Acclimatizing to automation: driver workload and stress during partially automated car following in real traffic

Daniël D. Heikoopa\textsuperscript{a}, Joost C. F. de Winter\textsuperscript{b}, Bart van Arem\textsuperscript{c}, and Neville A. Stanton\textsuperscript{a}

\textsuperscript{a}Corresponding author

\textsuperscript{a}Transportation Research Group, Faculty of Engineering and the Environment, Boldrewood Innovation Campus, University of Southampton, Burgess Road, Southampton, SO16 7QF, United Kingdom

\textsuperscript{b}Department of BioMechanical Engineering, Delft University of Technology, the Netherlands

\textsuperscript{c}Department of Transport & Planning, Delft University of Technology, the Netherlands

Automated driving systems are increasingly prevalent on public roads, but there is currently little knowledge on the level of workload and stress of drivers operating an automated vehicle in a real environment. The present study aimed to measure driver workload and stress during partially automated driving in real traffic. We recorded heart rate, heart rate variability, respiratory rate, and subjective responses of nine test drivers in the Tesla Model S with Autopilot. The participants, who were experienced with driver assistance systems but naïve to the Tesla, drove a 32 min motorway route back and forth while following a lead car in regular traffic. In one of the two drives, participants performed a heads-up detection task of bridges they went underneath. Averaged across the two drives, the participants' mean self-reported overall workload score on the NASA Task Load Index was 19%. Moreover, the participants showed a reduction of heart rate and self-reported workload over time, suggesting that the participants became accustomed to the experiment and technology. The mean hit (i.e., pressing the button near a bridge) rate in the detection task was 88%. In conclusion, driving with the Tesla Autopilot on a motorway involved a low level of workload that decreased with time on task.

1. Introduction

1.1. Workload and Stress in Automated Driving

Cars that provide combined longitudinal and lateral automated control support have recently been introduced on the market. Automated driving may be expected to reduce workload and stress as compared to manual driving because the driver does not have to control the vehicle. However, unless the driving task is fully automated (SAE level 5), automated driving may cause high workload and stress, because the driver needs to supervise both the human-machine interface and the state of the car in relation to the outside environment (for an illustration, see Fig. 1). More specifically, the driver of an automated car has to remain attentive to reclaim manual control if required [1, 2] a task that may be demanding and stressful [3]. Furthermore, the type of supervisory control shown in Figure 1 may cause out-of-the-loop problems, such as loss of situation awareness and mode errors, which resemble those observed in aviation and process control [4, 5, 6; see also 7]. A survey by Dikmen and Burns [8] among 121 Tesla owners found that automation failures (e.g., failure to detect lanes) were frequent but not perceived as risky. Furthermore, the majority of respondents indicated that it is important to remain alert and to be aware of the automation’s limitations.

1.2. Prior research on workload and stress in automated driving.

The majority of Human Factors research on driver workload in automated vehicles has been conducted in driving simulators [see 9 for a review]. Overall, the results indicate that the self-reported workload as assessed with the NASA Task Load Index (TLX) is substantially lower in automated driving than in manual driving [see 9 for a review], and below 50% on a scale from 0 to 100% (see Table 1 for an overview).
Figure 1. Manual control (left), supervisory control (middle), and fully automatic control (right). In manual control, the
driver controls the car via manipulators (steering wheel & pedals) and continuously receives information from the car in the
environment (i.e., task). In fully automatic driving, the human has no contribution to the driving task other than to set a
destination (or to press an emergency stop button). Hence, the driver is out of the control loop completely. In supervisory
control, the driver interacts with a computer that closes the control loop via sensors and actuators, while the driver
intermittently (1) provides instructions to the computer, (2) receives information via displays, and (3) receives information
from the car in the environment [from 12].

A small number of on-road studies are available. Recently, Endsley [10] conducted a single-subject naturalistic
driving study using a Tesla Model S over a six-month period. She reported that her situation awareness increased
when using automation, because less focus was needed on controlling the vehicle, and more attention could be
devoted to looking at traffic and road signage. However, Endsley also experienced various issues of mode
confusion and unexpected automation transitions, as well as loss of attention. Endsley further found that ratings
of satisfaction, usefulness, and trust gradually increased from months 1–2 towards months 5–6, which is in line
with the results of a longitudinal naturalistic driving study on adaptive cruise control (ACC) with 15 participants
[11]. Additionally, overall self-reported workload was low, averaging at about 1.3 during months 1–2 and 1.0
during months 3–4, on a scale from 0 to 5 [10].

Eriksson et al. [13] let 12 test drivers use the Tesla autopilot for about 20 minutes per participant. Participants
each experienced approximately 12 automation-to-manual control transitions, and completed the NASA-T LX
after the ride. The mean overall workload was 19%. Stapel et al. [14] conducted an on-road highway driving
study in which 15 participants used the Tesla Autopilot for about 20 minutes. The authors found overall low
levels of workload among participants (between 10% and 43%), with the type of road (busy city ring versus
relatively empty highway) and prior experience with the Tesla Model S being moderator variables (Table 1). In
another on-road study, Banks and Stanton [15] tested a prototype version of automated longitudinal and lateral
control in addition to a driver-initiated auto-overtaking system. These authors found relatively high workload on
the NASA-T LX (median of 42%) during 9 minutes of automated driving per participant.

The discrepancy between the results of Banks and Stanton [15] and the findings of Eriksson et al. [13] and Stapel
et al. [14] may be caused by the fact that the prototype system tested by Banks and Stanton, which included a
heads-up display and offered overtake suggestions, was difficult to use or that participants were still learning
how to use it. Because the participants in Banks and Stanton [15] drove only 9 minutes with the automation
system, the high workload levels “may be a simple reflection of the fact that these ratings were collected during
first-time use of the automated system”, p. 393.

McDowell et al. [16] and Davis et al. [17] performed on-road trials with automated military convoys. In these
studies, where there was no other traffic and, because they were military experiments, object detection was of
primary importance. The results showed that automated driving reduced workload and improved performance in object detection in comparison to manual driving.

On-road studies may be expected to yield higher workload than simulator studies, because the latter involve no physical risks of accidents. However, in some cases, on-road studies actually yielded lower workload than simulator-based studies. For example, the reported workload in Eriksson et al. [13] was 19%, compared to the 33% in the simulator study of Manawadu et al. [18]. This difference might be attributable to the participant pools, added events, or secondary tasks. Specifically, the study of Eriksson et al. [13] involved experienced test drivers and did not include a secondary task; participants were merely required to take over and relinquish control of the vehicle throughout the experiment. In Manawadu et al. [18], critical events were triggered, to which the participants had to respond.

Table 1
Overview of workload measurements presenting a NASA-TLX overall workload (TLX OW) score in automated driving studies.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Simulator/road</th>
<th>Sample size</th>
<th>Mean TLX OW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banks and Stanton [15]</td>
<td>On-road</td>
<td>32</td>
<td>42% (median)</td>
</tr>
<tr>
<td>Borjenci et al. [19]</td>
<td>Simulator</td>
<td>21</td>
<td>30%</td>
</tr>
<tr>
<td>Damböck et al. [20]</td>
<td>Simulator</td>
<td>24</td>
<td>33%</td>
</tr>
<tr>
<td>De Winter et al. [21]</td>
<td>Simulator</td>
<td>24</td>
<td>31% (exp. 1)</td>
</tr>
<tr>
<td></td>
<td>Simulator</td>
<td>27</td>
<td>31% (exp. 2)</td>
</tr>
<tr>
<td>Eriksson et al. [13]</td>
<td>On-road (Tesla)</td>
<td>12</td>
<td>19%</td>
</tr>
<tr>
<td>Eriksson and Stanton [22]</td>
<td>Simulator</td>
<td>26</td>
<td>21%</td>
</tr>
<tr>
<td>Heikoop et al. [23]</td>
<td>Simulator</td>
<td>22</td>
<td>28%</td>
</tr>
<tr>
<td>Large et al. [24]</td>
<td>Simulator</td>
<td>30</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(partial automation), 21% (high automation)</td>
</tr>
<tr>
<td>Manawadu et al. [18]</td>
<td>Simulator</td>
<td>6 (novices)</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td>Simulator</td>
<td>6 (experienced)</td>
<td>30%</td>
</tr>
<tr>
<td>McDowell et al. [16]</td>
<td>On-road (military)</td>
<td>11</td>
<td>40%</td>
</tr>
<tr>
<td>Petermeijer et al. [25]</td>
<td>Simulator</td>
<td>24</td>
<td>28% (with auditory and vibrotactile feedback)</td>
</tr>
<tr>
<td>Petermeijer et al. [26]</td>
<td>Simulator</td>
<td>18</td>
<td>22%, 36% (with N-Back task)</td>
</tr>
<tr>
<td>Saxby et al. [27]</td>
<td>Simulator</td>
<td>36</td>
<td>34% (exp. 1)</td>
</tr>
<tr>
<td>Schwalk et al. [28]</td>
<td>Simulator</td>
<td>56</td>
<td>27% (exp. 2)</td>
</tr>
<tr>
<td>Stapel et al. [14]</td>
<td>On-road (Tesla)</td>
<td>8 (no experience with Tesla)</td>
<td>25% (empty highway), 43% (city ring)</td>
</tr>
<tr>
<td></td>
<td>On-road (Tesla)</td>
<td>7 (experienced with Tesla)</td>
<td>10% (empty highway), 24% (city ring)</td>
</tr>
<tr>
<td>Young [29]</td>
<td>Simulator</td>
<td>18</td>
<td>23%</td>
</tr>
<tr>
<td>Young and Stanton [30]</td>
<td>Simulator</td>
<td>12</td>
<td>12% (exp. 1)</td>
</tr>
<tr>
<td></td>
<td>Simulator</td>
<td>12</td>
<td>12% (exp. 2)</td>
</tr>
<tr>
<td>Young and Stanton [31]</td>
<td>Simulator</td>
<td>24 (novice drivers)</td>
<td>11%</td>
</tr>
<tr>
<td></td>
<td>Simulator</td>
<td>30 (learner drivers)</td>
<td>13%</td>
</tr>
<tr>
<td></td>
<td>Simulator</td>
<td>30 (expert drivers)</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>Simulator</td>
<td>30 (advanced drivers)</td>
<td>24%</td>
</tr>
</tbody>
</table>

1.3. Aim of the present study

The present study aimed to assess whether on-road automated driving with the Tesla Model S alleviates driver workload over time. Both the on-road studies of Eriksson et al. [13] and Stapel et al. [14] consisted of approximately 20 min of highway driving with the automation engaged (excluding a familiarization drive) and did not report on temporal effects. Our study consisted of 64 minutes of automated highway driving per participant.

Additionally, in our study, a simple detection task was used to add task demands on top of the regular monitoring demands during automated driving. More specifically, participants were instructed to press a handheld button when driving underneath a bridge. The task of detecting bridges is practically convenient in an on-road experiment because bridges are irregularly spaced, and the locations of bridges are retrievable from Google Maps. This detection task is conceptually similar to the approach taken in a previous platooning experiment in a driving simulator [23]. In Heikoop et al. [23], it was found that the detection task (i.e., to detect red cars on the
road) increased self-reported mental demands compared to not performing a detection task. We expected to find a similar effect in this study.

2. Methods

2.1. Participants
Nine participants (seven males, two females) aged between 25 and 47 years ($M = 35.44; SD = 8.26$) with 6 to 30 years of self-reported driving experience ($M = 17.56; SD = 8.46$) took part in this experiment. The participants were employees of a large automotive company. Eight participants indicated that they drove every day and one participant indicated driving 4–6 days a week. Two participants indicated they drove up to 10,000 miles, five up to 20,000, one up to 30,000, and one up to 50,000 miles in the past year. All participants had completed level-2 driver training, an extended driver training specifically designed for people who drive as part of their job and which serves as a legal requirement for insurance purposes. All participants had driven various supercars before, and had experience with advanced driver assistance systems (e.g., adaptive cruise control, lane keeping assist), but had no experience with the Tesla Autopilot. We refrained from recruiting participants who had experience with automated driving systems, such as the Tesla’s Autopilot, because participants in our previous experiment were not experienced with automated driving systems either [23] (and see [14] for a comparison of self-reported workload between automation-experienced drivers and automation-inexperienced drivers).

No incentive was provided to the participants, and all participants gave written informed consent. The study was approved by the Ethics Research Governance Office of the University of Southampton under submission ERGO number 19091.

2.2. Apparatus
The experiment was performed with a Tesla Model S 90D with Autopilot as the participants’ vehicle (PV) and a Jaguar XF as a lead vehicle (LV). The LV was used for safety reasons and for creating a persistent car following task. With a forward-looking radar, forward-facing camera, and ultrasonic sensors, the Autopilot can steer, adjust speed, detect obstacles, and apply brakes automatically [32]. The Tesla Autopilot can be characterised as SAE J3016 level 2 automation (i.e., partial automation) because both steering and speed control are automated, and the driver is still expected to monitor the driving environment [33].

The Traffic-Aware Cruise Control (TACC) of the PV was set to 1, which was the closest following distance and which translates to a time headway of about 1 second. This headway corresponds to common headways in highway traffic [34, 35, 36, 37], and was sufficiently short to have a low likelihood of other cars merging in between the PV and LV.

Participants wore electrocardiography (ECG) equipment linked to LabChart 8 that captured their cardiovascular and respiratory activity. This ECG equipment consisted of the AD Instruments PowerLab 26T Teaching Series, three MLA2505 biopotential electrodes and lead wires with disposable ECG electrode patches, and the MLT1132 respiratory belt transducer. The electrodes were placed in a triangular configuration. For male participants, one electrode was placed over the xiphoid process, and two electrodes below the far ends of the collar. For female participants, one electrode was placed at the top of the sternum and two electrodes below the ribs on both sides. This gender-based distinction was mainly made for comfort purposes [see e.g., 38]. The respiratory belt was placed over the clothes around the chest.

2.3. Environment
The experiment took place on March 14–18, 2016. Participants drove on the left (slow) lanes of the British dual three-lane motorways M40, M42, and M5, for which the speed limit is 70 mph (112 km/h). Participants completed two drives during daytime outside of rush hours. The first drive was completed between entry point 14 of the M40 northbound and M5 northbound exit point 3 (Fig. 2). In the second drive, the participants drove back to the starting point. Specifically, the second drive was completed between the motorway entry point at the service stations after entry point 3 of the M5 southbound and the M40 southbound until exit point 14.
2.4. Procedure

All participants received a training trial and completed two drives of approximately 32 minutes each. Vigilance research has shown that detection performance exhibits a decay function with time on task [39]. Furthermore, it has been found that after 15 minutes the most substantial deterioration of detection performance has taken place (see a review by [40], reporting that “at least half of the final loss is completed within the first 15 min). Because the average driving trip in Europe and the U.S. is between 20 and 30 minutes [41, 42], it may be assumed that the present study is representative of the first exposures to a new automated driving system on public roads.

Before the experiment, the participant performed a test drive on a test track. Upon arrival at the test track site, the participant received paper instructions explaining that he/she would be driving within a highly automated platoon. Furthermore, a consent form, a demographics questionnaire, and the pre-task Dundee Stress State Questionnaire (DSSQ) were provided. After having completed these questionnaires, the participant was taken to the passenger seat of the PV and introduced to the safety driver, who was a professional driver, trained to intervene in emergencies. The safety driver sat in the passenger seat throughout the experiment for legal and safety reasons. The safety driver performed a lap on the test track and showcased the Autopilot, as well as several details of the car. After that lap, the participant and safety driver changed seats, and the participant drove the car until they were comfortable driving manually and with the Autopilot feature. Then the ECG electrodes were attached after which the participant drove to the selected motorway entry point, following the LV. After entering the motorway, the Autopilot was engaged by the participant, and the experiment started. The safety driver sat in the seat next to the participant, and verbally intervened if the participant did not act appropriately or safely (e.g., when the participant did not override the automation when he/she should). The experimenter sat in the rear seat, monitoring the equipment and making notes of events during the experiment. Before the first drive, the participant was discouraged from interacting with the safety driver or the experimenter for the duration of the experiment. Thus, the interaction between the safety driver and the experimenter was kept to a minimum.

In the occasions where another vehicle merged in between the PV and the LV, the participants were instructed by the safety driver to remain in automated mode and follow this other vehicle. However, if the gap with the LV became large, then the participants were calmly instructed to follow the LV again by overtaking the outside traffic while it was emphasised to try to remain in automated mode. An automated lane change could be performed by using the indicator stalk while holding the steering wheel. All events such as lane changes, merges, and Autopilot (dis)engagements were recorded by the experimenter using paper and pencil. Summed across the nine participants, a total of 33 and 37 lane changes (of which 16 and 21 automated) occurred for Drive 1 and 2, respectively. A manual override occurred 9 and 4 times during Drive 1 and Drive 2, respectively.

At the end of the first drive, the participants exited the motorway and stopped at a nearby parking lot. They were then provided with the post-task DSSQ and the NASA Task Load Index (TLX). Once completed, the
participants followed the LV to the motorway again and performed the second drive. At the end of the second drive, the participants were again provided with the post-task DSSQ and TLX.

2.5. Independent variables
The experiment consisted of two drives, either with (DT) or without (NT) a detection task, in counterbalanced order. Specifically, five participants completed the second ‘southbound’ drive with the detection task, and four participants completed the first ‘northbound’ drive with the detection task. Without the detection task, participants had to follow the LV as their only objective. With the detection task, they also had to detect the bridges they went underneath by pressing a handheld button (Fig. 3). In the first drive, participants drove underneath 50 bridges, and during the second drive, participants drove underneath 47 bridges. Photos of the bridges are available as supplementary material.

Learning/acclimatization effects were assessed by comparing the results of the first drive with the results of the second drive.

2.6. Dependent measures
The following measures were calculated per participant for each of the two drives:
- Duration of the drive (s).
- Mean speed (km/h), recorded with a GPS application on a smartphone.
- Hit rate of the bridges (% of bridges detected). The hit rate was calculated by linking the known locations of the bridges (as retrieved from Google Maps) with the locations at the moments of button presses (recorded with a GPS application on a smartphone).
  An algorithm was written that matched bridges with the nearest button press in terms of radial distance until all bridges were assigned to a button press or no button presses were left (each button press could be assigned to one bridge only). Button presses which followed each other within ⅔ seconds (i.e., accidental double pressing of the button) were discarded. Furthermore, if the nearest button press was more than 1,250 m from the bridge, then this bridge was marked as a miss. The liberal threshold of 1,250 was used because there were several sources of inaccuracy in the locations of the button presses. Specifically, (1) The GPS signal had a limited temporal resolution (0.2 Hz, which at an average speed of 86 km/h amounts to a travelled distance of about 120 m), (2) Some participants pressed the button late (i.e., when being beneath a bridge) while others pressed the button early (i.e., when the bridge could first be seen), and (3) The GPS recording had limited accuracy (the 50th, 95th, and 99th percentile of the estimated accuracy were 24 m, 249 m, and 965 m, respectively). By definition, the miss rate equals 100% minus the hit rate [43].
- False alarm rate (% of false alarms relative to the number of bridges). A button press was considered a false alarm when after determining the hits, there were still button presses unaccounted for. Figure 4 provides an illustration of the hits and false alarms for Participant #1.
- Heart rate (bpm). The heart rate was regarded as a measure of stress [44].
- SDNN (ms), a time-domain measure of mental workload [e.g., 45, 46, 47]. The SDNN was defined as the mean of the standard deviation (SD) of all Normal to Normal peak intervals (NN) in the ECG signal per 5-min segment along the drive [SDNN index, see 48]. A low SDNN value is interpreted as high workload [49; see also 23]. See Figure 5 for an illustration of the calculation process.
- LF/HF ratio, a frequency-domain measure of mental workload. This spectral analysis of the NN interval calculates the power in the low-frequency (LF) 0.04–0.15 Hz range relative to the power in the high frequency (HF) 0.15–0.40 Hz range. A high LF/HF ratio is indicative of high workload [50, 51]. Both the SDNN and the LF/HF ratio were calculated from the NN intervals after a default NN artefact filter using an open-source MATLAB program [52].
- Respiratory rate (bpm). Because the respiratory belt transducer produced a noisy signal (presumably because of in-vehicle vibrations) and may contain drifts and other artefacts, the signal was filtered with a second-order Butterworth 0.1–1.0 Hz bandpass filter. This frequency range incorporates a typical human respiratory rate of 0.25 Hz. Next, the data were rank transformed to remove outliers, and subsequently, a discrete Fourier transformation was applied to retrieve the frequency with maximum amplitude (see Fig. 6 for illustration).
- DSSQ, a self-report measure of stress and fatigue [53]. In this experiment, version 1.3 of the DSSQ was used [54]. Standardized change scores for each scale of the DSSQ were calculated as follows: (post-score−pre-score)/(standard deviation of the pre-score) [55]. The scores for the three scales (Engagement, Distress, and Worry) were calculated by averaging four subscales and averaging them to result in one score for each element [based on 56, 57, 58]. Task Engagement consists of the subscales (1) Energetic Arousal, (2) Success Motivation, (3) Intrinsic Motivation, and (4) Concentration. Distress consists of (5) Tense Arousal, (6) Hedonic Tone, (7) Control & Confidence, and (8) Anger/Frustration. Finally, Worry consists of
(9) Self-Focused Attention, (10) Self-Esteem, (11) Task-Relevant Interference, and (12) Task-Irrelevant Interference. We imputed missing answers (4% of the total) using the nearest-neighbour method.

- NASA-TLX, a self-report measure to assess workload [59]. The ‘raw’ approach was used, also known as the Raw TLX (RTLX). This approach does not apply weights to the scales [60].

The mean speed, duration, heart rate, SDNN, LF/HF ratio, and respiratory rate were calculated from the moment that the participant was 200 m in front of the first bridge until 200 m after the participant passed the last bridge. SDNN was calculated as the average across six available 5-min segments.

![Figure 3. Photo taken during the experiment: the participant on the right presses the handheld button when detecting a bridge. The safety driver is sitting in the left (passenger) seat. The vehicle in front is the LV the participant had to follow during the experiment.](image)

3. Results
Table 2 presents results per individual participant. Due to the small number of participants in this study, statistical tests are not reported, as these were deemed unreliable.

Participants drove on average about 32 km per drive, at a mean speed of 86 km/h (Table 2), which is well below the speed limit of 112 km/h (the speed limit on British motorways is 70 mph, which equals 112 km/h). The difference in duration between Drive 1 and Drive 2 is caused by the fact that Drive 1 was about 4 km longer than Drive 2. This was due to the respective entry- and exit points being in different locations.
Figure 4. Illustration of GPS data and the detection task. In this case, the participant detected all 47 bridges (hit rate = 100%) and had 1 false alarm (false alarm rate = 2.1%). Where a single dot is visible for two bridges, two button presses appeared in the same GPS sample (the GPS recorded the position every 5 seconds).

Figure 5. Illustration of the calculation of SDNN. A) ECG signal with extracted NN intervals (first 10 s of Drive 1 of Participant #1). B) Distribution of NN intervals with the mean and standard deviation of the NN intervals (SDNN = 34.6 ms; based on the first 300 s of Drive 1 of Participant #1).
The 9 participants together manually took control of the automated driving system 13 times, of which 6 were due to the Autopilot failing to anticipate on traffic merging between the LV and PV, 2 to the Autopilot following the undesired line at an exit point, 2 to the participant not trusting the Autopilot to perform correctly, 1 to an unexpected disengagement of the Autopilot, and 1 to the participant disengaging the Autopilot without apparent reason. The remaining Autopilot disengagement occurred for unknown reasons. Furthermore, lane changes were performed 29 and 41 times during the no task (NT) and detection task (DT) condition, respectively, of which 16 and 17 were manual.

Table 3 shows the results of the self-report questionnaires. Averaged across the two drives, the mean self-reported overall workload was 19% (21% in Drive 1, 16% in Drive 2; 18% for NT drives, 19% for DT drives). The mean (SD) per TLX item was 26% (19%) for Physical Demand, 8% (6%) for Mental Demand, 12% (9%) for Temporal Demand, 27% (29%) for Performance, 17% (13%) for Effort, and 21% (23%) for Frustration. The DSSQ results showed that participants exhibited an overall disengagement from the task, and a worriless attitude towards the task as compared to the pre-task DSSQ (i.e., the standardized change scores are smaller than 0).

Table 4 shows the results of the physiological measures. An acclimatization effect can be seen, with the heart rate being lower in Drive 2 than in Drive 1 for 8 out of 9 participants. The SDNN exhibits a negative correlation with the heart rate [see also 23]. Here, for 7 of 9 participants, SDNN was higher in Drive 2 than in Drive 1. The LF/HF ratio and respiratory rate remained relatively constant throughout the two drives.
Table 2
Descriptive results (time of day, duration, mean speed, manual overrides, manual and automated lane changes) per participant and drive number.

<table>
<thead>
<tr>
<th>PP</th>
<th>Time of day</th>
<th>Duration (s)</th>
<th>Mean speed (km/h)</th>
<th>Manual overrides</th>
<th>Manual lane changes</th>
<th>Automated lane changes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drive 1</td>
<td>Drive 2</td>
<td>Drive 1</td>
<td>Drive 2</td>
<td>Drive 1</td>
<td>Drive 2</td>
</tr>
<tr>
<td>1 (NT, DT)</td>
<td>10:52</td>
<td>11:55</td>
<td>2158</td>
<td>1925</td>
<td>80.5</td>
<td>82.9</td>
</tr>
<tr>
<td>2 (DT, NT)</td>
<td>14:29</td>
<td>15:18</td>
<td>2032</td>
<td>1892</td>
<td>85.2</td>
<td>84.1</td>
</tr>
<tr>
<td>3 (NT, DT)</td>
<td>10:43</td>
<td>11:46</td>
<td>2057</td>
<td>1653</td>
<td>85.4</td>
<td>96.5</td>
</tr>
<tr>
<td>4 (DT, NT)</td>
<td>14:40</td>
<td>15:20</td>
<td>1440</td>
<td>1405</td>
<td>85.6</td>
<td>86.1</td>
</tr>
<tr>
<td>5 (NT, DT)</td>
<td>10:33</td>
<td>11:28</td>
<td>2073</td>
<td>1871</td>
<td>83.6</td>
<td>85.1</td>
</tr>
<tr>
<td>6 (DT, NT)</td>
<td>14:41</td>
<td>15:33</td>
<td>2013</td>
<td>1834</td>
<td>86.1</td>
<td>87.0</td>
</tr>
<tr>
<td>7 (NT, DT)</td>
<td>10:25</td>
<td>11:18</td>
<td>1991</td>
<td>1813</td>
<td>87.1</td>
<td>87.9</td>
</tr>
<tr>
<td>8 (DT, NT)</td>
<td>14:34</td>
<td>15:34</td>
<td>1928</td>
<td>800</td>
<td>88.9</td>
<td>88.2</td>
</tr>
<tr>
<td>9 (NT, DT)</td>
<td>10:40</td>
<td>11:42</td>
<td>2008</td>
<td>1886</td>
<td>86.1</td>
<td>84.6</td>
</tr>
<tr>
<td>Average</td>
<td>2033</td>
<td>1834</td>
<td>86.9</td>
<td>1.00</td>
<td>0.44</td>
<td>1.56</td>
</tr>
</tbody>
</table>

Note. NT = No Task, DT = Detection Task. Participant #4 did not complete the entire route because the batteries of the car were emptying and the car needed to be charged. This participant was excluded from the calculation of the average duration.

Table 3
Self-reported overall workload (TLX OW) and standardised change scores of self-reported stress (DSSQ) per participant and drive number.

<table>
<thead>
<tr>
<th>PP</th>
<th>TLX OW (%)</th>
<th>DSSQ engagement</th>
<th>DSSQ distress</th>
<th>DSSQ worry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drive 1</td>
<td>Drive 2</td>
<td>Drive 1</td>
<td>Drive 2</td>
</tr>
<tr>
<td>1 (NT, DT)</td>
<td>12</td>
<td>8</td>
<td>-0.10</td>
<td>-0.21</td>
</tr>
<tr>
<td>2 (DT, NT)</td>
<td>18</td>
<td>14</td>
<td>-0.12</td>
<td>-0.18</td>
</tr>
<tr>
<td>3 (NT, DT)</td>
<td>45</td>
<td>46</td>
<td>-1.10</td>
<td>-1.19</td>
</tr>
<tr>
<td>4 (DT, NT)</td>
<td>17</td>
<td>3</td>
<td>-0.01</td>
<td>0.26</td>
</tr>
<tr>
<td>5 (NT, DT)</td>
<td>31</td>
<td>20</td>
<td>-0.37</td>
<td>-0.61</td>
</tr>
<tr>
<td>6 (DT, NT)</td>
<td>14</td>
<td>14</td>
<td>0.18</td>
<td>0.46</td>
</tr>
<tr>
<td>7 (NT, DT)</td>
<td>9</td>
<td>6</td>
<td>-1.48</td>
<td>-0.52</td>
</tr>
<tr>
<td>8 (DT, NT)</td>
<td>40</td>
<td>25</td>
<td>-2.60</td>
<td>-1.88</td>
</tr>
<tr>
<td>9 (NT, DT)</td>
<td>5</td>
<td>7</td>
<td>-1.43</td>
<td>-0.02</td>
</tr>
<tr>
<td>Average</td>
<td>21</td>
<td>16</td>
<td>-0.78</td>
<td>-0.32</td>
</tr>
</tbody>
</table>
Table 4
Cardiovascular and respiratory results per participant and drive number

<table>
<thead>
<tr>
<th>PP</th>
<th>Heart rate (bpm)</th>
<th>SDNN (ms)</th>
<th>LF/HF ratio</th>
<th>Respiratory rate (bpm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drive 1</td>
<td>Drive 2</td>
<td>Drive 1</td>
<td>Drive 2</td>
</tr>
<tr>
<td>1</td>
<td>(NT, DT)</td>
<td>61.3</td>
<td>61.9</td>
<td>44.7</td>
</tr>
<tr>
<td>2</td>
<td>(DT, NT)</td>
<td>99.4</td>
<td>90.4</td>
<td>20.3</td>
</tr>
<tr>
<td>3</td>
<td>(NT, DT)</td>
<td>59.7</td>
<td>56.6</td>
<td>50.2</td>
</tr>
<tr>
<td>4</td>
<td>(DT, NT)</td>
<td>72.2</td>
<td>67.5</td>
<td>48.3</td>
</tr>
<tr>
<td>5</td>
<td>(NT, DT)</td>
<td>69.0</td>
<td>66.4</td>
<td>55.7</td>
</tr>
<tr>
<td>6</td>
<td>(DT, NT)</td>
<td>82.1</td>
<td>73.0</td>
<td>43.6</td>
</tr>
<tr>
<td>7</td>
<td>(NT, DT)</td>
<td>59.2</td>
<td>58.9</td>
<td>44.9</td>
</tr>
<tr>
<td>8</td>
<td>(DT, NT)</td>
<td>80.4</td>
<td>70.7</td>
<td>35.3</td>
</tr>
<tr>
<td>9</td>
<td>(NT, DT)</td>
<td>56.6</td>
<td>52.2</td>
<td>93.2</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>71.1</td>
<td>66.4</td>
<td>48.5</td>
</tr>
</tbody>
</table>

Note. The respiratory rate of Participant #2 is not provided, as there was no clear peak value to be identified from the Fourier transformation.

Due to technical issues, participants #3, 4, and 6 had no button press data for 14, 16, and 10 bridges respectively. Hit rates and false alarm rates were calculated for the remaining number of bridges for these participants. The mean (SD) hit and false alarm rates of the bridges were 88.0% (16.0%) and 0.9% (1.5%), respectively. The lowest hit rate was 47.1%, whereas two participants had hit rates of 100% (Figure 7). The mean hit rate and mean false alarm rate correspond to a perceptual sensitivity ($d'$) of 3.52 and response bias ($\beta$) of 7.89 [61]. These results indicate that participants were well able to distinguish the bridges from the non-bridges with a conservative response strategy.

![Figure 7](image_url)

**Figure 7.** Number of hits and misses per participant. The number of false alarms was 1, 0, 0, 1, 0, 0, 0, 2 for participants 1–9.

4. Discussion
This study aimed to measure levels of workload and stress during automated driving with the Tesla Autopilot. The literature has shown that automated driving yields low ratings of self-reported overall workload (averaging at 23%, see [9]). A previous study using the Tesla Autopilot has found overall workload scores ranging from 10% for experienced Tesla drivers on an empty highway to 43% on a city ring with drivers who had not driven in the Tesla before [Table 1; 14]. Another on-road experiment found high workload for automated driving compared to
manual driving, with overall workload scores for automated driving being 42% [15]. It was unclear whether the novelty of the automation in Banks and Stanton [15] created elevated levels of workload, so they proposed extended exposure to automation, which was the purpose of the current study.

Our results of a 2 x 32 min of automated driving showed that the mean overall workload dropped from 21% in Drive 1 to 16% in Drive 2. In other words, automated driving involves a level of self-reported workload that is within the range of the workload observed in driving simulators (see Table 1, which shows a minimum overall workload of 11% and a maximum of 36%). The fact that the participants had to follow a lead vehicle may have contributed to the low overall workload by limiting their decision-making requirements. The workload was particularly low for the Physical Demand item, which may be because the participants did not have to move the pedals or steering wheel for large portions of the time. Similarly, the average heart rate and respiratory rate were close to the resting rates of a typical person [62, 63]. Thus, automated driving in the present experiment could not be considered demanding or stressful.

Compared to the detection task in Heikoop et al. [23], participants in the current study performed somewhat worse (hit rate of 95% in [23]; 88% in the present study). Furthermore, self-reported workload in the current study was not substantially different between the DT and NT conditions. It is possible that, for safety reasons or due to the presence of the safety driver sitting next to them, participants in the present study were trying harder to stay alert to the primary driving task as compared to the participants in the simulator study, thereby having a lower incentive to detect the bridges. Indeed, with the Tesla Autopilot, participants do have to remain alert due to the potential need to intervene and take manual control of the vehicle. In a survey study by Dikmen and Burns [8], 62% of Autopilot users reported that they had experienced at least one unexpected or unusual behaviour when driving in automated mode. In our experiment, the 9 participants took over manual control 9 times, for a variety of reasons, and performed 70 lane changes, of which 47% were manual. The relatively large percentage of manual lane changes may be due to the somewhat cumbersome technique required to perform an automated lane change. For an automated lane change to succeed, the driver has to press the indicator stalk while having his/her hands on the steering wheel with enough weight for the Autopilot to recognize their presence. In our experiment, this often resulted in a slight turn of the steering wheel, disengaging the Autopilot, after which the lane change had to be performed manually.

Comparing participants’ self-reported engagement during a previous driving simulator study featuring similar methods [23] and another more recent simulator-based platooning study [64] with the current on-road study (Figure 8), it can be seen that participants felt relatively engaged during the on-road study. The relatively high level of engagement may be because the participants in the present study prioritized safety and tried to stay alert, as noted above. Participants of this on-road study reported relatively low levels of distress and worry compared to the participants in the simulator study. This may be because the participants in the on-road study were professional drivers who were used to driving with advanced driver assistance systems. Furthermore, the overall self-reported workload was low (a mean score of 19% on a scale from 0% to 100%). Both the on-road and simulator-based studies found that self-reported workload remained approximately constant, heart rate decreased, and SDNN increased as a function of time. Previous on-road studies into manual driving also found that the heart rate tends to decrease with time on task [65, 66]. Finally, the LF/HF ratio in the simulator and on the road were within the same range (Fig. 8).

5. Conclusions and Recommendations
This experiment complements existing research on stress and workload during automated driving (cf. Table 1), and may form a basis for more extensive research on this topic. The results point to an effect of acclimatization as demonstrated by a drop in perceived workload over time, and a decrease in heart rate. Our sample was small ($N = 9$); in order to acquire greater statistical power, replication studies with more participants are advised. Nonetheless, our study produced insights into the effects of workload and acclimatization to automation, and could serve as a foundation for future research into this phenomenon.

This experiment used expert drivers, and it remains to be investigated how the results translate to less trained drivers. It is likely that the participants in this study, who were experienced with various supercars and had completed advanced driver training, are more adept to novel technology and less likely to be stressed than the general population. Stapel et al. [14] found in their on-road study that people who have experience with driving in a Tesla reported a lower workload than people who had not driven a Tesla before, both during manual and automated driving.
Figure 8. Comparisons of three independent experiments. Top: The Dundee Stress State Questionnaire (DSSQ). Middle: Cardiovascular measures per 5-minute interval. Bottom: Self-reported overall workload. Experiment 1: Heikoop et al. (2017) [23] \((N = 22; 3 \times 40\text{ minutes of simulated driving})\). Experiment 2: Heikoop et al. (2018) [64] \((N = 33, \text{ but } N = 29 \text{ for heart rate measures}; 3 \times 40\text{ minutes of simulated driving})\). Experiment 3: present experiment \((N = 9; 2 \times 32\text{ minutes of on-road driving})\). Note: The DSSQ used in Experiment 2 is the short version of the DSSQ. The DSSQ and workload scores were scaled from 0% (minimum possible) to 100% (maximum possible).

The methods used for psychophysiological measurement need consideration. Although the LF/HF ratio is regarded as a valid measure of workload [50, 51], the results of the LF/HF ratio in this experiment did not show the same time-on-task effects as the self-reported overall workload. The lack of sensitivity of the LF/HF ratio could be due to various confounding effects such as driver posture, vibrations in the vehicle, or a dependency on the heart rate itself [as also discussed by 23]. Also, the use of a smartphone-based GPS app did not allow for high accuracy data (see Section 2.6). It is recommended that future researchers use CAN bus data and differential GPS (DGPS).

The focus of this experiment was on how drivers are affected by automated driving over time. Future on-road research could include a control condition in which people drive manually [cf. 14] or include different levels of automation such as driving with ACC, or driving with ACC and steer assist [see 66]. In our experiment, control transitions were not wanted, yet occurred several times per participant. A closed-track on-road experiment [cf. 68, 69] could investigate the psychological effects of transition of control to and from automated driving in a controlled manner. Furthermore, it could be investigated whether the present findings generalize to driving tasks situations without safety driver, and with actual hazards for which manual intervention is necessary. It is possible that drivers would score better in an environment in which the target stimuli represent actual safety-critical events. Finally, it remains to be studied how our findings generalize to other driving scenarios, such as driving during rush hours.

6. Acknowledgements

This experiment has been performed in collaboration with Jaguar Land Rover. The authors are involved in the Marie Curie Initial Training Network (ITN): HFAuto (PITN-GA-2013-605817).

References


Supplementary materials

Photos of the bridges obtained with Google Streetview. https://www.dropbox.com/sh/bln7zbs0i1h9a7/AAX2SFO2Z9vEfqu-dX4Wpib?dl=0

GPS data of the bridges locations obtained with Google Maps, and GPS data of the button press locations of the participants. Locations in green are hits, red fields are misses, and ... are no data available. Locations in black are not applicable, i.e., those button presses fall outside of the experimental boundaries, and GPS positives have been produced by participant #1 (at 52.3584, -2.0695), #5 (at 52.2675, -1.6380), and #9 (at 52.3056, -1.7004 and 52.2615, -1.6163).

补充材料

获取桥梁的照片使用了Google Streetview。https://www.dropbox.com/sh/bln7zbs0i1h9a7/AAX2SFO2Z9vEfqu-dX4Wpib?dl=0

根据Google Maps获取了桥梁的位置信息，以及根据GPS数据获取了按钮操控位置的信息。绿色的区域表示正确命中，红色的区域表示错误不命中，...表示无数据可用。黑色区域的值不适用，即这些按钮的按压位置超出了实验的边界，GPS的正例由第1位（在52.3584, -2.0695）, #5（在52.2675, -1.6380），和 #9（在52.3056, -1.7004 和 52.2615, -1.6163）提供。