in any trial, and was also least likely to initiate a takeover alert.

Car 2's behavior during highway testing was also 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. Curiously, this pinball behavior diminished somewhat 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.

Other abnormalities were noted including Car 2 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. Future research efforts should investigate whether there is an 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 the interval of hands-free driving prior to an alert in the highway task and the location of takeover alert 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.

These tests raise significant issues with L2+ systems and safety. 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. In the lane departure test, 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 at least acknowledged the imminent lane departure but could not provide steering.

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. Indeed, even though only 3.4% of events were unexpected disengagements, if a similar proportion of disengagements occurs across the fleet of Teslas, this means millions of disengagements could be happening, which may not be noticed by drivers.

When the results in Table III are compared with the original hypotheses in Fig. 5 in terms of execution, complexity and alerting performance, the clear pattern of unpredictable variation both between and within cars emerges. The only tests that met our expectations were the highway tests for cars 1 and 3. The road departure execution results were not surprising, but the lack of warning for half the road departures is concerning.

Cars 1 and 3 performed better than expected in the construction zone tests, suggesting that camera vision systems are improving. However, such progress is overshadowed by the fact that one of the cars failed to respond to either the cones or lane shift and if this occurred in the real world, would have likely led to harm to both the driver and the personnel working in such a zone.

Another important finding from Table III is the inconsistent and lack of alerting provided to the driver across a number of scenarios. Tesla is very clear in every owner's manual that drivers are responsible for safe operation but they are not consistently informed of important information regarding the state and capabilities of Autopilot. This fact, coupled with the research showing people have difficulty maintaining sustained attention in such settings, means that more work is needed to develop predictable and accurate alerting systems.

The high consistency of alerts in the highway tests for Cars 1 and 3 seems encouraging, however the small percentages of unexplained autopilot shutoffs in both cars raises the possibility that resulting complacency and mode confusion could lead to real-world transportation safety problems. Complacency is a growing problem in cars with L2+ systems [36], with drivers developing poor monitoring habits and extending periods of distraction. Even though there was a small percentage of unexpected autopilot disengagements (3.4%), complacent drivers can easily miss such an event that does not generate a salient warning.

Complacency opens the door to mode confusion, which has been documented in vehicles with partial autonomy [37]. Mode confusion occurs when a driver thinks the car is in one mode but is actually in a different mode. In the case of inadvertent autopilot disengagements, already complacent drivers may think the car is capably handling the driving task, so then they engage in a distracting task right after briefly putting their hands back on the steering wheel, not realizing they, not the Autopilot system, are in charge of the driving task. Mode confusion can lead to no one driving the car, which can and has led to fatalities.

In their 2020 review of Tesla Model 3s, Euro NCAP awarded the Tesla Model 3 a 95% rating in safety backup behaviors, but only 36% in driver assistance competence [38]. While the competence findings were similar to ours, none of the Euro NCAP results were based on tests akin to the ones conducted for this study. Given the demonstration by Consumer Reports that these cars can easily be driven with no one in the driver's seat [39], more work is needed to develop effective safety backup and driver monitoring tests in the presence of L2+ systems.

While this research focused on Teslas because they were the only platform that could support such a wide range of tests, these lessons also apply to other manufactures with similar L2+ systems. Honda, GM, Ford and Mercedes have recently started advertising "hands-free" ADAS systems similar to Tesla's Autopilot. The results from these tests show that if the underlying computer vision systems and associated alerting systems are flawed and the driver is hands-free, even small deviations in attention could lead to negative outcomes.

VI. Limitations

Small sample size was a limitation of this study. Choosing three cars of the same make and model from one region does not reflect the diversity in performance that would occur across the fleet of vehicles. However, given that this study was designed to determine if and how much variation existed between and within L2+ ADAS alerting systems, it serves as a baseline against which other future studies could be measured.

Variation in car software was another issue. One potential confound across the three vehicles 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)*.

These upgrades affected primarily non-driving aspects of the vehicle software by increasing the fidelity of the console visualization, improving the user interface for vehicle service monitoring, adding map support for navigating to non-Tesla sponsored charging stations, adjusting the default settings for Bluetooth device connection, increasing the versatility of voice-based interaction with the vehicle, and increasing the number of languages supported in console documentation. Additionally, both updates made slight modifications to the regenerative braking software, which was the only driving-related feature affected. No aspect of Autopilot was impacted by either upgrade.

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.

These tests were also only conducted in daylight, under ideal weather conditions and not in the presence of other vehicles. Several of the high-profile Tesla crashes as well as the Uber pedestrian death happened in darkness, so additional testing is needed to determine how increased complexity in the environment. Current tests are underway to determine the impact of low sun angle in similar conditions.

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 or the impacts of over-the-air software updates, so significant between-vehicles differences are not currently accounted for regardless of their source. Significantly more research is needed in determining if and how over-the-air updates should be regulated, particularly if they affect safety-critical functions.

VII. Conclusion

The goal of this research was to assess between- and within vehicle variation for an L2+ system, including driver monitoring, in three key scenarios. To this end, three Tesla Model 3 vehicles displayed significant between- and within-vehicle variation on a number of metrics related to driver monitoring, alerting, and safe operation of the underlying autonomy.

These results suggest that the performance of the underlying artificial intelligence and computer vision systems was extremely variable, and this variation was likely responsible for many of the delays in alerting a driver whose hands were not on the steering wheel. Ironically, in some cases the cars seemed to perform the best in the most challenging driving scenarios (navigating a construction zone), but performed worse on seemingly simpler scenarios like detecting a road departure.

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 cones were more easily detected in one software version as opposed to the road edges in a much more gradual drift in the road departure test. Another possibility is that 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 is allowed to operate without humans in direct control. These results also suggest that more effort is need on developing consistent and accurate alerts when L2+ systems are not performing as expected.

These results should be interpreted in light of 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.

Lastly, these results highlight that the post-deployment regulatory process that NHTSA uses in Fig. 1 to protect the public against unsafe vehicle technologies is ill-equipped to flag significant issues with L2+, or in the future, self-driving cars. These results dramatically illustrate that testing a single car, or even a single version of deployed software, is not likely to reveal serious deficiencies. Waiting until after new autonomous software has been deployed find flaws can be deadly and can be avoided by adaptable regulatory processes. The recent series of fatal Tesla crashes underscores this issue. It may be that any transportation system (or any safety-critical system) with embedded artificial intelligence should undergo a much more stringent certification process across numerous platforms and software versions before it should be released for widespread deployment. To this end, our current derivative efforts are focused on developing risk models based on such results.

Acknowledgment

At the time of this research, Prof. Cummings was a member of the Veoneer Board of Directors, a Tier 1 automotive supplier. The authors aver that these research results and findings are free from any commercialism and are totally objective findings, regardless of any affiliations. Lastly, they were assisted by Sam True, the Director of the North Carolina Center for Automotive Research, and Matthew Seong, Kausthub Ramachandran, and Vishwa Alaparthy in the collection of the data. The anonymous reviewers also provided very helpful feedback and recommendations.

References

[1] *Advanced Driver Assistance Technology Names*, AAA, American Automobile Association, Heathrow, FL, USA, 2019.

[2] *Taxonomy and Definitions for Terms Related to Driving Automation Systems for on-Road Motor Vehicles*, Standard SAE, J3016_201806, 2018.

[3] S. E. Shladover, "Connected and automated vehicle systems: Introduction and overview," *J. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 190–200, 2018.

[4] *Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator Mountain View, California March 23, 2018*, document NTSB/HAR-20/01, NTSB, Nat. Transp. Saf. Board, Washington, DC, USA, 2020.

[5] E. E. Miller and L. N. Boyle, "Adaptations in attention allocation: Implications for takeover in an automated vehicle," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 66, pp. 101–110, Oct. 2019.

[6] F. M. Favarò, P. Seewald, M. Scholtes, and S. Eurich, "Quality of control takeover following disengagements in semi-automated vehicles," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 64, pp. 196–212, Jul. 2019.

[7] H. J. Kim and J. H. Yang, "Takeover requests in simulated partially autonomous vehicles considering human factors," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 5, pp. 735–740, Oct. 2017.

[8] M. L. Cummings, "Adaptation of licensing examinations to the certification of autonomous systems," in *Safe, Autonomous and Intelligent Vehicles* (Unmanned System Technologies), H. Yu, X. Li, R. Murray, S. Ramesh, and C. J. Tomlin, Eds. Basel, Switzerland: Springer, 2019, pp. 145–162.

[9] *Test Procedures*, Vehicle Manufacturers, Department of Transportation, NHTSA, Washington, DC, USA, 2021.

[10] *Test Protocols and Technical Information*, IIHS, Insurance Institute for Highway Safety, Arlington, VA, USA, 2020.

[11] J. Jung, *Korean Government Announces Safety Standards for Level 3 Automated Vehicles*. Heilbronn, Germany: MOLIT, 2020.

[12] *Assessment Protocol—Safety Assist*, Euro NCAP, European New Car Assessment Programme, Leuven, Belgium, 2020.

[13] Euro NCAP. (Apr. 30, 2021). *What's New for 2020*. [Online]. Available: https://www.euroncap.com/en/vehicle-safety/safety-campaigns/2020-assisted-driving-tests/whats-new/
[14] *Active Driving Assistance Systems: Test Results and Design Recommendations*, Consumer Reports, Yonkers, NY, USA, 2020.
[15] *IIHS Examines Driver Assistance Features in Road, Track Tests*, IIHS, Insurance Institute for Highway Safety, Arlington, VA, USA, 2018.
[16] I. J. Reagan *et al.*, "Disengagement from driving when using automation during a 4-week field trial," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 82, pp. 400–411, Oct. 2021.
[17] *Evaluation of Active Driving Assistance Systems*, AAA, American Automobile Association, Heathrow, FL, USA, 2020.
[18] B. Templeton. (Jan. 4, 2020). *New Tesla Autopilot Statistics Show it's Almost as Safe Driving With it as Without*. [Online]. Available: https://www.forbes.com/sites/bradtempleton/2020/10/28/new-tesla-autopilot-statistics-show-its-almost-as-safe-driving-with-it-as-without/?sh=77ddf3ab1794
[19] *Preparing for the Future of Transportation: Automated Vehicles 3.0*, US DOT, US Department of Transportation, Washington, DC, USA, 2018.
[20] Coalition for Future Mobility. (Jan. 4, 2021). *Highly Automated Technologies, Often Called Self-Driving Cars, Promise a Range of Potential Benefits*. [Online]. Available: https://coalitionforfuturemobility.com/benefits-of-self-driving-vehicles/
[21] T. Cohen and C. Cavoli, "Automated vehicles: Exploring possible consequences of government (non)intervention for congestion and accessibility," *Transp. Rev.*, vol. 39, no. 1, pp. 129–151, Jan. 2019.
[22] B. Canis, *Issues in Autonomous Vehicle Testing and Deployment Updated*, document R45985, Congressional Research Service, Washington, DC, USA, 2020.
[23] IIHS. (Jan. 4, 2021). *Automated Systems Need Stronger Safeguards to Keep Drivers Focused on the Road*. [Online]. Available: https://www.iihs.org/news/detail/automated-systems-need-stronger-safeguards-to-keep-drivers-focused-on-the-road
[24] J. Gaspar and C. Carney, "The effect of partial automation on driver attention: A naturalistic driving study," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 61, no. 8, pp. 1261–1276, Dec. 2019.
[25] J. C. F. de Winter, R. Happee, M. H. Martens, and N. A. Stanton, "Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 27, pp. 196–217, Nov. 2014.
[26] A. Morando, P. Gershon, B. Mehler, and B. Reimer, "Driver-initiated Tesla autopilot disengagements in naturalistic driving," in *Proc. 12th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, Sep. 2020, pp. 57–65.
[27] E. Tivesten, A. Morando, and T. Victor, "The time course of driver visual attention in naturalistic driving with adaptive cruise control and forward collision warning," in *Proc. 4th Int. Conf. Driver Distraction Inattention*, Sydney, NSW, Australia, 2015, pp. 1–14.
[28] *Collision Between Vehicle Controlled by Developmental Automated Driving System and PedestrianTempe, Arizona March 18, 2018*, NTSB, National Transportation Safety Board, Washington, DC, USA, 2019.
[29] R. Elliott, "Congressmen, consumer reports raise concerns over Tesla's autopilot," *Wall Street J.* Accessed: Apr. 22, 2021.
[30] *Systems Engineering Guide*, MITRE Corporation, McLean, VA, USA, 2014.
[31] M. L. Cummings and D. Britton, "Regulating safety-critical autonomous systems: Past, present, and future perspectives," in *The Psychology of Interacting With Robots*, R. Pak, E. D. Visser, and E. Rovira, Eds. London, U.K.: Elsevier, 2019.
[32] M. A. Alcorn *et al.*, "Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects," 2018, *arXiv:1811.11553*.
[33] *Manual on Uniform Traffic Control Devices—Part 3: Markings*, US DOT, US Department of Transportation, Washington, DC, USA, 2000.
[34] *Test Protocol—Lane Support Systems, European New Car Assessment Programme*, Euro NCAP, Leuven, Belgium, 2017.
[35] C. Wiacek, G. Forkenbrock, and M. Mynatt, "Applying lane keeping support test track performance to real-world crash data," in *Proc. 26th Int. Tech. Conf. Enhanced Saf. Vehicles (ESV), Technol., Enabling Safer TomorrowNat. Highway Traffic Saf. Admin.*, 2019, pp. 1–19.
[36] *Drivers Let Their Focus Slip as They Get Used to Partial Automation*, IIHS, Insurance Institute for Highway Safety, Arlington, VA, USA, 2020.
[37] K. M. Wilson, S. Yang, T. Roady, J. Kuo, and M. G. Lenné, "Driver trust & mode confusion in an on-road study of level-2 automated vehicle technology," *Saf. Sci.*, vol. 130, Oct. 2020, Art. no. 104845.
[38] Euro NCAP. (Apr. 30, 2021). *Assisted Driving 2020: Tesla Model 3*. [Online]. Available: https://www.euroncap.com/en/ratings-rewards/assisted-driving-gradings/?ratingId=41020
[39] K. Barry. (Apr. 30, 2021). *CR Engineers Show a Tesla Will Drive With no One in the Driver's Seat*. [Online]. Available: https://www.consumerreports.org/autonomous-driving/cr-engineers-show-tesla-will-drive-with-no-one-in-drivers-seat/

Mary (Missy) L. Cummings (Senior Member, IEEE) received the Ph.D. degree in systems engineering from the University of Virginia in 2004. She is currently a Professor with the Department of Electrical and Computer Engineering and the Department of Computer Science, Duke Institute for Brain Sciences (DIBS), Duke University. She is also the Director of the Humans and Autonomy Laboratory and became a Safety Advisor at the National Highway Traffic Safety Administration in November 2021.

Ben Bauchwitz received the bachelor's degree in brain and cognitive science from the Massachusetts Institute of Technology. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Duke University. His research focuses on evaluating safety and establishing performance bounds for autonomous systems.