Human Factors Assessment of On-road L2 Driving

Recommendations for the implementation of partially-automated vehicles.

By Dr. Francesco N. Biondi
Human Systems Lab, University of Windsor

University of Windsor
uwindsor.ca

human systems lab
hslab.org

Title

Human Factors Assessment of On-Road L2 Driving: Recommendations for the implementation of partially-automated vehicles.

Authors

Francesco N. Biondi, Ph.D., & Noor Jajo
francesco.biondi@uwindsor.ca

Human Systems Lab, University of Windsor

Table of Contents

Executive Summary

The current study investigates drivers' cognitive workload, physiological activation, and visual attention allocation during real-world SAE L2 driving. Drivers drove a vehicle in either manual or L2 mode for a total of approximately 80 minutes. Performance to the ISO Detection Response Task, heart rate and heart rate variability, pupil size and blink rate, and off-road-glances were recorded with drivers driving a 2022 Tesla Model 3 on the section of Ontario Highway 401 between Windsor, ON and Chatham, ON. Results showed that, while cognitive workload was unaltered by L2 driving, significant changes were observed in drivers' attention allocation. In particular, when the L2 system was operational, drivers were more likely to look away from the forward roadway toward the vehicle's touchscreen. Individual glance durations to the vehicle's touchscreen were also longer suggesting that drivers' visual attention was allocated away from the road for longer. We use these findings and the data in the existing literature to suggest potential road safety recommendations to government regulators.

Introduction

The introduction of SAE level-2 (L2) or partially automated systems is intended to benefit road safety by aiding the human driver in the driving task. It is estimated that there are thousands of vehicles equipped with L2 systems already driving on our roads. Yet there is very limited knowledge on how motorists use this technology and the potential safety risks associated with operating L2 systems on our roads. The current study helps fill this gap by investigating how drivers use a vehicle equipped with an L2 system in two modes: manual and partially-automated. 30 drivers drove a 2022 Tesla Model 3 equipped with Tesla's proprietary L2 system 'Autopilot' for up to 2 hours on Ontario Highway 401 between Windsor, ON, and Chatham-Kent, ON. Physiological, behavioral, and ocular markers were recorded. Results from this study adds to our understanding of how drivers use L2 systems in the real-world, and highlights some unintended safety risks associated with their adoption.

Levels of Automation

The Society of Automotive Engineers defines six levels of ADS from fully-manual (level 0) to fully-automated (level 5) [1]. A level 2 (L2) or partially-automated system maintains control of the vehicle's longitudinal (speed) and lateral (lane position) behavior and the human driver is responsible to actively monitor its functioning and resume manual control whenever necessary (figure 1). The presence of L2 systems is rapidly increasing with the share of vehicles equipped with L2 systems being estimated to reach 60% of new vehicles sold in 2025 [2].

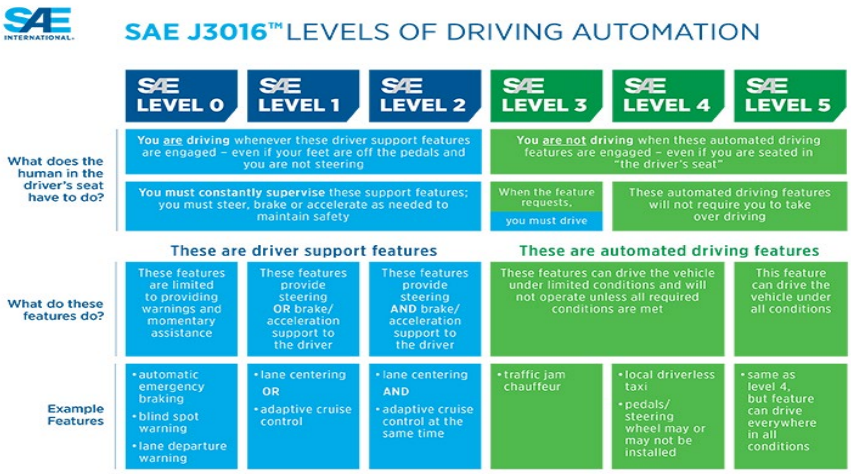

Figure 1. Six levels of driving automation.

The adoption of L2 systems comes with intended safety benefits. For example, driving with an ADS that is capable of maintaining the vehicle safely within the lane and at a safe distance from the vehicle in front may help mitigate the safety risks of driving in conditions of high driver workload resulting from, e.g., a congested traffic environment or poor visibility. Despite their intended safety benefits, a still limited yet growing body of research has shown some potential human factors safety risks of operating L2 systems.

Driver workload and engagement.

Cognitive workload is the "demand for cognitive control imposed by a task" [3]. There is a direct relationship between cognitive workload and performance. The modified version of the Yerkes-Dodson law (figure 2) stipulates that optimal performance in the task of driving results from intermediate levels of workload. High levels of workload or overload occur when overall task demands exceed the driver's cognitive capacity. Low levels of workload or underload occur when the demand imposed by the driving task is far less than the driver's capacity. While the two conditions manifest themselves differently - overload is typically associated with distraction whereas underload results in boredom or drowsiness - both are detrimental for driver behavior and road safety. Distracted driving research has shown an increase in braking times as well as a substantial reduction in the driver's field of view (i.e., the size of the visual area being actively scanned by the driver) under conditions of cognitive overload [4]–[6]**.** Likewise, drowsy driving studies have also observed delayed driver reactions to safety-critical events in conditions of underload [7], [8].

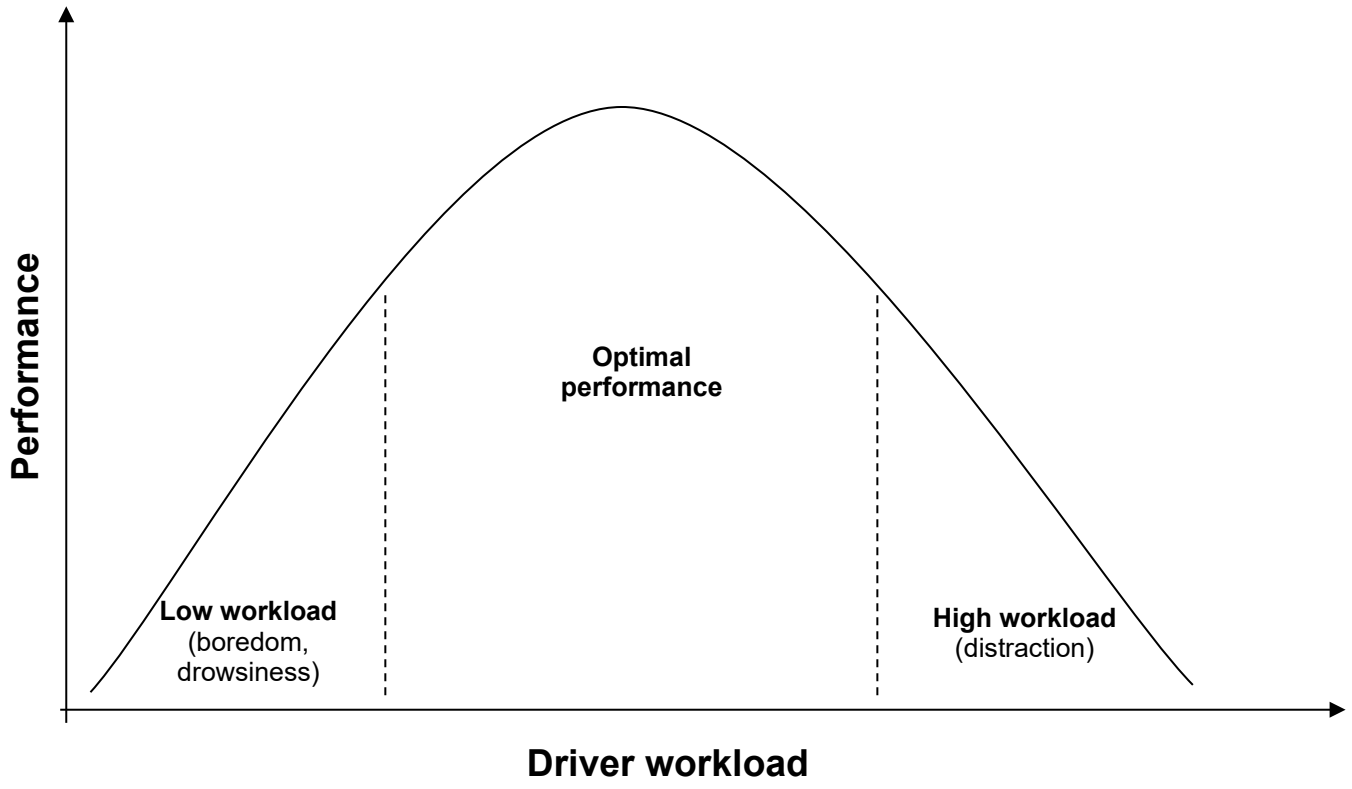

Figure 2. Modified version of the Yerkes-Dodson law.

Borrowing from Human Factors research in aviation [9], [10], Biondi et al. [11] and Strayer et al. [12] posit that the adoption of L2 systems leads to the role of the human driver switching from *operator* to *supervisor.* Whereas manual driving requires the human to maintain hands-on control of the vehicle, it is anticipated that the drastic switch in driver responsibilities will lead to drivers playing a more hands-off role in supervising the system functioning, which may result in impaired performance in resuming manual control of the vehicle whenever necessary.

Safety

The last several years have witnessed a rise in incidents involving drivers being caught at the wheel of L2 system vehicles while seemingly being distracted or completely disengaged from the active task of driving. Investigations conducted by the US National Transportation Safety Board on numerous crashes involving L2 systems list driver inattentiveness and poor monitoring of the automated system as probable causes of these accidents [13], [14]**.** To shed light on the safety risks of L2 systems, in 2021 the US National Highway Traffic Safety Administration (NHTSA) started requiring manufacturers and operators to report certain crashes involving L2 systems. In 2022, preliminary data recorded over a period of 11 months were published [15]**.** A total of 392 crashes were reported as of May 2022. Of the total number of accidents, 273 involved Tesla vehicles, 90 involved Honda vehicles, with the remainder involving vehicles from other manufacturers (figure 3).

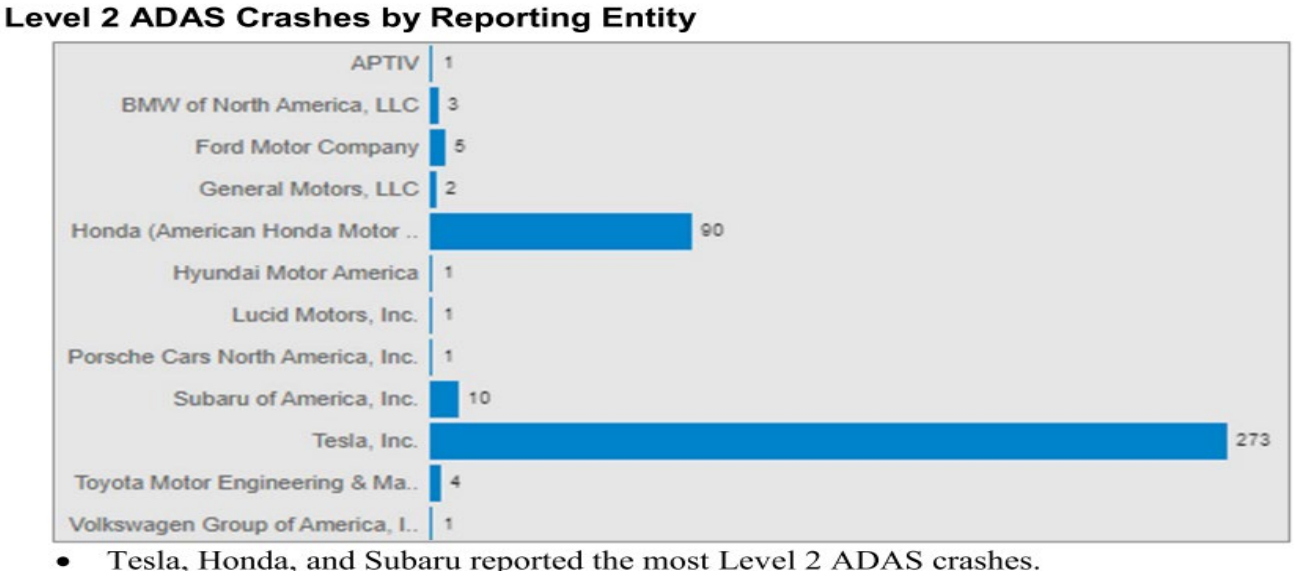

Figure 3. Crashes involving L2 systems [15].

NHTSA points out some of the limitations associated with the data collection. For example, the way data were collected may have been different across manufacturers. Likewise, some of the incident data may have been incomplete or unverified at the time of reporting. More importantly, the reported data are not normalized, i.e., they do not account for the total number of vehicles being manufactured or currently on the road.

Reports from NHTSA and NTSB add to the body of knowledge investigating the potential unintended safety risks of operating L2 systems. Most of the research conducted in this field has adopted driving simulators to investigate driver use of L2 systems. Simulated studies have identified the potential for drivers to become less attentive toward driving when the L2 system is engaged [16], [17]**.** Likewise, drivers show delayed responses when control of the vehicle is transferred from the automated system back to the human [18]**.** It is worth noting that, while driving simulator studies are useful to build foundational knowledge, they often lack the degree of ecological validity that is necessary to extend findings to the real-world [19], [20].

Studies investigating driver interaction with real-world L2 systems are limited and have produced mixed results. In previous work, Gaspar and Carney [21] had a small sample of 12 drivers operate a vehicle equipped with an L2 system over multiple trips, each with an average duration of 13 minutes. Participants drove the vehicle in manual and L2 mode. Compared to manual driving, operating the vehicle in L2 mode resulted in overall longer glances away from the forward roadway, indicating a potential safety risk of driving in partially-automated mode. In a similar study, Biondi et al. [22] measured drivers' responses to a detection task with participants driving in manual and L2 mode. A reduction in detection task performance was found in the latter suggesting a decline in driver vigilance when the L2 system was engaged. McDonnell et al. [23] and Lohani et al. [24] recorded participants' physiological and neural activity while driving an L2 system-equipped vehicle over two 20-minute manual and partially-automated drives. No significant differences were found in neurophysiological activations between the two modes, a pattern that is seemingly at odds with the findings by Biondi et al. [22] and Gaspar and Carney [21].

Current study.

The current study adopts a combination of physiological (heart rate, heart rate variability), eye-tracking (pupil size, blink rate, glance distribution), behavioral (ISO Detection Response Task performance [3]), and subjective (self-reported workload ratings and drivers' subjective assessments) metrics to investigate changes in driver behavior between manual and L2 mode. Unlike some of the previous studies where participants drove the vehicle for 20 minutes or less, our study investigates driver behavior over longer manual and L2 drives of approximately 40 minutes each. Dunn et al. [25]posit the amount of exposure to the ADS has a direct effect on how drivers use it and their overall behavior. In particular, as drivers become more familiar with it, this will lead to them becoming more over trusting of the system and potentially more inattentive. With this in mind, the current study aims to:

1. **Investigate the effect that operating an L2 system on the road has on drivers' cognitive workload.** It is hypothesized that as the role of the driver switches to *system supervisor* during L2 driving, this will lead to a reduction in cognitive workload. The ISO Detection Response Task (DRT), blink rate and pupil size are recorded as metrics of cognitive load.
2. **Explore differences in drivers' physiological activation between manual and L2 driving**. It is hypothesized that drivers' physiological activation may decline during L2 driving. Heart rate (HR) and heart rate variability (HRV) are used as metrics of drivers' physiological activation.
3. **Investigate the effect of L2 driving on attention allocation.** Should drivers become less attentive toward driving and the task of supervising the functioning of the L2 system, this will lead to a reduction of glances executed toward the forward roadway.
4. **Measure drivers' subjective experience when driving in L2 and manual mode.** Drivers' subjective experience is recorded to investigate how it changes between manual and L2 driving.

Method

Participants

30 volunteers (13 females) participated in this study. Their average age was 22 years old and standard deviation of age was 4.36 years. Inclusion criteria included: be a fluent English speaker; have normal or corrected-to-normal vision and hearing; hold a valid driver's license; have proof of car insurance; have not been the at-fault driver in an accident within the past 2 years; provide a valid Ontario Driver's Abstract; complete a 30-minute defensive driving course; be affiliated with the University of Windsor (e.g., students, researchers, faculty, staff). A University of Windsor Research Ethics Board approval (REB #20-141) was obtained for the study.

Experimental Design

We adopted a factorial design with one independent factor: driving mode. Participants drove the vehicle in one of two modes: manual (SAE level-0) or L2. Dependent measures included: performance to the ISO DRT; HR and HRV; attention allocation on and off the forward roadway; self-reported ratings and driver experience.

Equipment, procedure and data processing

Vehicle

A 2022 Tesla Model 3 was used for the study (figure 4. In L2 mode, participants drove the vehicle with both Adaptive Cruise Control (ACC) and Lane Keeping Assist System (LKAS) engaged. ACC maintains the vehicle at a set speed and distance from the vehicle in front and, during the study, it was set to the maximum distance of 7 car lengths. LKAS maintains the vehicle within the lane and, during the study, participants were instructed to occupy the right lane when two lanes were available or the middle lane when three lanes were available. This was done to minimize lane changing and the need to overtake slower traveling vehicles. The joint functioning of ACC and LKAS meets the requirements of L2 ADS as per the SAE taxonomy [1]. In manual mode, participants drove the vehicle manually with no assistance from ACC or LKAS.

Figure 4. Exteriors of the vehicle used in the study.

The vehicle was equipped with a 15-inch horizontally-oriented touchscreen (figure 5) that displayed information on the functioning of the L2 system as well as other vehicle and infotainment information. During the study, the radio and infotainment system were turned off.

Figure 5. Interiors of the vehicle used in the study.

Route

Participants drove two routes. The familiarization route was used to help participants familiarize with the vehicle and the functioning of its L2 system. This route (figure 6) consisted of a loop through Huron Church road and Ontario Highway 3 between approximately the University of Windsor campus and the St. Clair College campus in Windsor, ON.

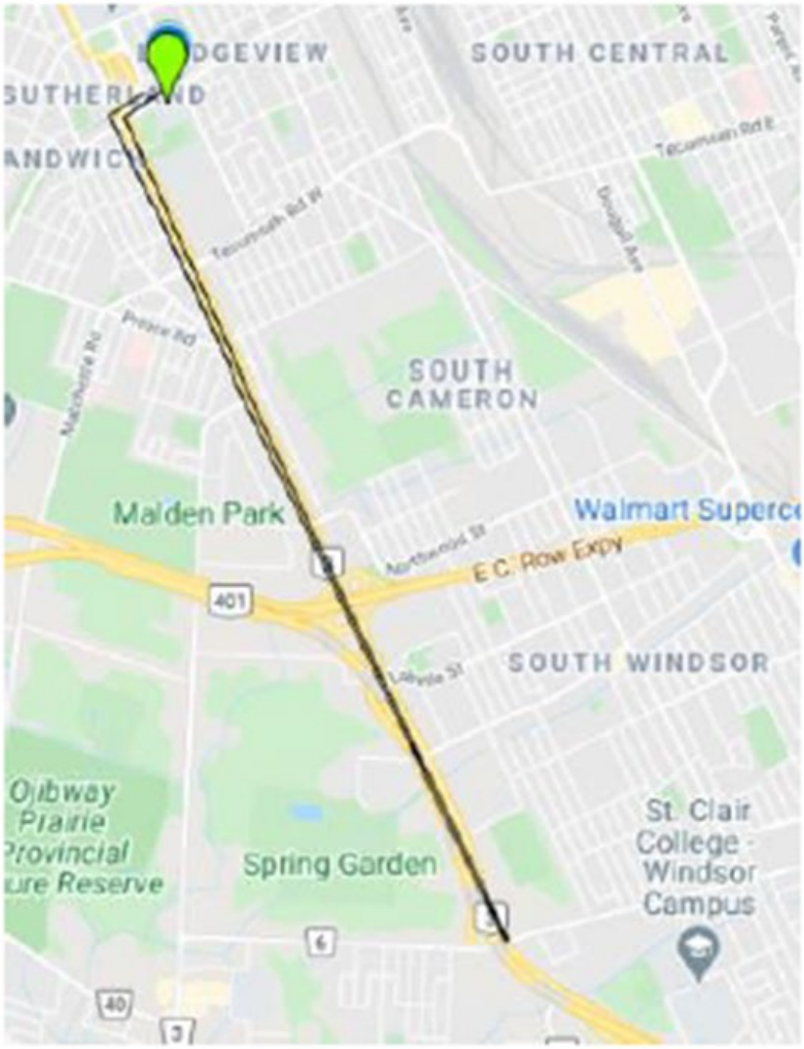

Figure 6. Familiarization route.

The experimental route consisted of a section of Ontario Highway 401 approximately between Exit 13 in Windsor, ON and Exit 81 near Chatham, ON (figure 7). The road uses a combination of three-lane and two-lane dual carriageways that provide sufficient space between vehicles. The daily traffic volume on the route was approximately 25,000 vehicles. The data collection ran during off-peak hours between 10am and 3pm weekdays and weekends. Participants were instructed to obey all traffic laws and never exceed the posted speed limit of 100 km/hour or 110 km/hour.

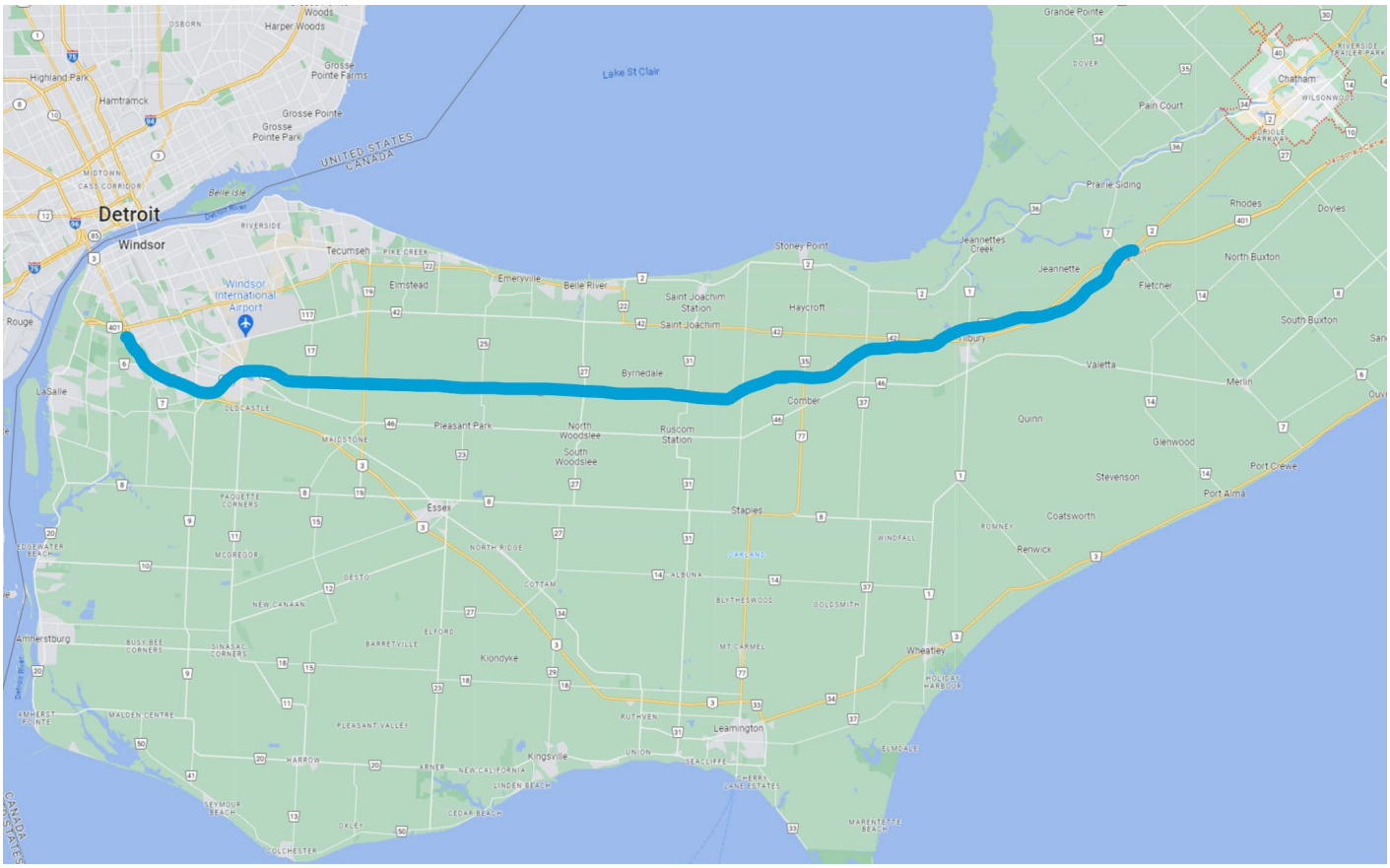

Figure 7. Experimental route.

Cameras

The vehicle was retrofitted with three GoPro HERO8 Black cameras. One camera was pointed at the driver to record eye glances and head movements (driver view). A second camera was pointed at the forward roadway (front view), A third camera was pointed at the vehicle's center stack touchscreen (touchscreen view). A diagram of view from the three cameras is available in figure 8. The cameras recorded at 1080p at 240 frames per seconds.

Figure 8. Views from each of the cameras: (A) touchscreen view; (B) front view; (C) driver view.

Eye glances

Videos recorded from the driver view camera were manually coded. Four areas of interest (AOI) were identified: forward roadway, side mirrors (both left and right), rearview mirror, instrument panel. A diagram of the areas of interest is presented in figure 9. The decision to consider these AOI is consistent with the work by Biondi et al. [4] and Gaspar and Carney [21]. Coders manually coded each video frame-by-frame to measure time each driver spent looking at each AOI during the manual and automated drive. Each drive was analyzed by two trained coders and any discrepancies in the coding were flagged and reviewed for consistency by a third coder. In general, coders were very accurate and only a small number of events needed to be double-checked. An Intra-class Correlation (ICC) analysis for ordinal variables was performed to measure the consistency. An ICC of 0.822 was found, indicating that there was strong consistency between the two coders.

Glance Coding Definitions	
Front View (Blue)	Any glance that is directed toward the forward roadway glance. This includes glances directed toward the center, left or right side of the road to, e.g., inspect the road or vehicle in front, check for potential hazards or road signs.
Instrument Panel (Purple)	Any glance that is directed at the instrument panel where information about the vehicle's functioning and SAE level-2 system is displayed.
Rear view Mirror (Yellow)	Any glance that is directed at the rear view mirror.
Side Mirrors (Green)	Any glance that is directed at the left or right-side mirror.

Figure 9. Diagram and table describing the AOIs used for glance coding.

The total time spent glancing at each AOI was calculated to investigate any differences between the two driving modes. Total-eyes-off-road-time (TEORT) was also calculated as the total time spent glancing away from the forward roadway in each mode. Average and maximum glance durations were also calculated.

DRT

The vibrotactile version of the DRT manufactured by Red Scientific Ltd (Salt Lake City, UT, USA) as per [3] was used in the study. A vibrotactile motor was placed on the inside of the participants' left elbow area and a microswitch was attached to either the index or middle finger of the left hand. The vibrotactile motor emitted a short stimulus (1 s in duration), similar to a phone vibration. Upon its presentation which occurred every 3–5 s, participants were instructed to press the microswitch as fast as possible. Reaction times in milliseconds and hit rates were recorded. RT were recorded as the time interval between the onset of the vibrotactile stimulus and the depression of the microswitch. Accuracy was calculated as the ratio between the numbers of hits and the number of total presented stimuli. In accordance to ISO guidelines, responses faster than 100 ms or longer than 2,500 ms were eliminated from the calculation. nonresponses or responses produced later than 2,500 ms were considered as misses.

Physiological recording.

A Biopac MP 160 system manufactured by Biopac System Inc. (Goleta, CA, USA) was used for the physiological recording. Participants wore three electrodes: one on the left collarbone, one on the right collarbone, and one lower left rib cage to form a triangle around their heart area. Data cleaning and processing was conducted using Acknowledge software and consistently with the existing literature [24], [26]. Raw data was bandpass filtered at 1 and 35 Hz cutoffs, Upon conducting visual inspections of the ECG spectrograms, the research assistant manually added R-wave peaks that were missed by the software's automatic detection algorithm. Artifacts generated by head or body movements (e.g., when the participant inadvertently touched one of the electrodes to check the vehicle's blind spot) were manually removed, and R-wave peaks were interpolated. After data cleaning, average HR and HRV were calculated for the baseline, manual, and L2 conditions. The root mean square of successive differences between normal heartbeats (RMSSD) is commonly used as a measure of physiological activation in the related literature and was used in this study as a metric of HRV. RMSSD was calculated using 30-second epochs. For RMSSD and HR, data points exceeding 3 standard deviations from the mean of normal distribution were considered as outliers and removed from the analysis. Average HR and RMSSD were baseline corrected as follows: average HR and RMSSD recorded in the baseline condition were subtracted from the data points recorded in the manual and L2 driving conditions to account for interindividual differences. Baseline-corrected HR and RMSSD averages were then calculated for the two conditions.

Subjective workload

At the end of each experimental drive, participants were instructed to complete a NASA-Task Load Index (NASA-TLX) questionnaire on an Apple iPad. The NASA-TLX [27] is a widely-adopted tool for assessing

workload. It consists of six scales measuring: mental demand (how mentally demanding was the task?); physical demand (how physically demanding was the task?); temporal demand (how hurried or rushed was the pace of the task?); performance (how successful were you in accomplishing what you were asked to do?); effort (how hard did you have to work to accomplish your level of performance?); frustration (how insecure, discouraged, irritated, stressed, and annoyed were you?). Participants answered each question using a 21-point Likert scale with 1 being very low and 21 being very high. Average ratings for the six scales were calculated for the two conditions.

Eye-tracking

Eye-tracking data was processed using the Pupil Player software (v3.1.16, Pupil Labs, Berlin, Germany). The pupil detection algorithm adopted a three-dimensional model for estimating the center and the rotation of the eyeball, and detected the pupil by searching for the darkest region and creating an ellipse around this region. For blink rate, a filter length of 0.2 s and a confidence threshold onset/offset between 0.5 and 0.3 was adopted. The onset and offset thresholds are, respectively, the thresholds that the filter response must rise above or fall below to classify the onset and end of a blin. The filter length represents the time interval wherein the blink detector attempts to find confidence drops and gains. In line with the work by Biondi et al. [28]–[30], blinks detected with a confidence level below 20% were not classified as blinks and excluded from the analysis. Blink rate was calculated for each of the two experimental conditions as the number of blinks recorded per minute.

For pupil size, data below 2 mm and above 9 mm were considered outliers (3.32% of the total data) and then removed from further analysis (for a similar procedure, see [31], [32]). The mean pupil size (total area) recorded in the baseline condition was used to normalize the data in the experimental conditions as follows: $(x-\mu)/\sigma$, where x is the observed value, μ is the mean in the baseline condition, and σ is the standard deviation in the baseline condition. Normalized pupil size for each condition was then calculated.

Procedure

Intake and pre-study questionnaires

Participant recruitment occurred via email advertisement and word-of-mouth. Prior to the study, participants were contacted by the research associate to complete a demographic form wherein they provided information about:(1) their age, (2) whether they held a driver license and how long they have had their driver license for, (3) the class of their license (G1,G2, G), (4) their eye vision and hearing(normal or corrected) and if they wore lenses or glasses, (5) whether English was their native language, and (6) their familiarity with advanced driver assistance systems. One week prior to the day of the study, participants were required to provide their driver license and driving record of the past 3 years, proof of insurance, drivers' insurance forms, and successfully complete a 30-minute Stantec defensive driving fundamentals course accompanied by a certification test. Upon providing this information, participants were added to the University of Windsor's vehicle's insurance. They were next

required to sign the consent form and the REB checklist. Following the confirmation of these prerequisites, they were provided with a 1-minute video tutorial on the Tesla Model 3 which is available at https://www.youtube.com/watch?v=IkSw2SZQENU. The video contained a step-by-step guide on how to engage the vehicle's L2 system, the driver's requirements to monitor and stay attentive during the functioning of the L2 system, and how to regain manual control of the vehicle.

On the day of the study, participants met with the research associate in front of the Centre for Engineering Innovation (CEI) at University of Windsor and were directed to the CEI garage where the vehicle was parked. When inside the vehicle, participants were asked to review the signed consent form and REB checklist and ask any questions they may have. They also had to fill out a questionnaire that screened for the use of alcohol, drugs, marijuana, and the excessive amount of coffee. Note that the study's procedure complied with the research protocol approved by the University of Windsor's Research Ethics Board (REB #20-141).

Equipment setup, training and experimental drives.

Once in the vehicle, the research associate instructed participants on how to adjust the vehicle controls and engage the L2 system. Participants were then instructed on how to use the DRT by having them complete a 1 minute practice session. The research associate began calibrating the eye tracker once they verbally stated that they were comfortable with the DRT. The eye-tracking calibration involved participants gazing at a marker affixed to a large black board and moving their heads in a spiral motion until the eye tracker software detected their gaze. Once the eye-tracker was calibrated, participants were asked to switch off or put their phones on mute. Baseline recordings for HR and HRV occurred with participants staring at a fixation cross for 3 minutes prior to the start of the study while the vehicle was parked. The familiarization process began when participants were asked to drive on the familiarization route (see figure 6). At this time, they practiced driving the vehicle in manual mode and operating the L2 system.

See figure 7 and above explanation for the experimental route. Once participants verbally confirmed they were comfortable driving the vehicle, they were instructed to drive to Highway 401 where the experimental phase begun. At this time they were told not to use any device including the vehicle's touchscreen and not to communicate with the research associate unless strictly necessary. Participants were also instructed to occupy the middle lane if three lanes were available or the right lane if two lanes were available, and to obey all traffic laws. The first experimental drive ended at exit 81 in Chatham, ON. After exiting the highway, participants took a 15-minute break at a gas station where they filled out the NASA-TLX and were given the option to drink and eat. Whenever participants felt ready to for the second drive, they were instructed to drive back to highway 401 where the second experimental drive begun. The drive concluded at exit 13 in Windsor, ON, at which point participants were instructed to drive back to the University of Windsor campus wherein they filled out the NASA-TLX.

Statistics and data analysis

Bayes factor analyses were conducted to investigate the effects of the factor driving mode on the dependent measures.. The Bayesian approach was preferred over the traditional null-hypothesis statistical testing (NHST). NHTS relies on the p-value associated with a test to determine whether the null hypotheses is accepted. Traditionally, an alpha threshold of 0.05 is set so that if the test's probability is greater than alpha, the null hypotheses is accepted, and vice versa. Bayesian analysis set up two competing models, one in favor of the null hypothesis and the other in favor of the alternative hypothesis, and estimate which of the two models is more likely to generate the data at hand. In details, the Bayesian approach transforms the p-values into direct evidence against the null hypotheses [33]. The Bayes Factor (BF), which is used the determine the likelihood of the data under either the null or the alternative hypotheses, is calculated as the ratio between the marginal likelihood of the null model and that of the alternative model [34]. A BF equal to X indicates that the data is X times more likely under the alternative hypotheses than under the null hypothesis. For example, a BF of 10 indicates that the given data is 10 times likelier under H1, whereas a BF = 0.01 indicates that the same data is 10 times likelier under H0. According to Dienes [35], BF varies between 0 and infinity. The bigger the BF (with BF > 1), the stronger the evidence in support of the alternative hypotheses. Likewise, the smaller the BF (with BF < 1), the stronger the evidence in support of the null hypotheses. BF = 1 indicate that the data is not supportive of either model. In short, unlike NHST which only yields a binary outcome (accept/reject H0), BF analysis allow for three separate conclusions (evidence in support of H0, evidence in support of H1, and insensitive evidence) as well as provides information on the strength of the evidence. Data processing and analyses were conducted using R (version 4.1.0) and RStudio (version 2023.03.0 [36]). The tidyverse (version 2.0) and BayesFactor (version 9.12) libraries were adopted for data processing and Bayesian analyses, respectively.

Results

DRT RT

Figure 10 shows the DRT RT by mode. RT in L2 mode averaged 541 milliseconds and RT in manual mode averaged 551 milliseconds. A Bayesian t-test was conducted to investigate the effect of mode on DRT RT. A BF of 0.33 was found suggesting that there is no difference between modes.

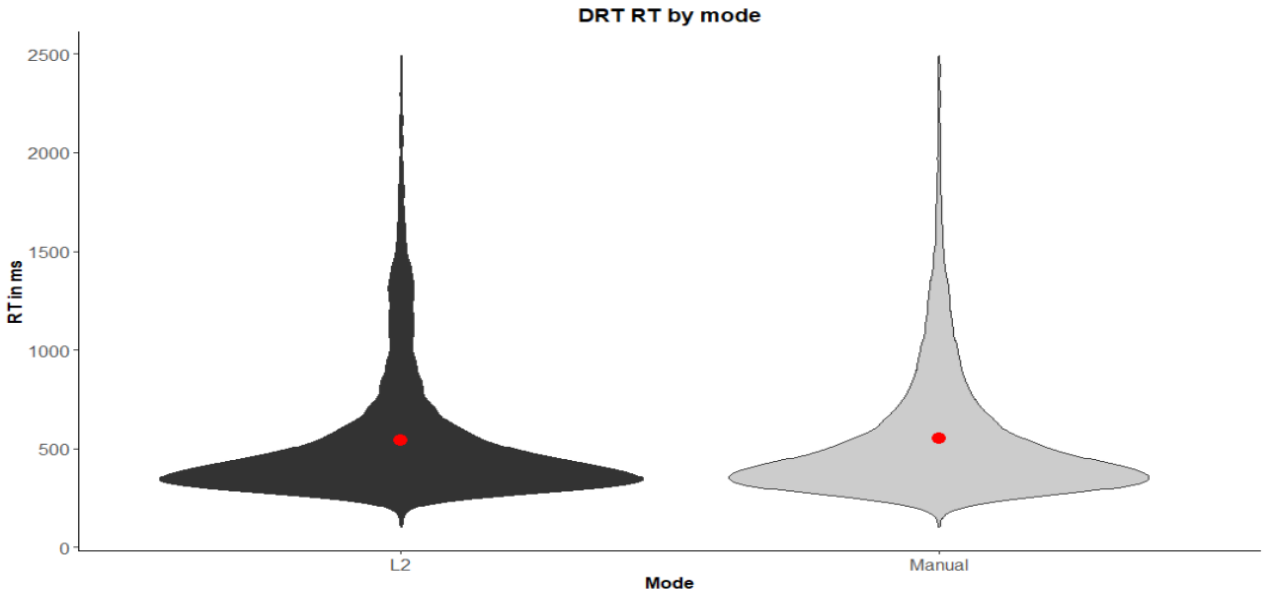

Figure 10. Violin plot showing DRT RT distribution by mode. The red dots represent average RT in the two conditions.

DRT accuracy

Average accuracies in percentages were 91.1% in L2 mode and 93.8% in manual mode. Figure 11 shows average and standard error of accuracy in the two conditions. A Bayesian t-test was conducted to investigate the effect of mode (2 level: L2 and manual) on DRT accuracy. A BF of 0.52 was found indicating anecdotal evidence in support of the null hypotheses that there is no difference between the two modes.

Mode	Average	Standard Error
L2	91.1%	0.01%
Manual	93.8%	0.01%

Figure 11. Average and standard error of DRT accuracy in percentages in the two modes.

HRV

Figure 12 shows average baseline-corrected HRV by mode. A Bayesian t-test was conducted to investigate the effect of mode (2 level: L2 and manual) on HRV accuracy. A BF of 0.29 was found indicating that HRV did not change between modes.

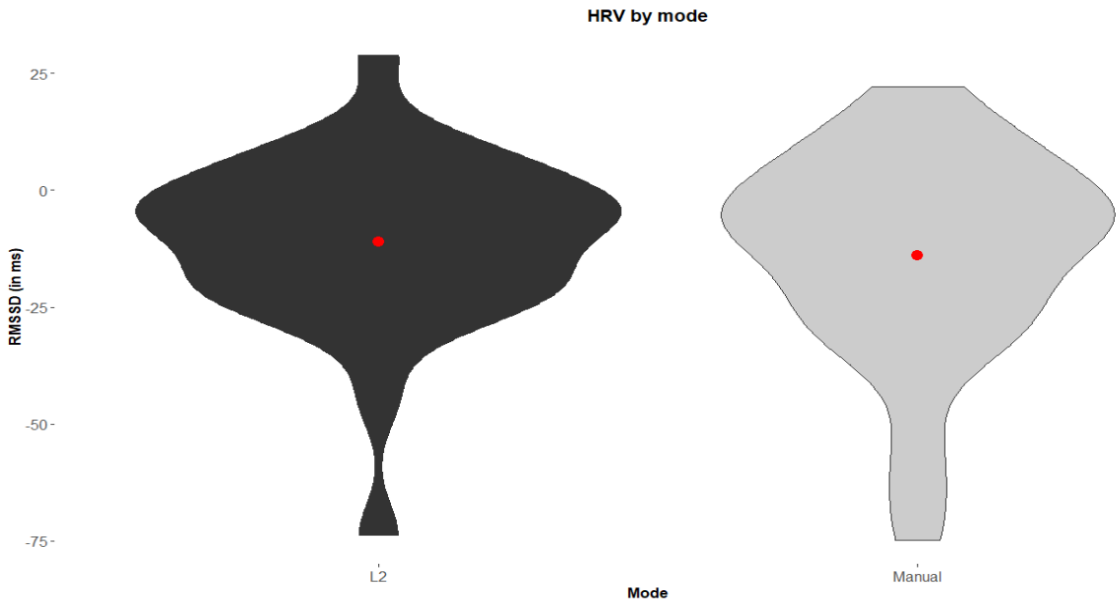

Figure 12. Violin plot showing baseline-corrected HRV distribution by mode. The red dots represent average HRV in the two conditions.

HR

Figure 13 shows average baseline-corrected HR by mode. A Bayesian t-test was conducted to investigate the effect of mode (2 level: L2 and manual) on HRV accuracy. A BF of 0.12 was found indicating that HR did not change between modes

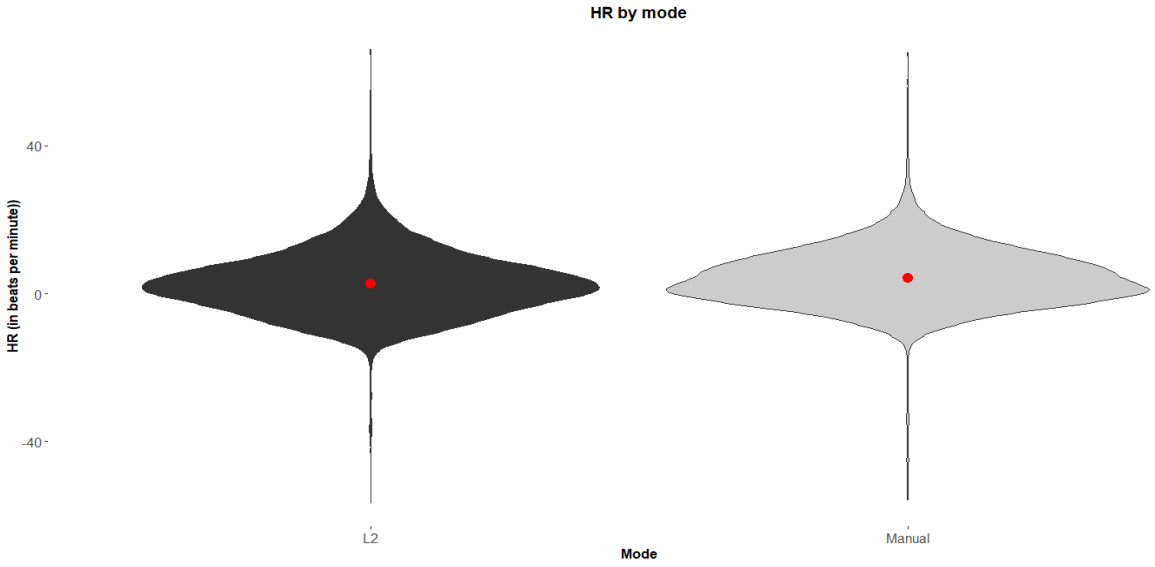

Figure 13. Violin plot showing baseline-corrected HR distribution by mode. The red dots represent average HRV in the two conditions.

NASA-TLX

Figure 14 shows average and standard error for the six scales of the NASA-TLX in the two modes. Bayesian t-tests were conducted to investigate the effect of mode (2 level: L2 and manual) on each individual scale. Analysis show evidence in support of the hypotheses that there are no difference between modes for the mental scale (BF = 0.57), the temporal scale (BF = 0.33), the performance scale (BF = 0.29), the effort scale (BF = 0.80), and the frustration scale (BF = 0.39). Analysis for the physical scale suggest that participants experienced a slightly higher physical demand in the manual mode relative to the L2 mode (BF = 2.08).

Scale	**Manual**		**L2**	
	Average	SE	Average	SE
Mental	9.93	0.78	8.45	0.76
Physical	7.66	0.84	5.24	0.67
Temporal	6.59	0.68	7.41	0.84
Performance	6.14	0.54	6.55	0.60
Effort	9.38	0.86	7.52	0.75
Frustration	6.79	0.87	7.93	0.75

Figure 14. Average and standard error (SE) for NASA-TLX ratings for the six scales in the two driving modes.

Eye glances

Figure 15 shows the total eyes off the road time (TEORT) by mode. Bayesian t-tests were conducted to investigate the effect of mode (2 level: L2 and manual) on TEORT. A BF of 180.53 was found indicating strong evidence that TEORT increased when drivers operated the vehicle in L2 mode.

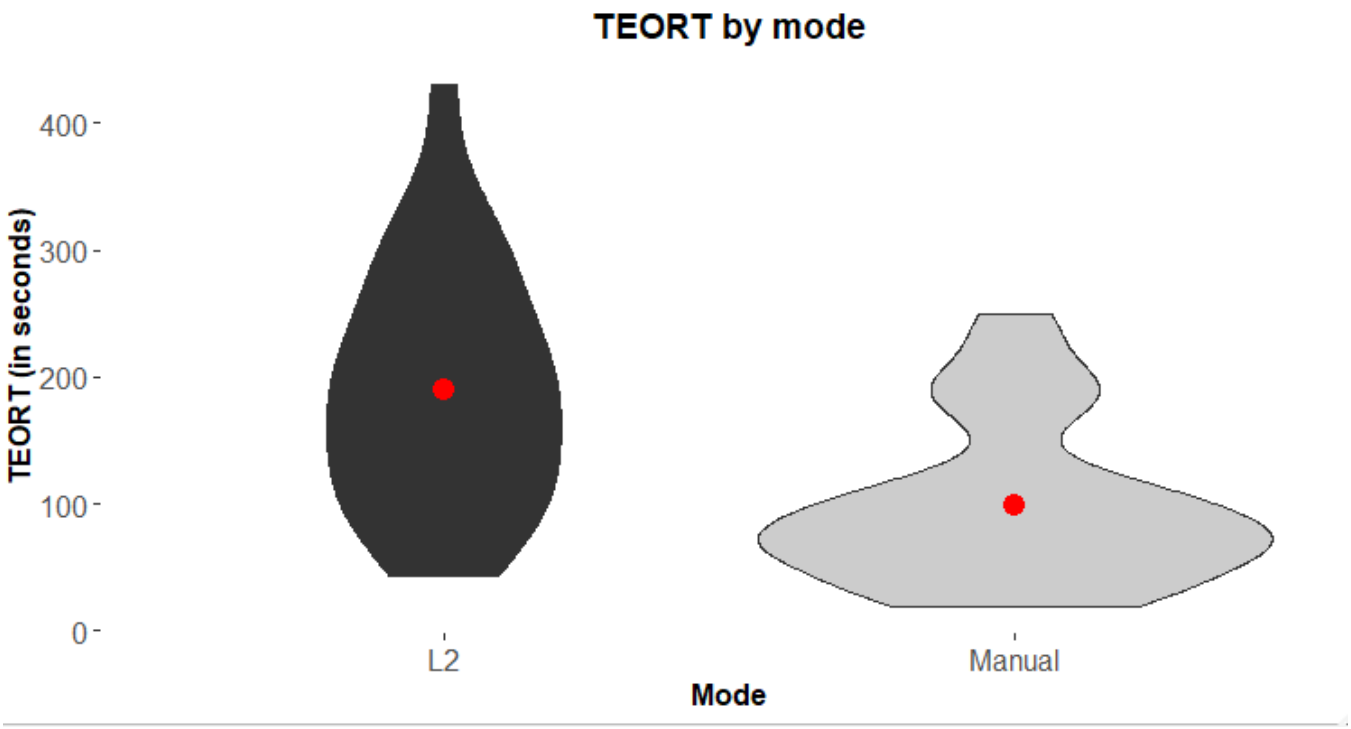

Figure 15. Violin plot showing TEORT (in seconds) distributions by mode. The red dots represent average RT in the two conditions.

Additional analyses were performed to investigate differences in total glance times by AOI. Figure 16 shows total glance times by AOI and mode. Bayesian t-tests were conducted to investigate the effect of mode (2 level: L2 and manual) and AOI (3 levels: rearview mirror, side mirrors, and touchscreen) on total glance times. Analysis revealed BF of 1.52 and 1.62 for side mirrors and rearview mirror, respectively, indicating that total glance time did not change between modes. A BF of 326.73 was found for touchscreen indicating that drivers spend more time looking at the touchscreen during L2 driving relative to manual driving. In particular, whereas drivers spent approximately 50 seconds glancing at the touchscreen during manual driving, the total glance time during L2 mode was close to 120 seconds.

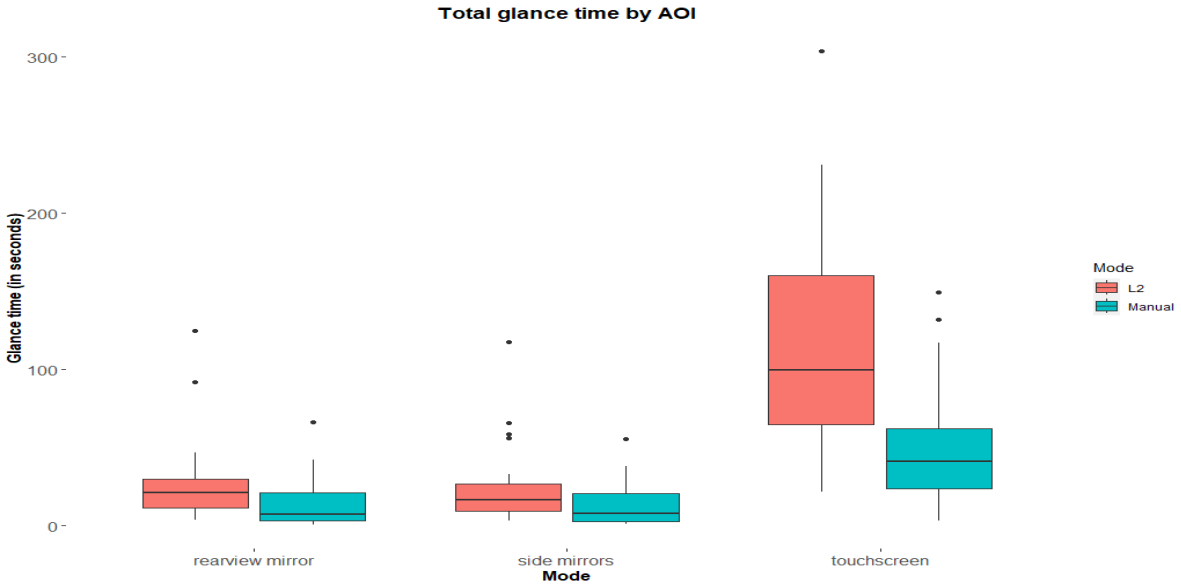

Figure 16. Bar plot showing glance time (in seconds) by AOI and mode. The lower and upper hinges correspond to the first and third quartile, dots represent outliers, and the black horizontal line within each box represents the mean.

Maximum glance duration was also calculated to investigate differences between modes. Figure 17 shows maximum glance durations by AOI and mode. Bayesian t-tests were conducted to investigate the effect of mode (2 level: L2 and manual) and AOI (3 levels: rearview mirror, side mirrors, and touchscreen) on maximum glance durations. No differences were found between modes for rearview mirror or side mirrors. However, consistently with the results found for glance time, longer maximum glance durations were found for touchscreen in L2 mode (BF=107.62). In particular, max glance duration increased from approximately 2 seconds during manual driving to 4 seconds during L2 driving.

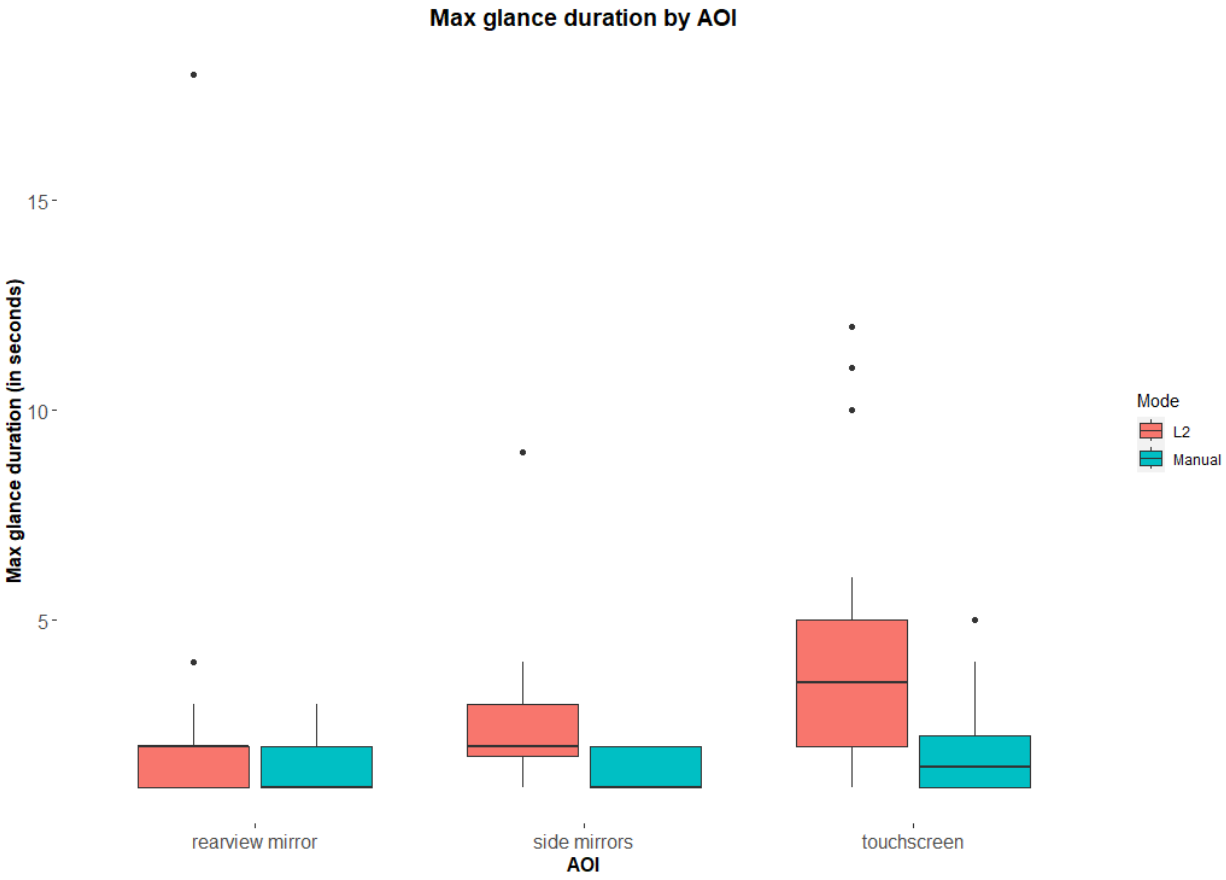

Figure 17. Boxplot showing maximum glance duration(in seconds) by AOI and mode. The lower and upper hinges correspond to the first and third quartile, dots represent outliers, and the black horizontal line within each box represents the mean.

Average glance duration was also calculated to investigate differences between modes. Figure 18 shows average glance durations by AOI and mode. Bayesian t-tests were conducted to investigate the effect of mode (2 level: L2 and manual) and AOI (3 levels: rearview mirror, side mirrors, and touchscreen) on average glance durations. Results showed that average glance duration increased during L2 mode for touchscreen (BF = 215.56) from approximately 0.2 seconds to 0.4 seconds, and side mirrors (BF = 10.62) from 0.2 seconds and 0.3 seconds between manual and L2 mode, respectively.

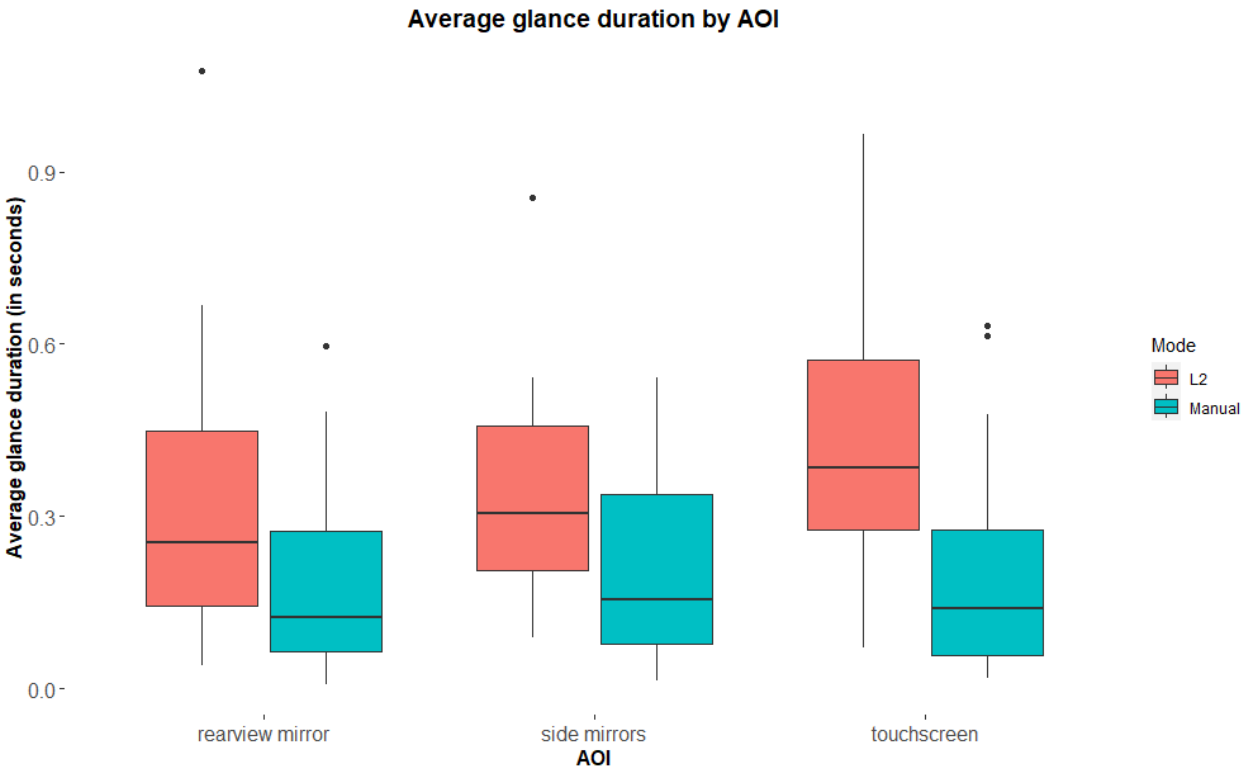

Figure 18. Boxplot showing average glance duration(in seconds) by AOI and mode. The lower and upper hinges correspond to the first and third quartile, dots represent outliers, and the black horizontal line within each box represents the mean.

Blink rate.

Figure 19 shows blink rate by driving mode. Bayesian t-test was conducted to investigate differences in normalized pupil size between driving modes. A BF of 0.34 was found suggesting no differences in pupil size between the two driving modes.

Mode	Blink rate	Standard Error of blink rate
L2	14.6	1.67
Manual	15.2	1.70

Figure 19. Blink rate and standard error of blink rate by driving mode.

Pupil size.

Figure 20 shows normalized pupil size by driving mode. A Bayesian t-test was conducted to investigate differences in normalized pupil size between driving modes. A BF of 0.55 was found suggesting no differences in pupil size between the two driving modes.

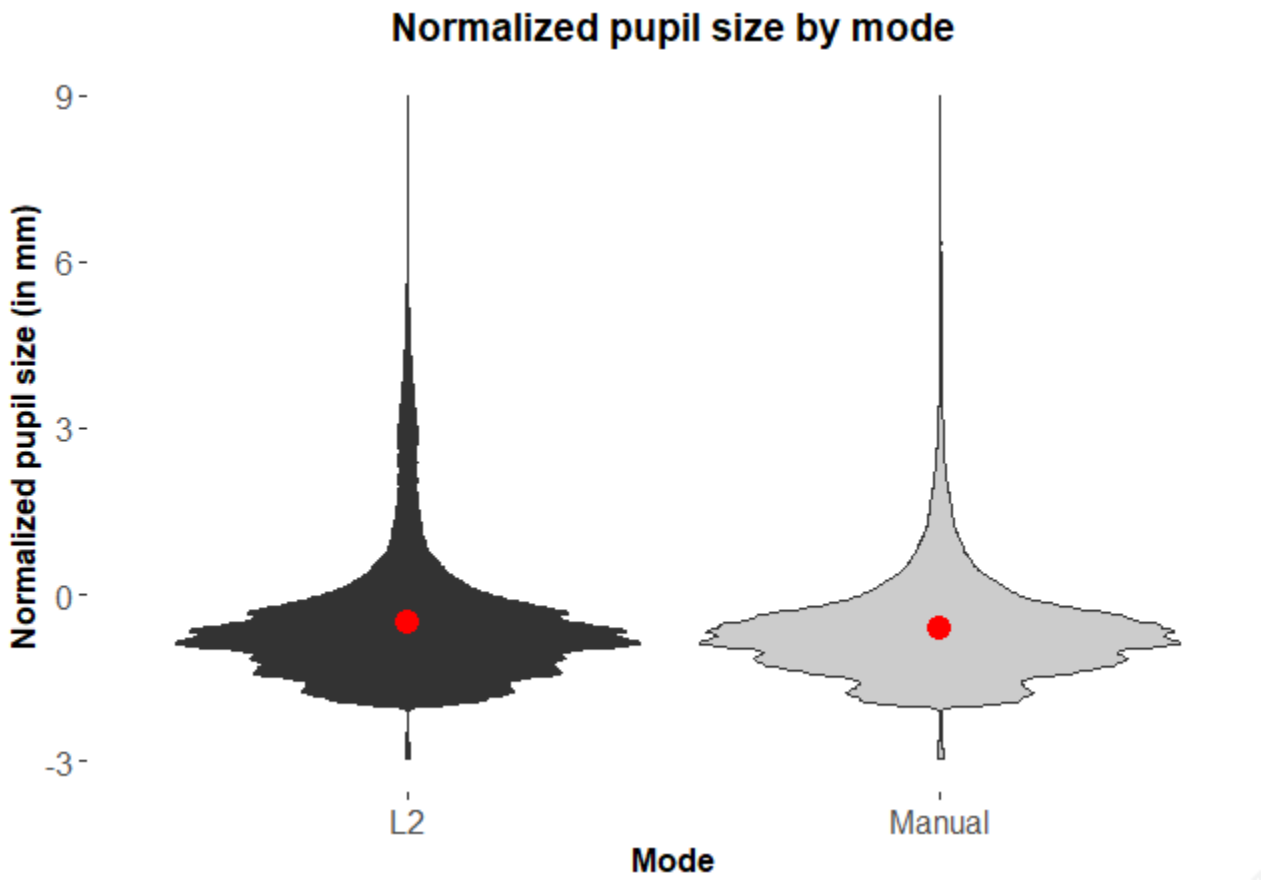

Figure 20. Violin plot showing normalized pupil size distribution by mode. The red dots represent average RT in the two conditions.

Discussion

The current study aimed to investigate changes in drivers' cognitive and visual workload, attention allocation, and subjective experiences between manual and L2 driving. Below we discuss our findings according to the four research objectives as presented in the introduction.

1. Investigate the effect that operating an L2 system on the road has on drivers' cognitive workload

Drivers' workload was measured through a combination of the ISO DRT which is a standard metric of cognitive workload, pupil size and blink rate. Results showed that no differences in cognitive workload were found between L2 and manual modes for any of these metrics. This is in contrast with our hypotheses that operating an automated system would reduce cognitive workload. In particular, considering the consistent patterns found across metrics, this strongly indicates that cognitive workload is unaltered by levels of vehicle automation. This is consistent with the findings by Strayer et al. [37] who observed no differences in drivers' cognitive workload across modes.

2. Explore differences in drivers' physiological activation between manual and L2 driving

HR and HRV were used as metrics of drivers' physiological activation. It was hypothesized that physiological activation may decline in the L2 mode relative to manual driving. In contrast with our hypotheses and in line with the findings for objective 1, no differences in physiological activation were found across modes. This pattern is consistent with the work by McDonnell et al. [23] who also failed to record any differences between modes.

3. Investigate the effect of L2 driving on attention allocation

Glance analysis revealed significant differences between L2 and manual driving. When the total eyes off the road (TEORT) metric was calculated, longer TEORT was found during L2 mode, relative to manual driving. This suggests that, upon the L2 system being operational, drivers felt sufficiently comfortable to relinquish control of the vehicle to the ADS. As a result, they started looking away from the forward roadway for longer. In particular, whereas TEORT approximately equaled 100 seconds or 4% of the total drive time in manual mode, it doubled on average to 200 seconds or 8% of the total drive time, with the TEORT for some drivers exceeding 400 seconds or 16% of the total drive time. This is concerning considering that drivers not only were trained on how to use the L2 system and stay vigilant during its functioning, but they also were aware that they were part of a controlled research study wherein they were consistently monitored, a research associate was present in the vehicle, and, most importantly, they were not allowed to use mobile devices or the vehicle's touchscreen and infotainment system. It is

then arguable that disengagements from the task of driving would occur more frequently during real-world driving.

Consistent with TEORT data, we also calculated the total glance time directed toward the side mirrors, rear view mirror, and touchscreen. No differences were found between driving modes for side and rear view mirrors. Instead, a sizeable increase in glance time was found for the touchscreen. Figure 16 shows that whereas the average total glance time was approximately 50 seconds during manual driving, it increased to approximately 100 seconds during L2 driving. Data distributions also show that whereas glance times were fairly homogenous during manual driving, a broader distribution with longer glance times was found in L2 mode.

Further analysis were conducted on average and maximum glance duration. Results showed that drivers not only spent more time looking away from the road during L2 driving, but their average glance duration also increased when the ADS was engaged. Similarly, the maximum glance duration toward the touchscreen also increased when driving in L2 mode.

4. Measure drivers' subjective experience when driving in L2 and manual mode

Drivers' subjective experience were collected at the end of the study through unstructured interviews. Several participants expressed a high level of comfort while driving in L2 mode, particularly when driving on highway sections without construction.

> *"Initially, I was skeptical, but later on, I started feeling safe and comfortable."*

> *"Driving on the highway felt remarkably comfortable during partially automated mode."*

Upon encountering the construction zone, the behavior of the L2 system changed dramatically according to some participants from "smooth" to "jolty" especially when it was unable to detect the lane markings as well as it did prior to the construction zone.

> *"I would never consider buying a Tesla. The car's failure to recognize the lane, the proximity of construction cones, and the inability to engage the automatic mode again made me feel unsafe."*

> *"I have mixed feelings about the automatic mode. While I enjoyed it at times, it became problematic whenever there was construction, and I wasn't paying full attention."*

It can be argued that the behavior of the vehicle was predictable and enjoyable in absence of constructions which may have led them to build trust toward the L2 system, but to some participants it became unpleasant during construction zones.

Altogether, study findings show that whereas cognitive workload was unaltered by driving in L2 mode, drivers' visual attention allocation significantly changed when the L2 system was operational resulting in longer time spent looking away from the forward roadway and longer glances. These findings advance our understanding of how drivers use this technology, and the Human Factors risks of operating L2 systems. Whereas existing studies explored cognitive workload and attention allocation separately, to our knowledge this is the first study that is addressing these two human factors of vehicle automation.

Policy recommendations to government regulators

Based on the study findings, below are some road policy recommendations for consideration:

1. Enhanced driver training of L2 systems.

Study findings suggest that drivers are more likely to disengage from the driving task when operating the L2 system, therefore increasing the risk of abusing the system's capabilities. Both the total time glancing away from the forward roadway and glance durations at the vehicle's touchscreen increased in L2 mode. Based on these findings and the available literature on drivers' use of ADS, we suggest government regulators review driver training policies to include specific modules aimed at instructing on the correct use of L2 systems. This could be achieved through a revision of the study material or by including a direct experience module wherein prospective drivers must show a correct use of the L2 system, or both. It is also arguable that a separate driver's license be designed for the use of L2 and higher-level ADS.

2. Training for sales/dealership staff at the point of sale.

The current study highlighted potential safety risks associated with operating L2 systems. Literature suggest that poor sales/dealership staff training may be a contributing factor to drivers' limited understanding of the capabilities and limitations of this technology, with customers often receiving limited or even inaccurate information about driver assistance systems at the point of sale [38]–[40]. It is recommended that the training for sales/car dealership staff be enhanced and/or standardized to include information on the use of L2 systems so that potential customers can receive comprehensive and accurate information at the point of sale.

3. Recording of crashes involving L2 systems.

In 2021, NHTSA issued a standing general order requiring manufacturers of L2 systems to report crashes involving vehicles equipped with L2 systems. A summary report containing data recorded over an 11 month period was issued in June of 2022**.** The report begins to shed light on the amount of real-world crashes involving L2 systems. Considering this effort by NHTSA, a similar effort could take place to further add to our understanding of the potential safety risks of L2 systems.

4. Additional research on longer-term driver use of L2 systems, and system functioning in unique weather and road conditions.

Whereas the current study investigated driver use of an L2 system over a 40-minute drive, it is recommended that more research be conducted that explores drivers' longer-term use of L2

systems. The literature suggests that drivers' interaction with vehicle automation morphs over time from a novelty phase wherein drivers explore the system's capabilities and limitations to an experienced phase where they tend to become over reliant toward the system. Likewise, whereas most of the research involving L2 systems take place in pristine weather and road conditions, little is known about how L2 systems behave and drivers use these systems in weather and road conditions that are typical in different regions (e.g., snow-covered roads, dirt roads or roads with faded lane markings). With this in mind, more research is necessary to develop a better understanding of how longer-term use of L2 systems and operating these systems in unique road and weather conditions affect driver behavior and road safety.

References

[1] SAE, “J3216: Taxonomy and Definitions for Terms Related to Cooperative Driving Automation for On-Road Motor Vehicles,” 2021.

[2] Statista, “Level 2-4 autonomous vehicle sales as a share of total vehicle sales in 2025 and 2030, by automation level,” 2022. [Online]. Available: https://www.statista.com/statistics/1230101/level-2-autonomous-vehicle-sales-worldwide-as-a-share-of-total-vehicle-shares-by-autonomous-vehicle-level/.

[3] International Organization for Standardization, “Detection-response task (DRT) for assessing attentional effects of cognitive load in driving, ISO/DIS 17488,” 2015.

[4] F. Biondi, J. Turrill, J. R. Coleman, J. M. Cooper, and D. L. Strayer, “Cognitive distraction impairs drivers’ anticipatory glances: an on-road study.,” *Proc. Eighth Int. Driv. Symp. Hum. Factors Driv. Assessment, Train. Veh. Des. distractive*, pp. 23–29, 2015.

[5] R. Rossi, M. Gastaldi, F. Biondi, and C. Mulatti, “Evaluating the Impact of Processing Spoken Words on Driving,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2321, no. 1, pp. 66–72, Dec. 2012, doi: 10.3141/2321-09.

[6] P. a Hancock, M. Lesch, and L. Simmons, “The distraction effects of phone use during a crucial driving maneuver.,” *Accid. Anal. Prev.*, vol. 35, no. 4, pp. 501–514, Jul. 2003.

[7] D. J. Saxby, G. Matthews, J. S. Warm, E. M. Hitchcock, and C. Neubauer, “Active and passive fatigue in simulated driving: Discriminating styles of workload regulation and their safety impacts,” *J. Exp. Psychol. Appl.*, vol. 19, no. 4, pp. 287–300, 2013, doi: 10.1037/a0034386.

[8] E. Petridou and M. Moustaki, “Human factors in the causation of road traffic crashes,” *Eur. J. Epidemiol.*, vol. 16, no. 9, pp. 819–826, 2000, doi: 10.1023/A:1007649804201.

[9] M. R. Endsley, “From Here to Autonomy: Lessons Learned from Human-Automation Research,” *Hum. Factors*, vol. 59, no. 1, pp. 5–27, 2017, doi: 10.1177/0018720816681350.

[10] R. Molloy and R. Parasuraman, “Monitoring an Automated System for a Single Failure: Vigilance and Task Complexity Effects,” *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 38, no. 2, pp. 311–322, 1996, doi: 10.1177/001872089606380211.

[11] F. Biondi, I. Alvarez, K. Jeong, F. Biondi, I. Alvarez, and K. Jeong, “Human - System Cooperation in Automated Driving,” *Int. J. Human–Computer Interact.*, vol. 00, no. 00, pp. 1–2, 2019, doi: 10.1080/10447318.2018.1561793.

[12] D. L. Strayer, D. Getty, F. Biondi, and J. M. Cooper, “The Multitasking Motorist and the Attention Economy,” in *Human Capacity in the Attention Economy*, S. M. Lane and P. Atchley, Eds. APA Press, 2020.

[13] NTSB, “Collision Between Car Operating with Partial Driving Automation and Truck-Tractor Semitrailer Delray Beach, Florida, March 1, 2019 HWY19FH008,” 2020.

[14] NTSB, “Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator, HWY18FH011, Mountain View, California,” pp. 1–9, 2020.

[15] NHTSA, “Summary Report : Standing General Order on Crash Reporting for Level 2 Advanced Driver Assistance Systems,” 2022.

[16] E. T. Greenlee, P. R. DeLucia, and D. C. Newton, “Driver Vigilance Decrement is More Severe During Automated Driving than Manual Driving,” *Hum. Factors*, 2022, doi: 10.1177/00187208221103922.

[17] G. Lu, J. Zhai, P. Li, F. Chen, and L. Liang, “Measuring drivers’ takeover performance in varying levels of automation: Considering the influence of cognitive secondary task,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 82, no. May, pp. 96–110, 2021, doi: 10.1016/j.trf.2021.08.005.

[18] A. Eriksson and N. A. Stanton, “Takeover Time in Highly Automated Vehicles: Noncritical

Transitions to and From Manual Control," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 59, no. 4, pp. 689–705, 2017, doi: 10.1177/0018720816685832.

[19] R. Rossi, M. Gastaldi, G. Gecchele, F. Biondi, and C. Mulatti, "Traffic-Calming Measures Affecting Perceived Speed in Approaching Bends On-Field Validated Virtual Environment," 2020, doi: 10.3141/2434-05.

[20] D. L. Strayer, J. Turrill, J. M. Cooper, J. R. Coleman, N. Medeiros-Ward, and F. Biondi, "Assessing Cognitive Distraction in the Automobile.," *Hum. Factors*, vol. 57, no. 8, pp. 1300–1324, 2015, doi: 10.1177/0018720815575149.

[21] J. Gaspar and C. Carney, "The Effect of Partial Automation on Driver Attention: A Naturalistic Driving Study," *Hum. Factors J. Hum. Factors Ergon. Soc.*, p. 001872081983631, 2019, doi: 10.1177/0018720819836310.

[22] F. N. Biondi, M. Lohani, R. Hopman, S. Mills, J. M. Cooper, and D. L. Strayer, "80 MPH and out-of-the-loop : Effects of real-world semi-automated driving on driver workload and arousal .," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, pp. 1878–1882, 2018, doi: https://doi.org/10.1177/1541931218621427.

[23] A. S. McDonnell, T. G. Simmons, G. G. Erickson, M. Lohani, J. M. Cooper, and D. L. Strayer, "This Is Your Brain on Autopilot: Neural Indices of Driver Workload and Engagement During Partial Vehicle Automation," *Hum. Factors*, 2021, doi: 10.1177/00187208211039091.

[24] M. Lohani *et al.*, "No Difference in Arousal or Cognitive Demands Between Manual and Partially Automated Driving: A Multi-Method On-Road Study," *Front. Neurosci.*, vol. 15, no. June, pp. 1–12, 2021, doi: 10.3389/fnins.2021.577418.

[25] N. Dunn, T. Dingus, and S. Soccolich, "Understanding the Impact of Technology : Do Advanced Driver Assistance and Semi- Automated Vehicle Systems Lead to Improper Driving Behavior ?," 2019.

[26] M. Lohani, B. R. Payne, and D. L. Strayer, "A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving," *Front. Hum. Neurosci.*, vol. 13, no. March, pp. 1–27, 2019, doi: 10.3389/fnhum.2019.00057.

[27] S. G. Hart and L. E. Staveland, "Development of Nasa TXL (Task Load Index): Results of Empirical and Theoretical Reseach," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds. Amsterdam: North Holland Press, 1988, pp. 239–250.

[28] B. Saberi, J. A. Cort, F. Graf, and F. N. Biondi, "A Cognitive Workload Toolkit for Ergonomic Professionals," *Hum. Factors Ergon. Soc. Annu. Meet. Proc.*, 2021.

[29] F. N. Biondi, F. Graf, and J. Cort, "On the potential of pupil size as a metric of physical fatigue during a repeated handle push/pull task," *Appl. Ergon.*, vol. 110, no. March, p. 104025, 2023, doi: 10.1016/j.apergo.2023.104025.

[30] F. N. Biondi, B. Saberi, F. Graf, J. Cort, and P. Pillai, "Distracted worker : Using pupil size and blink rate to detect cognitive load during manufacturing tasks," *Appl. Ergon.*, vol. 106, no. August 2022, p. 103867, 2023, doi: 10.1016/j.apergo.2022.103867.

[31] F. N. Biondi, B. Balasingam, and P. Ayare, "On the Cost of Detection Response Task Performance on Cognitive Load," *Hum. Factors J. Hum. Factors Ergon. Soc.*, 2020, doi: 10.1177/0018720820931628.

[32] M. E. Kret and E. E. Sjak-Shie, "Preprocessing pupil size data: Guidelines and code," *Behav. Res. Methods*, vol. 51, no. 3, pp. 1336–1342, 2019, doi: 10.3758/s13428-018-1075-y.

[33] L. Held and M. Ott, "On p-Values and Bayes Factors," *Annu. Rev. Stat. Its Appl.*, vol. 5, pp. 393–419, 2018, doi: 10.1146/annurev-statistics-031017-100307.

[34] D. S. Quintana and D. R. Williams, "Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using JASP," *BMC Psychiatry*, vol. 18, no. 1, pp. 1–8, 2018, doi: 10.1186/s12888-018-1761-4.

[35] Z. Dienes, “Using Bayes to get the most out of non-significant results,” *Front. Psychol.*, vol. 5, no. July, pp. 1–17, 2014, doi: 10.3389/fpsyg.2014.00781.
[36] J. S. Racine, “RStudio: A Platform-Independent IDE FOR R And Sweave,” *J. Appl. Econom.*, vol. 27, no. 1, pp. 167–172, 2012.
[37] D. L. Strayer *et al.*, “Driver’ s Arousal and Workload Under Partial Vehicle Automation,” 2020.
[38] H. Abraham, B. Reimer, and B. Mehler, “Learning to use in-vehicle technologies: Consumer preferences and effects on understanding,” *Proc. Hum. Factors Ergon. Soc.*, vol. 3, pp. 1589–1593, 2018, doi: 10.1177/1541931218621359.
[39] A. Boelhouwer, A. P. van den Beukel, M. C. van der Voort, C. Hottentot, R. Q. de Wit, and M. H. Martens, “How are car buyers and car sellers currently informed about ADAS? An investigation among drivers and car sellers in the Netherlands,” *Transp. Res. Interdiscip. Perspect.*, vol. 4, p. 100103, 2020, doi: 10.1016/j.trip.2020.100103.
[40] A. McDonald, C. Carney, and D. V McGehee, “Vehicle Owners ’ Experiences with and Reactions to Advanced Driver Assistance Systems,” no. September, 2018.