

# Driver State Monitoring: Manipulating Reliability Expectations in Simulated Automated Driving Scenarios

Jaume R. Perello-March<sup>1</sup>, Christopher G. Burns, Roger Woodman<sup>2</sup>, Mark T. Elliott<sup>3</sup>, and Stewart A. Birrell<sup>4</sup>

**Abstract**—Highly Automated Driving technology will be facing major challenges before being pervasively integrated across production vehicles. One of them will be monitoring drivers' state and determining whether they are ready to take over control under certain circumstances. Thus, we have explored their physiological responses and the effects on trust of different scenarios with varying traffic complexity in a driving simulator. Using a mixed repeated measures design, twenty-seven participants were divided in two reliability groups with opposite induced automation reliability expectations -low and high-. We hypothesized that expectations would modulate participants' trust in automation, and consequently, their physiological responses across different scenarios. That is, increasing traffic complexity would also increase participants' arousal, and this would be accentuated or mitigated by automation reliability expectations. Although reliability group differences could not be observed, our results show an increase of physiological activation within high complexity driving conditions (i.e., a mentally demanding non-driving related task and urban scenarios). In addition, we observed a modulation of trust in automation according to the group expectations delivered. These findings provide a background methodology from which further research in driver monitoring systems can benefit and be used to train machine learning methods to classify drivers' state in changing scenarios. This would potentially help mitigate inappropriate take-overs, calibrate trust and increase users' comfort and safety in future Highly Automated Vehicles.

**Index Terms**—Driver state monitoring, highly automated driving, take-over request, trust in automation.

## I. INTRODUCTION

**I**NCLUSION of highly automated driving (HAD) capability -SAE Level 4- [1] in future vehicles will entail a dramatic change in task allocation whilst driving. Henceforth, under certain scenarios, manual control will not be required and users will be able to engage in Non-Driving Related Tasks (NDRTs) [2]. However, there are situations where the conditions for autonomous control of the vehicle are not met and the

Manuscript received June 22, 2020; revised October 3, 2020; accepted December 17, 2020. Date of publication January 22, 2021; date of current version May 31, 2022. This work was supported by WMG, The University of Warwick. The Associate Editor for this article was M. Brackstone. (Corresponding author: Jaume R. Perello-March.)

Jaume R. Perello-March, Christopher G. Burns, Roger Woodman, and Mark T. Elliott are with WMG, The University of Warwick, Coventry CV4 7AL, U.K. (e-mail: jaume.perello-march@warwick.ac.uk; c.burns.2@warwick.ac.uk; r.woodman@warwick.ac.uk; m.t.elliott@warwick.ac.uk).

Stewart A. Birrell is with the National Transport Design Centre, Coventry University, Coventry CV1 5FB, U.K. (e-mail: stewart.birrell@coventry.ac.uk). Digital Object Identifier 10.1109/TITS.2021.3050518

human driver must take over control [1]. In this sense, the next generation of Driver State Monitoring (DSM) systems may need to adapt to drivers' temporary disengagement, position movements or engagement in NDRTs, but also be able to monitor the driver's state before, during and after the take-over process. Recently, [2] proposed the concept of *Driver Availability* in automated driving as a model to determine drivers' availability to resume manual control safely in real time. These authors state that traffic complexity, automation capability and NDRTs will determine the driver state required for each circumstance. This means that a wide range of take-over situations may occur, and therefore a constant monitoring of driver's physiology can provide real-time information of driver's availability to take control safely. When engaged with NDRTs, the current driver state will need to change to a target state of the driver available of taking control. Arousal levels and motivational conditions will modulate this process. Accordingly, certain NDRTs may facilitate the takeover, preventing drivers' drowsiness by maintaining a suitable arousal level before take-over. Therefore, DSM systems should also be able to detect when a driver is sleepy or fatigued and cannot take control [3]. Appropriate arousal levels will be necessary for an optimal transition performance. Thus, classifying individual physiological states for each context can provide useful knowledge to train machine-learning classifiers determine the suitable driver state required before resuming control safely under different scenarios.

## II. BACKGROUND

Relatedly, a driving simulator study, [4], recorded drivers' prefrontal hemodynamic responses, gaze behaviour, heart rate and skin responses across several road layouts of changing complexity. They observed that complex road layouts (i.e., city centre and suburbs) were associated with increased physiological activity compared to a dual-carriage way and interurban road. Thus, urban scenarios resulted in an increase of oxygenated haemoglobin concentration, skin conductance level, skin conductance responses, horizontal spread of search and a decrease of fixation duration. Similar findings were observed in a naturalistic driving study monitoring drivers stress using electrocardiogram, electromyogram, skin conductance, and respiration [5]. These authors observed an increase of stress measures during high traffic density and urban scenarios.

More recently, [6] highlighted the potential of using drivers' physiological changes to determine the quality of a take-over request in a SAE Level-3 driving simulator study. Results showed that increased pupil diameter and heart rate while performing a NDRT were predictors of response time and, consequently, the quality of the takeover. In a naturalistic manual-driving study, [7] found a reduction of long off-road glances and an increase in time spent looking at the road during turns and upcoming traffic both with and without a NDRT. [8] used electrocardiogram (ECG) to measure mental workload fluctuations across different simulated driving contexts with a traffic congestion assistant (SAE Level 2). They observed a reduced heart rate while driving with the traffic jam assistant engaged; however, heart rate variability (feature not reported) increased when taking-over manual control after the traffic jam. According to the authors, this suggests that mental workload decreased when driving with the assistant engaged but also when resuming control after the traffic jam; they also observed a workload increase when driving with fog. Similar findings were observed by [9] in a driving simulator study combining different road layouts and including an autonomous driving scenario. The authors highlighted that higher driving complexity scenarios reported the highest mental workload levels, as shown by a reduced HRV (LF/HF ratio and RMSSD) and increased heart rate and skin conductance levels compared to the less demanding scenarios –including an autonomous drive with moderate traffic. Although, such stress and workload assumptions based on HRV should be cautiously taken. The general consensus is that LF/HF ratio increases for high stress, but it is a controversial and unreliable measure [10]–[12]. A recent review of stress and HRV states that whereas lower RMSSD relates to higher stress, for LF/HF Ratio a higher value relates to higher stress [10]. Although [13] suggest that there could be some exceptions, an example being attentional tasks, which would explain the opposite results observed in [8], [9].

Whilst it has been considered that eye-trackers are currently the most suitable technique to monitor driver's state for SAE Level 3 vehicles [14], it could be argued that daydreaming, sleeping, reading or a changed seat position could be their major drawback in SAE Level 4 vehicles. Certain wearable devices capable of cardiac and skin conductance measurements are offering a potential complement to eye-tracking for driver state monitoring [9], [15]. These measures have proven their validity in monitoring drivers' stress and mental workload both in naturalistic [5] and simulated driving [9], [15]–[18] studies. A series of driving simulator studies, [16], [17], evaluated HR, HRV, blink ratio, pupil diameter, body motion and SCL variations of drivers reporting discomfort when facing several complex and uncertain situations under manual or automated driving conditions. In [16], they observed a decreased HR during discomfort periods, that returned to the prior level approximately 5 s after the reported discomfort. HRV measured by the RMSSD showed a u-shaped tendency decreasing during the discomfort intervals. In their follow-up study [17], they registered similar findings with HR decreasing during uncomfortable situations. Finally, [18] followed a similar study design evaluating drivers' discomfort under

different vehicle controllers across several scenarios varying in traffic complexity and layout. They reported a significant reduction of RMSSD and increased HR under manual driving compared to the less arousing HAD conditions. In addition, they registered SCRs per minute and found an increase of SCRs/min during manual drive in line with those reported by HRV, but also during rural scenario, and interaction effects for manual drive in rural environment. Suggesting not only that manual driving in rural scenarios generated the highest arousal among all variables, but also that compared to HRV, SCR values were more sensitive to continuous changes across varying scenarios.

However, to the authors' knowledge, it remains unclear how traffic complexity and NDRT engagement would affect psychophysiology during HAD and therefore, drivers' readiness to take-over under certain scenarios. Only a few studies have approached the topic of take-over under different traffic complexity scenarios in HAD. [19] evidenced the negative effects of mental NDRTs and high traffic density on time to take-over and increasing the number of collisions during a simulator study. Similar results were reported by [20] when examining the role of traffic density on take overs from HAVs. [21] reported that high traffic urban scenarios might have adverse effects on subjectively reported drivers' emotional state and attitudes, and therefore motivating take-over behaviours. Overall, these findings suggest that high traffic density, complex road scenarios and engagement in mentally demanding NDRTs may have a detrimental effect on take-over quality during HAD, thus, supporting the use of psychophysiological measures to monitor drivers' state and prepare the driver for the optimal state to take over.

### III. OBJECTIVES

The present research aims to explore the physiological effects of HAD scenarios, using increasing traffic complexity and a mentally demanding NDRT. A driving simulator study was designed to provide a variety of driving scenarios (highway, interurban, urban and a hazardous risky manoeuvre), traffic density and complexity (see Procedure section) According to previous studies [4], [5], [8], [9], it could be expected that driving through low traffic density scenarios would be less arousing than high traffic density environments. The lack of new stimuli and external stressors in a low complexity scenario (i.e., a highway with flowing traffic) would potentially generate a low arousal state when the user is not involved in any task as habituation. Otherwise, an urban and complex scenario with dense traffic, junctions, pedestrians (including those crossing inappropriately) bus stops, cyclists and emergency vehicles should generate higher arousal just because of new stimulus constantly appearing unexpectedly. Finally, given that risk tolerance has been considered to influence driver's state and consequently the transition process [22], including a risky manoeuvre as the ending scenario should provide insights on how drivers perceive and react physiologically to risk after developing some experience with the HAD system.

A mentally engaging NDRT was included to the highway low traffic density scenario - the verbal 2-back variant of the n-back task. This task has been validated to generate mental workload derived from verbal working memory [19], [23]–[25]. Whereas the task may lack ecological validity, it can be argued that it involves verbal working memory and therefore relates to ecological tasks such as a phone conversation or talking to other passengers, but in a controlled and standardised manner. In addition, previous literature has also explored the effect of n-back tasks on physiological measures [23], [26] validating its use in this study. It could be expected that performing the 2-back task during a highway low traffic density condition should generate mental workload [25], and consequently higher physiological activity [11], [27] compared to the same condition with no NDRT. What remains unclear is whether this situation would be comparable to that arousal generated in a high complexity situation without NDRT.

The present study focuses on electrocardiogram (ECG) and electro-dermal activity (EDA) measures as these sensors are well established and validated for monitoring psychophysiological states in Human-Computer Interaction domains [11]. The following hypotheses have been proposed:

- H1: The urban high complexity scenario will produce more arousal than pre-drive, highway, interurban and urban low complexity, resulting in faster cardiac activity (higher heart rate, LF/HF ratio and lower RMSSD) and greater EDA (more skin conductance responses associated with greater amplitudes and magnitudes) than the other conditions.

- H2: The 2-back scenario will generate comparable arousal (in line with H1) to that of the urban high complexity condition and, consequently, more arousal than pre-drive, highway, interurban and urban low complexity.

- H3: The risk scenario will produce more arousal (in line with H1) than urban high complexity and 2-back task.

Besides physiological measures, the Trust in Automated Systems Scale [28] was included to investigate if motivational aspects may affect the reliance behaviour of taking control [2]. Trust, as an attitude towards the system [29], can either motivate reliance behaviours or ignoring the take-over request due to mistrust of the HAD system [30]. A real-world example occurred recently when the operator of a HAD test-vehicle took over inappropriately causing an accident, likely due to system distrust [31]. It is important to note that trust and distrust comprise two distinct yet related concepts, and that one is not the opposite of the other [32], [33]. Thus, observing how trust and distrust fluctuate among our scenarios could potentially complement physiology. Based on previous literature [34], [35], we induced two opposite automation reliability expectations (high and low reliability) which served as a grouping variable. Accordingly, the following hypotheses were proposed:

- H4: The low reliability group would display more physiological arousal (see H1) than the high reliability group, and particularly during urban high complexity and risk scenarios.

- H5: The low reliability group will display lower trust, lower total scores and higher distrust than the high reliability group.

- H6: Distrust scores would increase for the low reliability expectations group across the experiment, while trust and total scores would decrease.

- H7: Trust and total scores would increase across the experiment for the high reliability expectations group, while distrust would decrease.

## IV. METHOD

### A. Participants

Twenty-seven participants were recruited to take part in this study (20 male and 7 female). All of them held a UK-EU driving license. Participants were recruited within the University of Warwick (UK) and included undergraduate students, postgraduate students, university staff and other professionals. Recruitment and data collection methods received approval from the Biomedical and Scientific Research Ethics Committee from the University of Warwick. Participants voluntarily agreed to take part in this experiment and were free to withdraw at any point. All of them received a £10 voucher after the experiment.

Participants were divided into two groups (high and low) of automation reliability expectations, as it was also intended to explore the effects of expectations on trust in automation along with the changing complexity of the driving environments. The low group (N = 12) was told that the vehicle was running a prototype HAD system capable of self-driving and adapting road conditions, although it was not fully reliable yet, since it was still under development. The high group (N = 15) was told that they were testing a fully reliable HAD system, capable of driving through any scenario and adjust to all road conditions effectively. However, vehicle-driving performance was exactly equal for both groups across all driving conditions. Both groups were told to not take control of the vehicle under any circumstances to generate vulnerability.

### B. Apparatus







This study was conducted using WMG's 3xD driving simulator, at the University of Warwick. The 3xD is a fixed-base high-fidelity driving simulator equipped with a full body Range Rover Evoque and 8 projectors generating a 360° image, projected into a cylindrical screen 8m in diameter and 3m high (Fig. 1). The simulated vehicle automation is capable of lateral and longitudinal control, adapting to speed limits, queuing leading vehicles, keeping safety distance, emergency braking and overtaking slower/stopped vehicles. The simulation also generated road motion vibration and environmental sound.

ECG and EDA measures were recorded using a BIOPAC MP160 with wearable remote Bio-Nomadix amplifiers [36], [37]. The MP160 base-station was mounted behind the driver's seat inside the simulator in order to achieve the best quality signal. Three ECG electrodes were placed following a 3-lead configuration on the participant's torso. The EDA device comprised two electrodes on the medial phalanx region on the first and second fingers of the participant's non-dominant hand to minimise movement artefacts.



Fig. 1. WMG 3xD driving simulator.

TABLE I  
SCENARIOS DESCRIPTION

<i>2 - Back</i>	<i>Highway</i>	<i>Interurban</i>	<i>Urban L. C.</i>	<i>Urban H. C.</i>	<i>Risk</i>
					
Sunny	Sunny	Cloudy	Cloudy	Rainy	Rainy
Low traffic density	Low traffic density	Medium traffic density	Medium traffic density	High traffic density	High traffic density
60 to 80mph	60 to 80mph	30 to 50 mph	Up to 30 mph	Up to 30 mph	Up to 30 mph
Highway	Highway	Dual-carriage way	City	City	City

Boxes include driving conditions details.

Self-reported measures included the Trust in Automated Systems Scale [28]. This scale includes 12 items in a 7-likert rating scale. Items 1 to 5 assess distrust and items 6 to 12 assess trust. A total score can also be obtained inverting those items corresponding to distrust. The scale has been widely used in research as a subjective measurement of operator's willingness to trust in an autonomous teammate in the military domain [38], propensity to trust automated vehicles [39], [40] or in adaptable automation environments [41].

### C. Driving Scenarios

Highway scenario was a relatively straight, triple-lane road, with high speed limits of 60 to 80mph and opposite traffic separated by a central reservation. Traffic density was bidirectional, low and regular, so no braking or overtaking was needed. Scenario included relatively few signs –including overhead gantries-, and no pedestrians, pedal cyclists nor buildings along the roadside. Weather conditions were set to sunny and clear. Details concerning all scenarios can be found in Table I.

Interurban scenario carried traffic to and from the highway to suburbs and city centre in a straight line with two roundabouts, two lanes per way separated by a central reservation. Speed was limited to 30 to 50mph, and medium levels of oncoming traffic. Weather conditions changed to cloudy.

Urban low complexity scenario began within a suburb layout defined as two lanes passing through residential areas at a 30mph limit including several left and rights turns, give-ways and with a medium volume of oncoming traffic, pedestrians, cyclists and parked cars on the roadside.

Urban high complexity passed through the city centre and the surrounding area with commercial buildings, signs, billboards and the highest levels of moving and parked vehicles (i.e., vans, motorcycles, buses, trucks and emergency vehicles) and pedestrians compared to the other scenarios. Some vehicles were parked in driveways; others were parked on the street, and buses waiting at a bus stop with pedestrians running to them whilst inappropriately crossing the street. Speed limit was 30mph and the participant's automated vehicle had to overtake these stopped vehicles with traffic approaching ahead and deal with T-junctions with traffic approaching from both directions. Additionally, the simulated weather conditions shifted to heavy rain, degrading the visual range.

The risk scenario occurred within the urban high complexity environment and involved the HAV following a van which, immediately after a left bend, both encountered a cyclist and proceeded to overtake while approaching to a junction with right-of-way. Immediately after the van passed the junction, and while our HAV was overtaking the cyclist, an ambulance with emergency lights and siren moves into view at high speed from the left side of the junction. Our vehicle had to perform an emergency braking and evasive manoeuvre to avoid crashing against the ambulance, and immediately after, a police vehicle followed the ambulance, so the HAV had to brake again. Overall, this event lasted for approximately 60 seconds and was designed to explore whether the initial grouping expectations (i.e., high expectations of performance safety vs. low expectations) would affect trust scores and generate different arousal responses between groups and within conditions.

### D. Procedure

Upon their arrival, participants were guided into the simulator control room. Room temperature was set at  $21 \pm 2^\circ\text{C}$ . Participants were briefed on lab safety procedures and advised to follow the experimenter's instructions at all times. Consent forms and demographics questionnaires were filled in the week before the trial. Once all sensors were connected, approximately five minutes were allowed for electrodes to stabilise before data recording began. During this time, participants were instructed to be particularly careful in not applying any pressure to the sensors or stretching the cables in order to avoid signal spikes and artefacts. Following this, participants were briefed on the 2-back task and performed a short practice session. After the 2-back training, the data telemetry from the wearable amplifiers were checked to ensure good quality data acquisition immediately before participants were guided inside the driving simulator and asked to remain seated in the driver's seat.

Participants were informed that the experiment would start recording their physiological states pre-drive for 2 minutes, followed by a familiarization manual drive task. They were instructed to drive cautiously to gain familiarity and up to

20 mph, respecting normal UK Highway Code rules. The vehicle had an automatic gearbox, so they only had to use accelerator, brakes and steering wheel.

The drive began with a very simple manual drive across a countryside area with light traffic density taking approximately 5 minutes to complete. This acted as a familiarization run to minimize the impact of motion sickness. This scenario led to a roundabout which connected to a highway, and in which participants had to engage automated driving after hearing an audio cue by pressing a button on the centre console. This was also explained to them previously during the vehicle controls description.

Once the HAD was engaged the experimental scenarios began. After two minutes, participants heard an audio cue announcing they were about to perform a 2-back task lasting four sets of 30 seconds each. This was the first experimental condition. Instructions about the task were provided again by the same audio file. After performing the 2-back, the highway HAD scenario continued for five more minutes until reaching a highway exit. A two-minute epoch was then extracted for the second experimental condition, namely highway scenario. The vehicle stopped at a red traffic light in a roundabout, which led to the interurban scenario.

At this point, the simulation paused as longer exposures to driving simulator tend to increase the risk of simulator sickness [42]. Participants left the vehicle and went into the control room to fill in trust questionnaires. Sensors were again checked before resuming the scenarios.

Upon resuming, the scenario began from the same stopping point, leading to a fully autonomous interurban low complexity drive for 2 minutes. After this, the vehicle entered the urban low complexity scenario, where traffic complexity increased throughout the scenario.

The high complexity urban scenario ended with our highly automated vehicle (HAV) performing an evasive manoeuvre, which we referred to as the risk scenario. After this, participants left the driving simulator and filled in the trust scale.

### E. Analysis

One epoch per condition was extracted for data analysis, generating seven epochs. The seven conditions were, in order of occurrence: pre-drive, 2-back, highway, interurban, urban low complexity, urban high complexity and risky manoeuvre. For each of the first six conditions, epochs of 120 seconds in duration were extracted. For the risky manoeuvre, a shorter 60 second long epoch was extracted, due to the risk condition being event related. A full description of each condition is provided in the Procedure section (see Table I). Data were extracted using the automated data analysis routines from Biopac's ACQKnowledge software (CA, USA; version: 5.0.2).

ECG data were sampled at 2000Hz and filtered applying Biopac's recommendations using a band pass filter with a 35Hz high frequency cut-off and a low frequency cut-off at 0.5Hz. Features extracted comprised heart rate (beats per minute) and two typically used frequency and time-domain heart rate variability (HRV) parameters, namely

the LF/HF ratio and Root Mean Square Standard Deviation (RMSSD) [11], [43].

Raw EDA signals were sampled at 62.5Hz and low-pass filtered to a frequency cut-off fixed at 1Hz, following standardised guidelines [44], [45]. Phasic EDA signals were extracted using a high pass filter at 0.05Hz; the skin conductance response (SCR) threshold level was set at  $0.03 \mu\text{S}$  and the SCRs rejection rate to 10%. Features extracted included three phasic-derived SCRs: SCR count (the number of SCRs within each epoch), SCR amplitude (the delta value from the offset to the peak of the SCR obtained across all non-zero SCRs) and SCR magnitude (the delta values including non-responses). Raw SCR amplitude was obtained from the delta values of the SCRs reported with amplitudes below  $0.01 \mu\text{S}$  rejected for analysis based on standardised criteria [44], [46]. SCRs from all conditions were non-specific (NS-SCR) except for the Risk condition, where they were event-related (ER-SCR). The established common practice for normalising raw SCR amplitudes applying the square root transformation and raw SCR magnitudes using the Log + 1 transformation to correct for the presence of skewness and kurtosis were applied [45], [47], [48]. All SCR data -including SCR count- was standardised for parametric statistical analysis to T-scores ( $m = 50$ ,  $SD = 10$ ), in order to allow inter-individual comparisons [44], [48]. Descriptive statistics used for the T-scoring were obtained separately within each individual to control for inter-individual variability. For example, instead of calculating the mean SCR amplitude based on all responses within an epoch, the amplitudes given for each individual across all epochs are those used to compute the individual mean amplitude and the standard deviation to obtain the T-score from a raw SCR amplitude [48]. This is a current and standardised common practice with studies reporting SCR data [49], [50].

Finally, trust was reported three times across the study: pre-experimental stage (before grouping expectations on vehicle reliability were provided), during the pause at the end of the highway driving, and at the end of the study (see Fig. 2).

## V. RESULTS

This study explored the effect of complexity-changing scenarios and a NDRT on driver physiology during simulated HAD scenarios. A mixed between-groups design with repeated measures was conducted - i.e., participants experienced the same stimuli but with their a priori expectations differing according to group. Six epochs of 120 seconds were selected across all conditions except the risk condition, which lasted 60 seconds as described above. Eight participants were excluded from the EDA analysis either because of missing data or substantial artefacts on the raw signal, with  $N = 19$  participants analysed. For ECG data, two participants were excluded from analysis due to missing data and excessive artefacts.

We hypothesised that pre-drive, highway, interurban and urban low complexity scenarios would generate lower arousal than urban high complexity and risk scenarios (H1 & H3),

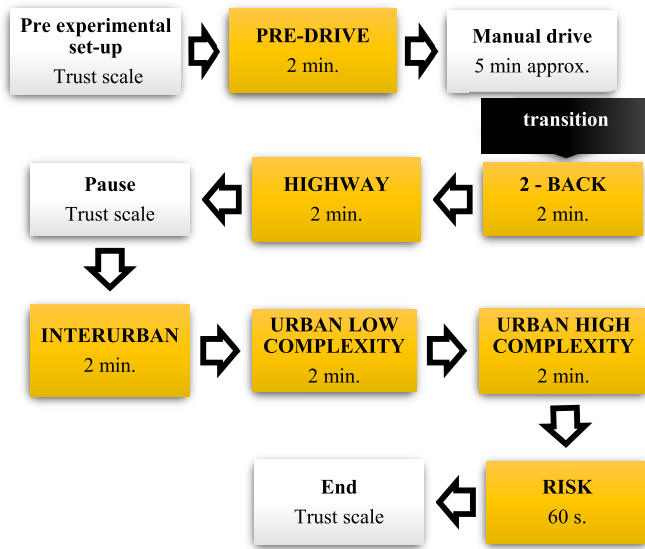


Fig. 2. Study layout. The orange shaded boxes represent the experimental conditions (epochs) recorded. No physiology data were recorded in the stages represented by the white boxes.

and that adding the 2-back task within a highway drive would generate similar arousal values than urban high complexity and risk scenarios without NDRT (H2). In addition, traffic complexity differences across conditions would also generate between groups differences derived from the reliability expectations given. These resulting in higher arousal for the low reliability group, particularly during urban high complexity and risk (H4). Expectations would also have a detrimental effect on trust scores on the low reliability group, and a positive effect on the high reliability group (H5, H6 and H7).

### A. Hypothesis 1

It was hypothesized that urban high complexity condition would be more arousing than pre-drive, highway, interurban and urban low complexity. The mixed repeated measures ANOVAs did not report any effects within driving conditions proving this assumption neither for ECG nor for EDA features. Nonetheless, cardiac mean values registered during urban high complexity scenario suggested that this condition elicited greater cardiac activity than interurban and urban low complexity, and even greater than pre-drive and highway for LFHF ratio (see Fig. 3). Aligned with these observations were all SCR features, where urban high complexity reported higher SCR count, amplitudes and magnitudes than highway, interurban and urban low complexity (see Fig.4). However, these means are not conclusive on their own so they will be compared with relevant literature in the next section.

### B. Hypothesis 2

This hypothesis tested whether performing a 2-Back task within a highway low complexity scenario would produce comparable arousal to that registered during urban high complexity and risk scenarios. Evidence supporting this hypothesis was found for all ECG and EDA features.

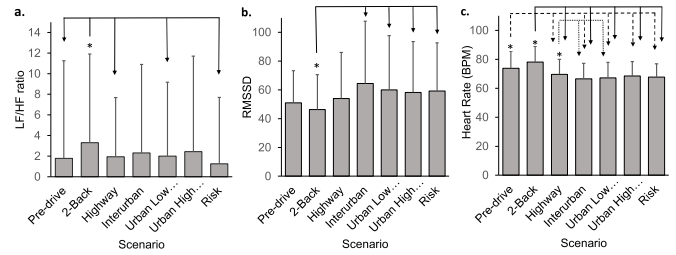


Fig. 3. Cardiac features of LF/HF ratio (a.), RMSSD (b.) and Heart Rate (c.) X-axis represents all experimental conditions in order of occurrence. Y-axis represents mean LF/HF ratio (a.), RMSSD (b.) and Heart Rate (c.) scores. Statistically significant effects are indicated with asterisks.

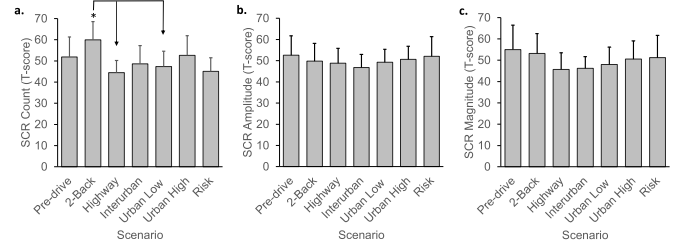


Fig. 4. SCR measures of count (a.), amplitude (b.) and magnitude (c.) for each scenario. X-axis represents all experimental conditions in order of occurrence. Y-axis represents mean T-scores.

Heart rate main effects ( $F(6,138) = 34.47, p < 0.001, \eta_p^2 = 0.600$ ; Fig. 3) were reported. Follow-up pairwise comparisons indicate that the 2-back condition generated significantly higher rate ( $m = 78.125, SD = 10.881$ ) than highway ( $m = 69.664, SD = 10.357, p < 0.001$ ), interurban ( $m = 66.501, SD = 10.705, p < 0.001$ ), urban low complexity ( $m = 67.154, SD = 10.725, p < 0.001$ ), urban high complexity ( $m = 68.556, SD = 9.907, p < 0.001$ ), and risk ( $m = 67.800, SD = 9.153, p < 0.001$ ). Similar findings were reported by LF/HF ratio ( $F(6,138) = 4.935, p = 0.003, \eta_p^2 = 0.177$ ; Fig. 3). Pairwise comparisons reveal a higher ratio during the 2-back condition ( $m = 3.297, SD = 3.093$ ) than pre-drive ( $m = 1.779, SD = 1.912, p < 0.001$ ), highway ( $m = 1.942, SD = 1.857, p = 0.001$ ), urban low complexity ( $m = 1.992, SD = 2.418, p < 0.001$ ), and risk ( $m = 1.251, SD = 1.032, p = 0.002$ ). RMSSD reported aligned main effects ( $F(6,138) = 7.048, p = 0.001, \eta_p^2 = 0.235$ ; Fig. 3). The 2-back scenario showed significantly lower time variability ( $m = 46.320, SD = 24.284$ ) relative to interurban ( $m = 64.3649, SD = 43.222, p = 0.001$ ), urban low complexity ( $m = 60.032, SD = 37.647, p = 0.002$ ), urban high complexity ( $m = 58.243, SD = 35.269, p = 0.001$ ) and risk ( $m = 59.243, SD = 33.413, p < 0.001$ ) (see Table II for a summary of all effects).

EDA data analysis also supported H2. SCR count varied across scenarios ( $F(6, 102) = 7.034, p < 0.001, \eta_p^2 = 0.293$ ; Fig. 4). Post-hoc comparisons showed the 2-back condition ( $m = 59.978, SD = 8.608$ ) resulted in a significantly higher SCR count than highway ( $m = 44.464, SD = 5.743, p < 0.001$ ) and urban low complexity ( $m = 47.381, SD = 7.183, p = 0.007$ ) scenarios. No effects were observed for SCR amplitudes and magnitudes.

### C. Hypothesis 3

We expected that the risk scenario would be more arousing than pre-drive, highway, interurban and urban low complexity.

TABLE II  
PHYSIOLOGY STATISTICALLY SIGNIFICANT RESULTS

Main findings			<i>Pre-drive</i>	<i>2-back</i>	<i>Highway</i>	<i>Interurban</i>	<i>Urban L.C.</i>	<i>Urban H.C.</i>	<i>Risk</i>
EDA N = 19	SCR count	Mean		<b>59.978</b>	44.464		47.381		
		(SD)		<b>(8.608)</b>	(5.743)		(7.183)		
		P value			< 0.001		= 0.007		
ECG N = 25	LF/HF	Mean	1.779	<b>3.297</b>	1.942		1.992		1.251
		(SD)	(1.912)	<b>(3.093)</b>	(1.857)		(2.418)		(1.032)
		P value	< 0.001		= 0.001		< 0.001		= 0.002
	RMSSD	Mean		<b>46.320</b>		64.365	60.032	58.243	59.243
		(SD)		<b>(24.284)</b>		(43.222)	(37.647)	(35.269)	(33.413)
		P value				= 0.001	= 0.002	= 0.001	< 0.001
	BPM	Mean	<b>73.849*<sup>1</sup></b>	<b>78.125*<sup>2</sup></b>	<b>69.664*<sup>3</sup></b>	66.501	67.154	68.556	67.800
		(SD)	<b>(11.790)</b>	<b>(10.881)</b>	<b>(10.357)</b>	(10.705)	(10.725)	(9.907)	(9.153)
		P value			= 0.001* <sup>1</sup> < 0.001* <sup>2</sup>	< 0.001 * <sup>1*2*3</sup>	< 0.001 * <sup>1*2</sup>	< 0.001 * <sup>1*2</sup>	< 0.001 * <sup>1*2</sup>

Bold values represent the highest means. Asterisks represent pairwise significant combinations.

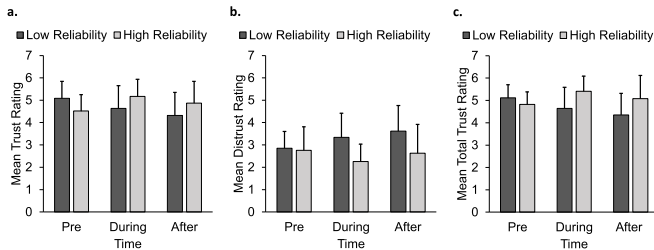


Fig. 5. Trust in automation scale ratings. Y-axis includes mean trust (a.), distrust (b.) and mean total trust (c.) ratings for both groups. X-axis represents time of administration, i.e. before any experimental manipulation (pre), during the experiment (during), and after the trial (after).

No statistical effects were observed supporting this hypothesis for either ECG or EDA features. However, the risk condition reported one of the highest mean SCR amplitudes and magnitudes among experimental conditions, only comparable with those observed during the 2-Back scenario (see Fig. 4). The relevance of this trend will be discussed in the next section.

D. Hypothesis 4

It was expected that reliability expectations as a grouping factor would generate different arousal levels between both groups. The lack of findings in this way will be further discussed in the next section.

E. Hypothesis 5

Aside from variations in arousal levels, reliability groups were also expected to generate opposite trust and distrust ratings. The trust in automated systems scale, consisting of 12 questions rated on a 7-point Likert scale [25], was included to explore the effect of reliability expectations on

participants’ trust and distrust across the different scenarios (N = 27). Trust subscale ratings reported interaction effects (F (2, 50) = 4.823, p = 0.012,  $\eta_p^2 = 0.162$ ), but these vanish after follow-up tests. Interaction effects were also observed for distrust subscale ratings (F (2, 50) = 4.961, p = 0.011,  $\eta_p^2 = 0.166$ ) indicating that the low reliability group (m = 3.333, SD = 1.086) developed more distrust than the high reliability group, p = 0.006), (m = 2.253, SD = 0.787) during the experiment. This trend remained after the experiment, evidencing the effect of induced low reliability expectations (m = 3.617, SD = 1.146) compared to the high reliability group (m = 2.627, SD = 1.289, p = 0.048) for distrust. These findings were aligned by the Total score (F (2, 50) = 6.136, p = 0.004,  $\eta_p^2 = 0.197$ ), highlighting the detrimental effect of low reliability expectations (m = 4.646, SD = 0.944) compared to high reliability expectations (m = 5.411, SD = 0.682, p = 0.022) during the study.

F. Hypothesis 6

Besides the expected differences between reliability groups, we also expected a detrimental effect on trust and increase on distrust for the low reliability group compared to the other group along the experiment, particularly after the exposure to urban high complexity and risk scenarios. This hypothesis was confirmed as distrust increased significantly within the low reliability group (F (2, 50) = 4.961, p = 0.011,  $\eta_p^2 = 0.166$ ). Particularly, before any expectations were given (m = 2.850, SD = 0.749) distrust increased compared to after the study (m = 3.617, SD = 1.146, p = 0.041). Similarly, Total trust scores also decreased for the low reliability group (F (2, 50) = 6.136, p = 0.004,  $\eta_p^2 = 0.197$ ) along the study. Although these effects vanish after post-hoc corrections.

### G. Hypothesis 7

On the other hand, opposite effects were expected within the high reliability group. This hypothesis was also confirmed as Total trust scores increased across the experiment for the high reliability group ( $F(2, 50) = 6.136, p = 0.004, \eta_p^2 = 0.197$ ). Particularly, total scores significantly increased during the experiment ( $m = 5.411, SD = .682$ ) compared to prior any expectations were delivered ( $m = 4.817, SD = 0.564, p = 0.031$ ).

### H. Other Findings

Unexpectedly, the highway scenario generated significantly higher heart rate ( $p < 0.001$ ) than interurban and urban low complexity; and the pre-drive ( $m = 73.849, SD = 11.790$ ) higher rate than all other experimental conditions, except for the 2-back (see Table II).

## VI. DISCUSSION

The present study explored the effects of different traffic-complexity scenarios and a Non-Driving Related Task (NDRT) on Highly Automated Vehicle (HAV) users' psychophysiology. It was expected that higher traffic complexity would generate higher arousal levels and that adding a NDRT to a low complexity scenario would generate arousal levels comparable to that of a high complexity scenario. Besides, two opposing automation reliability expectations served as a grouping variable, expecting that the high reliability group would be less aroused than the low reliability group, particularly during the 2-back task, the urban high complexity and the risk scenarios. Finally, trust in automation was reported before, during and after the experiment in order to observe whether reliability expectations had any effect on self-reported trust and on arousal states.

Results partially support the initial hypotheses. Although H1 could not be proved, overall means reflected an increase of arousal during both urban and risk scenarios, registering the greatest SCR amplitudes and magnitudes, and the urban high complexity registering the second highest SCR count, aligned with those findings reported by [4], [5], [9], [18]. The lack of physiological effects here does not disregard the need to further explore the effects of changing traffic complexity scenarios for HAVs. It still needs to be explored whether the length and the order of the scenarios had a detrimental effect in physiology, or whether all the instructions, manual driving induction and an extremely mentally demanding NDRT did so.

H2 has been clearly supported as the 2-back scenario generated higher arousal than pre-drive (higher LF/HF ratio), highway (increased heart rate and LF/HF ratio, and more SCRs), interurban (increased heart rate and lower RMSSD) and urban low complexity scenarios (increased heart rate, LF/HF ratio and SCRs and reduced RMSSD). Our results agree with the wider literature that LF/HF ratio increases with greater stress [10]–[13], but appears to contradict other driver behaviour studies who have used this measure and found the opposite result under high stress driving scenarios [8], [9]. The LF/HF ratio is recognised to be a controversial measure,

with variations in studies showing both increases and decreases in value for high stress scenarios [13]. Therefore, RMSSD is likely to be a more robust indicator in this study. Our findings are in line with previous literature [23], [26], reporting that the 2-back task increases physiological activity as a result of mental workload during a simulated driving scenario. These results suggest that users involved in verbal working memory tasks –i.e., a demanding phone conversation or talking with passengers - during a low traffic complexity scenario may develop physiological levels comparable to an urban high traffic complexity scenario without carrying out any NDRT. Moreover, our results also agree with those from [16], [18] reporting the lowest RMSSD values and HR increasing during mentally demanding driving scenarios [18].

Regarding H3, there may be several reasons why the risk scenario did not report significant findings. Even though the HR values observed during risk condition are similar to those from [16], [17], who noted a decrease during discomfort periods, and that a similar scenario involving a cyclist and a lateral event was rated as highly risky in [35]. Perhaps our driving simulation was not realistic enough to generate arousal derived from risk perception. Henceforth, further research exploring risk perception in HAVs in driving simulator should design more dramatic -or more realistic- scenarios to induce observable physiological responses. However, SCR amplitude and magnitude mean values reported during risk scenario were the highest among the experimental conditions, only comparable to the 2-back scenario, which suggest that urban scenarios lead to a progressive arousal increase, where further significant increases were not observed when the risk event occurred. This might present a problem for the implementation of DSM systems, where changes in arousal levels from low to moderate or low to high can be detected physiologically, but not from moderate to high arousal scenarios. Certainly, [17] noticed a similar issue but in the opposite direction as: “physiological reactions could be observed for situations with specific events that provoke moderate to high discomfort”. [...] Longer lasting and slowly evolving situations with moderate to low reported discomfort did not show associated changes in physiology and can therefore hardly be detected by these parameters.” pp. 454-455. Thus, perhaps this was due to the nature of their event-related design in contrast to our longer lasting and slowly evolving situations, with the exception of the risk event which, noticeably, generated a similar outcome.

Hence, machine-learning methods could provide a potential solution to overcome these issues. Otherwise, the limitations of ECG and EDA methods could also be supplemented with eye-tracking techniques. This would also explain why H4 could not be accepted as no group differences were observed across physiological data.

Even though no physiological differences were observed between reliability groups, self-reports evidenced that the low reliability group developed more distrust, and the high reliability group higher total trust scores, confirming H5. In addition, H6 and H7 were clearly supported by the fact that distrust significantly arose within the low reliability group during the study, and that total trust scores significantly increased within the high reliability group. Henceforth, it could be assumed that



our reliability expectations had an effect, but prior expectations or knowledge may become irrelevant in highly arousing events like the risk scenario. Another explanation could be the way expectations were given was not sufficient to elicit observable physiological differences between reliability groups and across all conditions. However, the method used was based on recent research, which successfully induced automation reliability expectations on their participants using introductory information [34], [35], and even successfully manipulated risk perceptions based on induced expectations [35]. We followed the qualitative method to generate expectations applied by [34], but expectations were delivered by the researcher instead of playing a video, hoping that the role of the experimenter would be enough to manipulate participants' beliefs. Perhaps quantitative methods are a better way to induce reliability expectations [35], but these results were only based on self-reported data, and do not necessarily apply to participants' physiology or behaviour, as suggested by [51]. Thus, further research should explore the correlation of induced expectations and risk perception on self-reports, physiology and behaviour under HAD conditions. Finally, it is worth mentioning the numerous similarities observed between the design and findings from [16]–[18] and our paper, for which we have reasons to believe that these psychophysiological findings attributed to discomfort in [16]–[18], could also be transferable to the Trust in Automation (TiA) literature, in particular to distrust:

First, [16], [18] state that discomfort could lead to safety-critical situations in automated driving, particularly due to unnecessary take-overs. In the TiA literature, unnecessary disengaging from an automated system in critical situations is known as *disuse*, and produced by distrust [52]–[54].

Second, the theoretical concepts associated with discomfort in [16]–[18] are extraordinarily related to those historically attributed to TiA literature [29], [55]. E.g., “the definition of comfort is rather broad, and shows similarities and overlap with related concepts of stress, mental workload, alertness, anxiety, fear, motion sickness or anger. [...] As the human role in automated driving shifts from active driver to user, additional psychological determinants of driving comfort are discussed, such as apparent safety, trust in the system, feelings of control, familiarity of driving manoeuvres, and information about system states and actions.” [17] p. 446.

Third, [16], [17] also instructed their participants to not take manual control, in order to induce driver vulnerability. Vulnerability has been identified as a determinant factor of leading to trust/distrust [29], [56].

Fourth, [16]–[18] also assume that drivers' arousal will increase along the complexity and unpredictability of the situation, and the uncertainty about the vehicle capability to deal with the task. Uncertainty about the system capability is another key factor leading to trust/distrust [29], [56] as we have actually manipulated in this study with reliability expectations. Moreover, we have also hypothesised that distrust would be analogous and increase under stress situations, and we have observed an increase of distrust/decrease of trust after the experiment, along with the higher complexity conditions. Finally, yet importantly, the commonalities between our driving scenarios are equally remarkable. We all

have followed an almost identical approach creating traffic complexity through infrastructure-related factors (i.e. complex intersections, roundabouts, highway exists, etc.), unclear behaviours from other road users -some of them vulnerable such as children inappropriately crossing and bicycles-, or unpredictable behaviours from the ego vehicle like avoiding obstacles, overtaking buses at the bus-stop with traffic ahead. Likewise, we manipulated external factors like adverse weather conditions in some scenarios. Overall, this makes the results from all four studies very transferable and suggest that discomfort and distrust may manifest similarly on psychophysiology, or perhaps that the burdens between both constructs are not so clear.

Finally, an interesting issue arose with the pre-drive physiological measurements, which was the generally high levels of arousal when participants were seated in the simulator vehicle prior to the scenarios starting, which represented our baseline for comparison. This initial high arousal was most probably due to a combination of the Hawthorne effect (i.e., the overall novelty of the experience), and the amount of initial instructions delivered. However, the complexity of this experiment required detailed instructions and further research could consider longer baseline periods when measuring physiology in driving simulators to avoid similar problems. Recommendations for future work would be to extend the period of this ‘active baseline’ to around 10 minutes, or take baseline readings out of the simulator vehicle. However, in this particular study it was felt that a resting baseline would have led to an artificially increase state of arousal, due to the points stated above, for the first, and subsequent driving conditions. Hence, a recommendation from this paper would favour extending the ‘active baseline’ to at least 10 minutes rather than removing it.

Our findings have answered some questions but more importantly, have raised more. They represent another step towards the next generation of DSM systems for HAVs. These systems should be able to distinguish and interpret several physiological states derived from different contexts when assessing the appropriateness of the driver's state to take over. For example, the physiological activation generated by an emergency scenario or an engaging NDRT may induce false negatives as our DSM system could not differentiate between moderate and high arousal, and even false positives – i.e., based on estimated “appropriate” arousal levels, the system assumes the driver is ready to take over. The system may interpret that users are engaged, active and aware of the situation when they are just aroused because of an emergency vehicle passing by, but may not be aware of the whole driving situation as they were immersed in a NDRT. A related real-world situation already occurred with an experimental vehicle as a consequence of distrust [31], [57]. In this case, the operator was probably aroused and attentive, but not fully aware of the driving situation and definitely not able to take over. Therefore, future DSM systems should cope with such situations and avoid inappropriate takeover manoeuvres based on “the bigger picture”, which includes driver's psychophysiology, traffic context and vehicle capabilities. A first step could be identifying physiological patterns related to simple contextual variables as we did here. This knowledge would

allow applying machine-learning techniques to generate individual baselines for different driving contexts. Furthermore, infotainment systems could potentially benefit from these data, managing in-vehicle settings to increase occupants' comfort and safety, such as the Attentive User Interfaces proposed by [49].

## VII. CONCLUSION

To conclude, our findings evidence the potential advantages and limitations of EDA and ECG based DSM systems to detect and classify different driver states based on contextual changes and determine their readiness to take over control. Whereas these methods may be useful for detecting low to moderate, and low to high arousal fluctuations, detecting changes from moderate to high arousal may be their major drawback. This knowledge should provide the grounds for further multimodal machine learning-based DSM systems to classify drivers' states before take-over and monitor the transition process until reaching the optimal state to perform the transition successfully. In this process, trust in automation will be determinant and here we have demonstrated that prior expectations can calibrate it, but these may become irrelevant during "fight or flight" situations. Suggesting that in highly arousing situations, trust calibration -and subsequent reliance behaviour- would be mainly headed by affective rather than analytic or analogic processes [29]. Therefore, future DSM systems should be also capable to monitor drivers trust in automation based on physiological and behavioural data, in order to mitigate inappropriate take-overs or unsafe manual transitions.

## REFERENCES

- [1] *Taxonomy and Definitions for Terms Related to Driving Automation Systems for on-Road Motor Vehicles*, SAE, Warrendale, PA, USA, 2018.
- [2] C. Marberger, H. Mielenz, F. Naujoks, J. Radlmayr, K. Bengler, and B. Wandtner, "Understanding and applying the concept of 'driver availability' in automated driving," in *Advances in Human Aspects of Transportation (Advances in Intelligent Systems and Computing)*, vol. 597. Cham, Switzerland: Springer, 2018, pp. 595–605.
- [3] J. Wörle, B. Metz, C. Thiele, and G. Weller, "Detecting sleep in drivers during highly automated driving: The potential of physiological parameters," *IET Intell. Transp. Syst.*, vol. 13, no. 8, pp. 1241–1248, Aug. 2019.
- [4] H. J. Foy and P. Chapman, "Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation," *Appl. Ergonom.*, vol. 73, pp. 90–99, Nov. 2018.
- [5] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005.
- [6] M. T. Alrefaie, S. Summerskill, and T. W. Jackson, "In a heart beat: Using driver's physiological changes to determine the quality of a takeover in highly automated vehicles," *Accident Anal. Prevention*, vol. 131, pp. 180–190, Oct. 2019.
- [7] E. Tivesten and M. Dozza, "Driving context and visual-manual phone tasks influence glance behavior in naturalistic driving," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 26, pp. 258–272, Sep. 2014.
- [8] K. A. Brookhuis, C. J. G. van Driel, T. Hof, B. van Arem, and M. Hoedemaeker, "Driving with a congestion assistant: Mental workload and acceptance," *Appl. Ergonom.*, vol. 40, no. 6, pp. 1019–1025, Nov. 2009.
- [9] V. Melnicuk, S. Birrell, E. Crundall, and P. Jennings, "Employing consumer electronic devices in physiological and emotional evaluation of common driving activities," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1529–1534.
- [10] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Invest.*, vol. 15, no. 3, pp. 235–245, Mar. 2018.
- [11] B. Cowley *et al.*, "The psychophysiology primer: A guide to methods and a broad review with a focus on human-computer interaction," *Found. Trends Hum.-Comput. Interact.*, vol. 9, nos. 3–4, pp. 151–308, 2016.
- [12] Y. Kageyama, M. Odagaki, and H. Hosaka, "Wavelet analysis for quantification of mental stress stage by finger-tip photo-plethysmography," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2007, pp. 1846–1849.
- [13] G. G. Berntson and J. T. Cacioppo, "Heart rate variability: Stress and psychiatric conditions," in *Dynamic Electrocardiography*. Feb. 2004, pp. 57–64.
- [14] T. Hecht, A. Feldh, J. Radlmayr, Y. Nakano, Y. Miki, and C. Henle, "A review of driver state monitoring systems in the context of automated driving," in *Proc. 20th Congr. Int. Ergonom. Assoc. (IEA)*, vol. 1, 2019, pp. 398–408.
- [15] V. Melnicuk, S. Birrell, E. Crundall, and P. Jennings, "Towards hybrid driver state monitoring: Review, future perspectives and the role of consumer electronics," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2016, pp. 1392–1397.
- [16] M. Beggiato, F. Hartwich, and J. Krems, "Using smartbands, pupilometry and body motion to detect discomfort in automated driving," *Frontiers Hum. Neurosci.*, vol. 12, p. 338, Sep. 2018.
- [17] M. Beggiato, F. Hartwich, and J. Krems, "Physiological correlates of discomfort in automated driving," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 66, pp. 445–458, Oct. 2019.
- [18] V. Radhakrishnan *et al.*, "Measuring drivers' physiological response to different vehicle controllers in highly automated driving (HAD): Opportunities for establishing real-time values of driver discomfort," *Information*, vol. 11, no. 8, p. 390, Aug. 2020.
- [19] J. Radlmayr, C. Gold, L. Lorenz, M. Farid, and K. Bengler, "How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 2014, vol. 58, no. 1, pp. 2063–2067.
- [20] C. Gold, M. Körber, D. Lechner, and K. Bengler, "Taking over control from highly automated vehicles in complex traffic situations," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 58, no. 4, pp. 642–652, Jun. 2016.
- [21] F. Techer *et al.*, "Anger and highly automated driving in urban areas: The role of time pressure," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 64, pp. 353–360, Jul. 2019.
- [22] F. Naujoks, C. Purucker, and A. Neukum, "Secondary task engagement and vehicle automation—Comparing the effects of different automation levels in an on-road experiment," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 38, pp. 67–82, Apr. 2016.
- [23] J. K. Lenneman and R. W. Backs, "Cardiac autonomic control during simulated driving with a concurrent verbal working memory task," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 51, no. 3, pp. 404–418, Jun. 2009.
- [24] S. M. Ko and Y. G. Ji, "How we can measure the non-driving-task engagement in automated driving: Comparing flow experience and workload," *Appl. Ergonom.*, vol. 67, pp. 237–245, Feb. 2018.
- [25] A. Unni, K. Ihme, M. Jipp, and J. Rieger, "Corrigendum: Assessing the driver's current level of working memory load with high density functional near-infrared spectroscopy: A realistic driving simulator study," *Frontiers Hum. Neurosci.*, vol. 12, p. 498, Dec. 2018.
- [26] T. M. Gable, A. L. Kun, B. N. Walker, and R. J. Winton, "Comparing heart rate and pupil size as objective measures of workload in the driving context: Initial look," in *Proc. Adjunct Proc. 7th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, Sep. 2015, pp. 20–25.
- [27] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neurosci. Biobehav. Rev.*, vol. 44, pp. 58–75, Jul. 2014.
- [28] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cognit. Ergonom.*, vol. 4, no. 1, pp. 53–71, Mar. 2000.
- [29] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 46, no. 1, pp. 50–80, 2004.
- [30] J. Perello-March, C. Burns, M. T. Elliott, and S. Birrell, "Integrating trust in automation into driver state monitoring systems," in *Human Interaction and Emerging Technologies (Advances in Intelligent Systems and Computing)*, vol. 1018. Cham, Switzerland: Springer, 2020, pp. 344–349.
- [31] Waymo. (Nov. 5, 2018). *The Very Human Challenge of Safe Driving*. Accessed: Dec. 10, 2018. [Online]. Available: <https://medium.com/waymo/the-very-human-challenge-of-safe-driving-58c4d2b4e8ee>

- [32] R. D. Spain, E. A. Bustamante, and J. P. Bliss, "Towards an empirically developed scale for system trust: Take two," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 52, no. 19, pp. 1335–1339, Sep. 2008.
- [33] R. J. Lewicki, D. J. McAllister, and R. I. Bies, "Trust and distrust: New relationships and realities," *Acad. Manage. Rev.*, vol. 23, no. 3, pp. 438–458, Jul. 1998.
- [34] M. Körber, E. Baseler, and K. Bengler, "Introduction matters: Manipulating trust in automation and reliance in automated driving," *Appl. Ergonom.*, vol. 66, pp. 18–31, Jan. 2018.
- [35] M. Li, B. E. Holthausen, R. E. Stuck, and B. N. Walker, "No risk no trust: Investigating perceived risk in highly automated driving," in *Proc. 11th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, Sep. 2019, pp. 177–185.
- [36] Biopac. (2018). *BN-PPGED—BioNomadix PPG and EDA*. [Online]. Available: <https://www.biopac.com/product/bionomadix-wireless-ppg-and-eda-transmitter/>
- [37] Biopac. (2018). *BN-RSPEC—BioNomadix Amplifier and Accessories for ECG and Optional Respiration*. [Online]. Available: <https://www.biopac.com/product/bionomadix-rsp-with-ecg-amplifier/>
- [38] K. Satterfield, C. Baldwin, E. de Visser, and T. Shaw, "The influence of risky conditions in trust in autonomous systems," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 61, no. 1, pp. 324–328, Sep. 2017.
- [39] Q. Zhang, L. P. Robert, N. Du, and X. J. Yang, "Trust in AVs: The impact of expectations and individual differences," in *Proc. Conf. Auto. Vehicles Soc., Building Res. Agenda*, Mar. 2018, pp. 1–6.
- [40] V. A. Banks and N. A. Stanton, "Keep the driver in control: Automating automobiles of the future," *Appl. Ergonom.*, vol. 53, pp. 389–395, Mar. 2016.
- [41] A. Chavaillaz, D. Wastell, and J. Sauer, "System reliability, performance and trust in adaptable automation," *Appl. Ergonom.*, vol. 52, pp. 333–342, Jan. 2016.
- [42] J. Smyth, P. Jennings, P. Bennett, and S. Birrell, "A novel method for reducing motion sickness susceptibility through training visuospatial ability—A two-part study," *Appl. Ergonom.*, vol. 90, Jan. 2021, Art. no. 103264.
- [43] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers Public Health*, vol. 5, pp. 1–17, Sep. 2017.
- [44] W. Boucsein, *Electrodermal Activity*, 2nd ed. New York, NY, USA: Springer, 2012.
- [45] J. Braithwaite, D. Watson, R. Jones, and M. Rowe, *A Guide for Analysing Electrodermal Activity (EDA) and Skin Conductance Responses (SCRs) for Psychological Experiments*. Birmingham, U.K.: Univ. of Birmingham, 2015.
- [46] D. C. Fowles, "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, Aug. 2012.
- [47] M. E. Dawson, A. M. Schell, and D. L. Filion, "The electrodermal system," in *Handbook Psychophysiology*, 4th ed. Cambridge, U.K.: Cambridge Univ. Press, 2016, pp. 217–243.
- [48] J. J. Braithwaite and D. G. Watson, "Issues surrounding the normalization and standardisation of skin conductance responses (SCRs)," Univ. Birmingham, Birmingham, U.K., Tech. Rep., 2015.
- [49] P. Wintersberger, A. Riener, C. Schartmüller, A.-K. Frison, and K. Weigl, "Let me finish before i take over: Towards attention aware device integration in highly automated vehicles," in *Proc. 10th Int. Conf. Automot. User Interface Interact. Veh. Appl.*, Sep. 2018, pp. 53–65.
- [50] J. Zaman, I. Van de Pavert, L. Van Oudenhove, and I. Van Diest, "The use of stimulus perception to account for variability in skin conductance responses to interoceptive stimuli," *Psychophysiology*, vol. 57, no. 3, Mar. 2020, Art. no. e13494.
- [51] D. Miller *et al.*, "Behavioral measurement of trust in automation: The trust fall," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2016, pp. 1842–1846.
- [52] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Mach. Stud.*, vol. 27, nos. 5–6, pp. 527–539, Nov. 1987.
- [53] V. Riley, "Operator reliance on automation: Theory and data," in *Automation and Human Performance Theory and Applications*, M. Mouloua and R. Parasuraman, Eds. New York, NY, USA: Routledge, 1996, pp. 19–35.
- [54] R. Parasuraman and V. Riley, "Humans and automation?: Use, misuse, disuse, abuse," *Hum. Factors*, vol. 39, no. 2, pp. 230–253, 1997.
- [55] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 57, no. 3, pp. 407–434, May 2015.
- [56] S. Khastgir, S. Birrell, G. Dhadyalla, and P. Jennings, "Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles," *Transp. Res. C, Emerg. Technol.*, vol. 96, pp. 290–303, 2018.
- [57] M. Moon, "Waymo blames self-driving collision on pesky human," in *Waymo Blames Self-Driving Collision on Pesky Human*. New York, NY, USA: Verizon Media, Jun. 2018.

**Jaume R. Perello-March** received the B.A. degree in psychology and the M.Sc. degree in human evolution and cognition from the University of the Balearic Islands, Spain, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree in human factors with the Intelligent Vehicles Group, WMG, The University of Warwick. He has been presenting his work at international conferences. He holds previous automotive research and industrial experience in leading European centers, such as the Swedish National Road Transport Research Institute (VTI) and the Galician Automotive Technology Centre (CTAG). His research is mainly focused in using psychophysiology and neuroimaging techniques for driver state monitoring, being particularly interested in trust in automated vehicles.

**Christopher G. Burns** received the B.A. degree (Hons.), the MRes degree in research and the Ph.D. degree in psychology from the University of Edinburgh, with a specialist background in quantitative research methods and statistics in individual differences with psychometric and psychophysiological methods.

Previously, he has worked on projects involving driving simulation, sustained attention, mental workload, and emotional and attitudinal responses. He has also worked on ultrasonography simulation and medical training in teaching contexts. In 2017, he joined the WMG's Human Factors Section as a Post-Doctoral Research Fellow on the UK Autodrive Project, embedded with the Research Team, Jaguar Land Rover. His involvement in UK Autodrive at WMG consisted of designing and refining methodologies, then conducting and analyzing practical experiments in passenger and pedestrian experiences of a prototype autonomous low-speed electric vehicle (L-SATS) operating in a controlled arena. These studies included user attitudes and intentions-to-use, trust formation, technology acceptance, and internal/external human-machine interface evaluations, using both quantitative and qualitative methods.

**Roger Woodman** received the B.Sc. degree in computer science from the University of Gloucestershire in 2006, the M.Sc. degree in robotics from the University of the West of England in 2008, and the Ph.D. degree from the Bristol Robotics Laboratory in 2013. He is currently an Assistant Professor with WMG, The University of Warwick, where he leads the Human Factors research within the Intelligent Vehicles Group. He has over six years of industrial experience, working in leading manufacturing, defense, and software companies. In 2013, he joined the University of Southampton as a Research Fellow, specializing in biomedical imaging. In 2017, he joined the WMG, The University of Warwick, as a Research Fellow, investigating human factors of low-speed autonomous transport. He has several scientific articles published in the field of autonomous vehicles and robotics. He lectures in the field of human-technology interaction. His research interests include future of mobility, provable AI, transport optimization, micro mobility, and last-mile logistics. He is an Associate Fellow of the HEA.

**Mark T. Elliott** received the Ph.D. degree from Aston University in 2007, developing intelligent systems to discriminate between different walking patterns. He is currently an Associate Professor with the Institute of Digital Healthcare, WMG, The University of Warwick. He subsequently spent a number of years as a Research Fellow with the Sensory Motor Neuroscience Laboratory, University of Birmingham, modeling multisensory integration in the context of human movement coordination. His core research focuses on measuring health, wellbeing and behavior through data-driven approaches. This primarily involves analyzing and modeling data from wearable and mobile devices that capture movement and physiological responses. His research is highly applied and involves collaborating with commercial and public-sector partners.

**Stewart A. Birrell** received the first-class degree in sport science in 2002 and the Ph.D. degree in ergonomics from Loughborough University, U.K., in 2007. He is currently a Professor of Human Factors for Future Transport with the National Transport Design Centre (ntdc), Coventry University. He has spent the previous 15 years working within the transportation sector within industry and academia, with expertise ranging from driver behavior and distraction, multimodal warnings, user state monitoring, and information requirements—all underpinned by the design of in-vehicle information systems, and their evaluation using simulators, virtual reality (VR), and field operational trials. He currently applies innovative Human Factors Engineering methodologies to enable real-world and virtual evaluation of user interaction with connected and autonomous vehicle (CAV), electric vehicle (EV), and urban air mobility (UAM) technologies and services. He has over 100 journals and conference papers, book sections, and articles published in his field to date. He is an Editor of the internationally renowned, Q1/4\* journal IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.