What were they thinking? Subjective experiences associated with automation expectation mismatch

Pär Gustavsson[1*], Trent W. Victor [1], Joel Johansson[1], Emma Tivesten[1], Regina Johansson[1] and Mikael Ljung Aust[1]

[1] Volvo Cars Safety Centre, Volvo Cars, Gothenburg, Sweden
[*]par.gustavsson@volvocars.com

Abstract: The aim of this paper is to gain insight into factors that affect drivers' mental processes and responses to a critical event while driving with supervised automation. Seventy-six drivers participated in a test track experiment ending with a conflict event that required active driver intervention to avoid a crash. About a third of the drivers crashed, despite being provided with instructions on system limitations as well as supervision reminders. Analysis of questionnaire and interview data from the drivers showed that crash outcomes could not be explained by factors often brought up as concerns when discussing supervised automation, such as sleepiness and inattention. Instead, the drivers who crashed did so due to expectation mismatch. This in turn seems to stem from learned trust. Crashers reported higher trust in the automation than non-crashers. Crashers also saw the conflict object but believed the vehicle would be able to resolve the situation on its own. For some drivers, 30 minutes of driving with highly reliable supervised automation thus seems to provide sufficient grounds for developing incorrect expectations on automated function capabilities. These conclusions are relevant to human-automation interaction in general, and for development of driver-state-adaptive supervised automation and advanced driver-assistance systems in particular.

1. Introduction

Development of successful automated driving will depend on recognizing and supporting the two new driver roles that come with driving automation – the delegated and the shared driving role, or unsupervised and supervised automated driving respectively [1].

In unsupervised automation, the driver delegates full control and responsibility to the vehicle, to be free to do something else (e.g. work, watch a film, or even sleep). This requires a vehicle designed for complete support and crash avoidance in all conflict situations (see e.g. [2] and [3]).

Supervised automation, on the other hand, only partly supports the driving task (e.g. headway control and some degree of steering assistance), and the driver is still required to supervise the driving and intervene at sensing or actuation limits (e.g. conflict situations). The driver is thus not free to disengage from the driving task.

In this context it is important to note that meta-analyses [4] has showed that there exists a general relationship between degree of automation and reduction in human performance (such as complacency, skill degradation, and loss of situation awareness). For instance, it has been found that while increased automation improves routine task performance, operators show difficulty troubleshooting and recovering when something unexpected happens [4].

This human performance reduction is largely attributed to operators' tendency to reduce their monitoring of highly reliable automation because of its ability to function properly for an extended period of time (e.g. [5] and [6]). It is simply difficult for humans to monitor automation, or be out of the loop for some time, and then suddenly solve critical issues [7].

Also, note that attention and understanding are often implicitly mixed together in descriptions of monitoring [8] [9], i.e. many assume that as long as a conflict object is perceived, it will be adequately acted upon. However, looking at an object or a road segment does not necessarily mean that cognitive control (top down selection processes) becomes engaged and actions are executed [10] [11].

The difficulty to monitor automation is often referred to as an Irony of Automation [12]. In short, as automation becomes more reliable during routine driving and the operational design domain expands (e.g. more situations, speeds, and road types), drivers may develop misconceptions that the automation can handle all safety conflict situations, leading to driver disengagement and performance reduction.

For road traffic, this means that the better the automation, the less attention drivers will pay to traffic and the system, and the less capable they will be to resume control [1]. Of particular concern are first failure effects. These are circumstances where the operator encounters perfect automation for some period of time, and then "complacency" or overtrust in automation is reflected by a very poor response when the automation fails [13] [14] [15]. In a simulator study it has also been found that takeover requests (from automation to drivers) were perceived as automation failures and temporarily reduced drivers' automation trust, but that trust still was higher in the end of the drive compared to the start. A possible explanation to this is that the takeover request illustrated that the system was not perfectly reliable, and in the long run might have helped drivers to understand the system, thereby increasing trust [16].

Several concepts have been proposed for understanding the human performance degradation associated with increased automation reliability. For example, a meta-review argued that trust formation is a key concept for understanding human relationships with automation [17].

This analysis describes three layers of variability in human-automation trust: dispositional, situational and

learned trust. Of particular interest here is learned trust, which can be further subdivided in two parts: initial learned trust (i.e. trust prior to interacting with the system) and dynamic learned trust (i.e. trust built during interaction). In the context of supervised automation, the former represents what you know about a vehicle automation function before starting to drive, while the latter represents what happens with your trust when you use that function while driving.

It has also been suggested that automation which fails to adapt, i.e. does not change to match the needs of the current situation, may be more susceptible to operator performance degradation [18] [19]. However, while the concept of adaptive automation seems to be viewed as positive, practical applications proving its value are scarce [17]. In the context of supervised automation, one possible implementation of adaptivity would be to let the function remind the supervising operator about monitoring and response readiness responsibilities, should the operator show tendencies to fail at these.

Another concept that is discussed is that of automation transparency. Transparency refers to the extent to which the inner workings or logic of an automated system are transparent to the operator [20]. The general idea is that transparent systems which provide accurate and useful feedback can reduce automation misuse or disuse [17].

However, it is less clear how to apply this in supervised automated driving. While transparency may be intrinsic to establishing initial learned trust (i.e. telling the future function users about function limitations), it is not clear what would constitute transparency in the dynamic phase. Ideas such as displaying which traffic elements the vehicle is actually tracking to the driver [1] to make it easy to spot tracking errors are abundant in online commentaries, but so far, the authors of the current paper are not aware of research where this has been further studied empirically.

The data used in the present paper comes from three experiments examining driver intervention response to conflicts after driving with highly reliable supervised automation following a lead vehicle on a test track [21]. In all experiments a conflict occurred after 30 minutes wherein the lead vehicle cut out of lane to reveal a conflict object in the form of either a stationary car or a garbage bag. Data from the second and third experiment reported in [21] are included in this paper.

In the first (baseline) experiment the test vehicle automatically braked and avoided the conflict. In this experiment, drivers displayed both extreme visual distraction and sleepiness, and few tried to intervene in the conflict. This raised the question of whether these drivers would have avoided the conflict, had the vehicle not intervened on its own.

In the second and third experiment, participants were given more detailed instructions on system limitations and driver responsibilities, Supervision reminders (Attention Reminder and Integrated Attention and Hands on Wheel reminder) were implemented, and the drivers needed to intervene to avoid a crash in the conflict event. Differences between experiments and conditions are described in Table 1 in section 2.1.

Supervision reminders effectively maintained eyes-on-path and hands-on-wheel. However, neither these reminders nor explicit instructions on system limitations and supervision responsibilities prevented 28% (21/76 drivers) from crashing with their eyes on the conflict object. The crash rates were similar across experiments and test conditions [21]. These results highlight the important role of expectation mismatches, showing that a key component of driver engagement is cognitive (understanding the need for action), rather than purely visual (looking at the threat), or having hands-on-wheel [21].

The aim of this paper is to gain insights into the cognitive mechanisms underlying the drivers' expectation mismatch by analysing questionnaire and interview data from the participants in the second and third experiment. Specifically, the perceived relevance of Supervision reminders and which factors that affect drivers' mental processes and actions in a critical event (what the drivers expected themselves to do and the car to do) is explored.

2. Method

This section covers the general test setup, the differences between experimental conditions, data collection, data processing, coding and review, and data analysis. A more detailed method description is provided in [21].

2.1. General method

The same general methodology was used in the two experiments included in this paper (experiment 2 and 3). The main differences between the experiments were the level of instruction detail, supervision reminder type, and conflict scenario type. Key differences between the experiments and conditions are summarized in table 1.

Table 1 *Overview of key differences between experiments and conditions. Supervision Reminder type is either AR (Attention Reminder) or AR&HoW (Integrated Attention and Hands on Wheel reminder).*

Experiment and condition	N	Level of instruction detail	Supervision reminder type	Conflict scenario type
1a	*15*	*Low*	*None*	*Stationary car fully in lane*
1b	*15*	*Low*	*None*	*Stationary car partially in lane*
2	16	Medium	AR	Drift & Garbage bag in lane
3a	15	High	AR	Stationary car partially in lane
3b	15	High	AR	Garbage bag in lane
3c	15	High	AR&HoW	Garbage bag in lane
3d	15	High	AR&HoW	Stationary car partially in lane

2.1.1. Participants:

All 76 participants were Volvo Cars employees. The experiments were set up to achieve a between-group design. The selected participants could not be involved in driving automation development, could not work as test drivers, had not participated in similar studies before, and had a minimum driving experience of at least 5000 km during the previous year.

In experiment 2, 16 participants were included in the final sample, 2 females and 14 males. Ages spanned from 27-66 years (M = 45.9, SD = 12.0) with driving experience spanning from 6-49 years (M = 26.9, SD = 13.7).

In experiment 3, 60 test participants were included in the final sample, 18 female and 42 male. Ages spanned from 26-65 years (M = 45.2, SD = 9.6) and driving experience from 1-47 years (M = 25.3, SD = 10.4).

2.1.2. Materials, procedure and scenario design:

On arrival, the participants received general information about the test and were asked to read through written participant information as well as sign an informed consent form. The participants were also asked to fill in a pre-drive questionnaire in order to provide driver background information. The stated purpose of the study was to evaluate driver experiences during automated driving.

Next, the participants were introduced to the test vehicle (TV), a Volvo XC90 (MY2016). The original XC90 Driver Information Module (DIM) was modified to display a customized supervised automation HMI which also could present attention reminders to participants according to predefined thresholds for visual inattention, a similar algorithm to [22], and the MDD algorithm [23]. The TV was equipped with special, test-unique software which had self-driving capability to precisely follow the road, maintain speed, and keep a constant headway with highly-reliable driving performance behind a robot-controlled XC90 lead vehicle (LV) on the AstaZero rural road test track.

While sitting in the driver's seat, the participants received further verbal information about the test and the vehicle, along with an introduction to the Karolinska Sleepiness Scale (KSS) [24]. The participants reported their sleepiness level (KSS 1-9) before the drive started and once every lap on the test track (approximately every 6 minutes). The participants were instructed to supervise the car throughout the drive and were also told that they could override the automation by steering or braking at any time. Two test leaders rode along in the backseat of the TV; one who administered the KSS scale and monitored a video stream of the driver to trigger supervision reminders according to pre-defined rules for eyes-off-path and hands-off-wheel, and another who acted as (back-up) safety driver. There was no conversation with the test participants during the drive, except when asking for KSS scores.

The TV followed behind the LV which kept a speed of 70 kph, except through some curves where speed was lowered by the LV to about 50 kph. The same (pre-recorded) LV path and velocity was used for all participants. After 30 minutes (five laps), the test vehicle encountered a conflict object placed in the driving lane; either a stuffed garbage bag (figure 1a) or the ADAC Advanced Emergency Braking System Stationary Target (stationary car) (figure 1b). The conflict object was positioned so that the participants could see it when passing through a curve and crest just prior to the event, 14.0 s before reaching the conflict object. The conflict object then became obscured again by the LV when the road straightened out. About 20 meters from the conflict object, the LV did a cut-out (an evasive steering manoeuvre around the object) revealing the conflict object in full to the participants, about three seconds before reaching the object. The TV did not brake or warn the drivers in any way, and the DIM displayed the same HMI throughout the whole drive.

After the conflict, the participants were asked to stop the car and fill in a post-drive questionnaire, which also served as a basis for a semi-structured interview. All interviews were recorded. After the interview, the full purpose of the study was disclosed.

Figure 1a - *Garbage bag used in experiment 2, 3b and 3c.*

Figure 1b – *Stationary car used in experiment 3a and 3d.*

2.2. Experiment 2

Experiment 2 (E2) examined if driver intervention performance is associated with first failure effects [13]. This was done by exposing the participants to a drift out of lane event after 15 minutes, and a conflict situation with a garbage bag inside the lane after 30 minutes. In the drift event, the vehicle drifted over into the left-adjacent lane and returned to the right lane after some time if the driver did not intervene (between 8 and 18 s for the participants who did not steer back).

In experiment 2, participants were given a medium level of instruction – written instructions that emphasized the driver's role as supervisor, the limitations of the vehicle, and the driver's responsibility for the safety of the vehicle even when the automation was engaged. The instruction also stated that the drivers needed to apply more force to override the steering when the automation was engaged compared to what is needed for normal steering in manual driving. A key excerpt from these instructions:

"The car you will drive is a so called Supervised automated drive car which means that the car itself, under certain circumstances and on chosen road stretches, can control steering and adapt speed and distance. Due to limitations in the car's sensor platform the driver can't yet engage in non-

driving activities, and you are instead expected to supervise the drive at all times, as you would in manual driving."

Participants received attention reminders in the DIM (warning messages in the instrument cluster behind the steering wheel) if they were visually inattentive (determined from patterns of off-path eye glances). Two levels of attention reminders were used. Both levels of reminders were presented for a maximum of 7.0 s without any notification sound. If participants looked back on the road for at least 2.0 s after the reminder or were judged to be more attentive, the system reset and the reminder disappeared from the display.

Level one reminders (figure 2) were issued if a single off-path glance longer than 3.4 s was detected, or if the driver had been looking predominantly off-path for a period of 12.0 s (total glance duration history).

Level two reminders were issued for single off-path glances longer than 7.0 s, if the driver's attention did not return to the road after having been issued a level one reminder, for eye-closures longer than 3.0 s, or if they received a new level one reminder within 10.0 s of a level one or two reminder. The only visual design difference for the level two reminder was a red icon. In addition, the level two reminder was combined with a soft deceleration of the test vehicle.

Figure 2 – *Level one Attention Reminder in E2.*

2.3. Experiment 3

Experiment 3 (E3) examined if more detailed instructions, updated Attention Reminder rules, and adding an Integrated Hands on Wheel and Attention Reminder improved driver intervention performance. The specific TV used in E3 had slightly improved lane keeping capability and slightly more steering wheel resistance when the automation was engaged than the TV in E2. All participants in E3 were instructed to override the steering approximately one minute into the drive to experience the steering wheel resistance and to minimise potential effects of this on crash outcomes.

The participants were exposed to a conflict situation after 30 minutes with either a stationary car partially inside the lane (condition 3a & 3d) or a garbage bag inside the lane (condition 3b & 3c).

In experiment 3, participants were given a high level of instruction by attending a 30 minute classroom training prior to their drive. The training covered these areas:

Driver responsibilities. The driver is responsible, should monitor, supervise and intervene whenever needed. The driver needs to be active and attentive at all times and supervise the traffic so that the car is driven in a safe manner for passengers in the vehicle and surrounding traffic. Sensors and cameras judge the driver's ability to actively supervise the automation and traffic and detects if the driver has their hands on the steering wheel or if the driver looks on the road. Drivers will get notifications after periods of inattentiveness or inactivity and the system will deactivate after a longer period of inactivity.

System limitations. Objects and obstacles in the traffic environment, such as potholes in the roadway, high curbs, and objects on road are not detected. Obstacles can also be falsely detected as lane markings and thus pose a risk that the car will collide with these obstacles. Cameras and sensors have a limited field of view. Indistinct lane markings might lead to erroneous steering by the automation. Other limitations may occur with road design (e.g. roadworks), oncoming vehicles, pedestrians, and animals. There are restrictions in the steering, braking, and acceleration force that can be applied by the system.

Instruction videos & risk scenarios. Videos of risk scenarios were shown, including a video showing when a car starts to depart from the roadway and the driver needs to steer back in lane and then let the function resume control. A risk scenario where the function does not detect obstacles in the roadway was explained, in which the driver needs to brake and/or steer away from the obstacle and after that let the function resume control.

All participants received attention reminders if they were visually inattentive. The Attention Reminder was updated based on feedback received in experiment 2. Participants in condition 3a and 3b were not required to have their hands on the steering wheel as long as they stayed visually attentive. Three levels of attention reminders were used.

Level one reminders (figure 3, left) were issued in the DIM if participants had been looking predominantly off-path during a period of > 17.0 s (total glance duration history).

Level two reminders used the same message but added a sound and were triggered either by a 3.4 s off-path glance, an eye closure longer than 3.0 s, or if they received a new reminder within 10.0 s of a level one or two reminder.

Level three reminders were issued as a text message *"Autopilot deactivated – Driver inattention"* with a hands-on-wheel icon and a more urgent sound if a 15.0 s glance off path was detected, or a 15.0 s eye closure, or if they were glancing more than 75% off path in a period of 20.0 s (glance history), or if they were to receive a third level two reminder within 15.0 s.

Participants in condition 3c and 3d were required to always keep their hands on the steering wheel and received Hands on Wheel (HoW) reminders if they failed to do so. Thus, drivers in these conditions could experience both attention reminders and Hands on Wheel reminders at different periods during the same trip.

Two levels of Hands on Wheel reminders (figure 3, right) were issued in the DIM.

Level one HoW reminders were issued if hands were off the steering wheel for more than 5.0 s.

Level two HoW reminders used the same message and icon but added a sound and were issued if hands were off the steering wheel for more than 10.0 s.

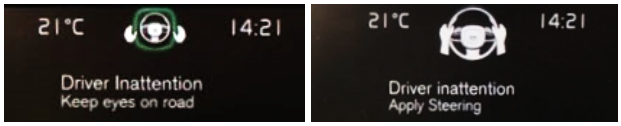

Figure 3 – *Supervision Reminder messages in E3. Left: Attention Reminder message. Right: Hands on Wheel reminder message.*

2.4. Data processing, coding and review

Video of the conflict event for all participants was reviewed in order to determine crash outcome [21]. Any contact between TV and the conflict object was classified as a crash.

All questionnaire responses and rating scales were compiled into a data set. The interview recordings (ranging between 8 and 36 minutes long per participant) were transcribed. The transcriptions were transformed into open codes separated by interview question and participant. In addition, the free text responses in the questionnaires were coded and added to the transcription codes. Themes were created by processing the codes related to one specific interview question or created by processing codes from different parts of the interview, according to content analysis methodology [25]. The themes and included codes were reviewed among the authors in order to reach consensus on the themes.

The process from transcription to themes is presented by the following example:

Questionnaire item: Did you perceive that the object was in the lane before the lead vehicle steered away (y/n)? If yes, when?

Questionnaire item response: No.

Transcription: "No I did not. It was not until the lead vehicle steered away that I understood that there was something there"

Open coding: Did not perceive object until LV manoeuvre

Final theme: I perceived the object late

The categorization of driver actions was based on responses to the interview question "*Did you intervene in the situation? If yes, how and why? If no, why?".* Responses indicating that the drivers intervened without delay were categorized as *I intervened*. Responses that explicitly described interventions as late or delayed waiting for intervention by the automation were categorized as *I intervened late*. Furthermore, if a participant stated that they realized the need to intervene late and intervened, they were categorized as intervening late as well. The category *I intervened too late* includes statements of acting too late to be able to avoid a crash. If a participant answered no to the interview question or stated that she only put her hands on the steering wheel, it was categorized as *I did not intervene*.

The categorization of realization of the need to act was based on responses to the interview question *"Did you realize that you needed to intervene to avoid a crash? If yes, when did you realize that?"*. The category *I realized the need* includes statements that the participants realized the need to intervene before or at the time of the LV cut-out manoeuvre. *I realized the need late* includes explicit statements that they realized it late, after the LV manoeuvre, or when they were close to the object. The category *I realized the need too late* includes explicit statements that the realization was too late to be able to act and avoid a crash. If a participant responded no to the interview question or stated that they realized the need to intervene after crashing it was categorized as *I did not realize the need*.

The categorization of conflict object perception was based on responses to the interview question "*Did you perceive that the object was in the lane before the lead vehicle steered away? If yes, when?".* The category *I perceived the object early* means before the LV cut-out manoeuvre (14 s before the conflict point), while *I perceived the object late* includes perception at the time of the LV manoeuvre or later. Some participants also responded that they perceived the object before the LV manoeuvre but were not sure if it was positioned in or outside the lane, creating the category *I perceived the object early, uncertain if in lane.*

The categorization of expectations on the automation was based on an analysis of the responses to several different interview questions, mainly *"Did you intervene in the situation? If yes, how and why? If no, why?", "Did you realize that you needed to intervene to avoid a crash? If yes, when did you realize that?"* and *"To what extent did you trust the automated vehicle to be able to handle the situation?".* Data from drivers that expressed that they were expecting an intervention from the automation were categorized as *I expected an intervention*. Drivers that expressed uncertainties regarding the automation's ability to intervene were categorized as *I was uncertain about an intervention*. The drivers that stated that they did not expect the automation to intervene were categorized as *I did not expect an intervention.*

The categorization of trust in the automation was based on ratings on a Likert scale (1 not at all, 7 completely) to the question *"To what extent did you trust the automated vehicle to be able to handle the situation?"*. The category *I had high trust* includes ratings 5-7, the category *I was neutral* includes rating 4, and the category *I had low trust* includes ratings 1-3.

2.5. Data analysis and visualizations

The participants' experimental condition, crash outcome, ratings and themes were compiled into a Microsoft Excel sheet. This enabled filtering of different variables in order to analyse the data using a combination of qualitative and quantitative methods.

The drivers' actions and mental processes related to the conflict situation were visualized in a format inspired by the graphical representation used in the DREAM methodology [26].

3. Results

This section presents the results from the analysis of the questionnaire and interview responses.

3.1. Attention Reminder relevance

To study the participants' experiences of attention reminders, the participants who were aware of receiving reminders in experiment 2 and 3 rated the relevance of them on a scale between 1 (not at all relevant) and 7 (very relevant). There was an increase in rated relevance of attention reminders from E2 (M=4.17, SD=1.70, N=12) to E3 (M=5.75, SD=1.25, N=48); t(14)=-3.03, p=0.009. The increase in relevance of the reminders in E3 was independent of whether hands were on wheel or not.

The drivers were also asked to explain their relevance rating of attention reminders. In E2 only 17% (2/12) of the participants expressed that the reminders were warranted or relevant compared to 77% (37/48) in E3. Further, 67% (8/12) of the participants in E2 found the system to be too sensitive or remind too frequently, which only 23% (11/48) expressed in E3.

Thus, the rated relevance of attention reminders increased on average from a neutral level in experiment 2 to a high level in experiment 3.

3.2. Hands on Wheel reminder relevance

The participants who were aware of receiving Hands on Wheel reminders in condition 3c and 3d rated the relevance of the reminders on a scale between 1 (not at all relevant) and 7 (very relevant). The Hands on Wheel reminders were on average rated as highly relevant (M=5.86, SD=1.07, N=7).

3.3. General themes

This section presents some general reoccurring themes in the data that are relevant to the supervising role drivers were instructed to take in these experiments.

3.3.1. Sleepiness:

An issue mentioned by 32% (n=24/76) of the participants was the risk of becoming sleepy while supervising, especially during longer drives. Only two participants stated in the interview that they were feeling less or equally sleepy compared to manual driving. Some example quotes that highlight problems with sleepiness are (words in [] represent test leader utterances while words in () are added for clarification):

"I felt that I got drowsy after a while… [if we had driven for thirty more minutes, do you think you would have stayed awake?] I would probably have ended up on an 8 on the scale (KSS), I definitely think so… I would have stayed awake but I probably would have needed to fight it a bit…"

"The more you trust the car, the sleepier you get"

"I was very tense the whole drive… it was comfortable to ride along but it felt more laborious than driving yourself since I know that I must control but not drive… know that I must have control but at the same time not have control because it is the car that drives. It is probably a matter of habit but I think that it took a lot of energy and effort since I also do not know how the car will react. When I drive myself I have complete control and I am aware that it is only me that has complete control. When I did not have that it felt like I strained myself very much more and became very tired in the end because it takes a lot of energy"

However, sleepiness does not seem to be a factor that explains why some drivers crashed. On average, the crashers in experiment 3 rated themselves as more tired on the KSS scale for the last lap (M=5.06, SD=1.18, N=16) than non-crashers (M=4.30, SD=1.55, N=44); t(34)=2.04, p=0.049, but the difference is small and in the middle of the scale. Also, there were no clear differences observed in E2, and only one participant (in E3) reported extreme sleepiness (KSS≥8) during the last lap.

3.3.2. Attentiveness:

Another common concern expressed by 24% (n=18/76) of the participants was problems staying attentive while supervising the drive. Some examples of this are:

"I would like something to keep me more active during the drive, because it is hard to keep your eyes on the road this long"

"It is hard to stay focused, your mind wanders"

"I thought it was hard to be focused but still not have an active role. Even if you sat and looked on the road I experienced that I had lots of other thoughts inside my head"

"When you feel safer you lose focus more and more. You lose the ability to concentrate".

However, in the current study, there was no clear mapping between concerns expressed around attention and crash involvement.

3.3.3. Issues with the supervising role:

A reoccurring theme expressed by 18% (n=14/76) of the participants was that they found the supervising role to be problematic to carry out or hard to grasp. Furthermore, some participants also said that they became passive (9%, n=7/76) or that they became more of a passenger in the driver's seat (7%, n=5/76). This is exemplified by the following quotes:

"It was exciting but I don't know for how long you are able to handle it. The better it becomes, the harder it gets… This is a very big problem because when something really happens it is catastrophic when you are too far away to react. Then it is basically over."

"… somehow you cannot relax completely as you can in the passenger seat when you know that someone else has the responsibility, now it feels like a middle stage… it is a bit unclear what you should do when there is something on the road"

"I had great trust in how the car acted and got lost in my active role and towards the end I sat and wondered why I should sit here and have such a passive role [and still need to have your hands on the wheel] yes"

"When you have to sit and supervise the whole time it detracts the whole thing I would say. Because if you still need to sit and supervise and check everything then you could just as well do it yourself. Because then you have something to do in the meantime."

"It felt a bit unaccustomed in the beginning, but after a lap you got used to the thought of actually being a passenger but behind the wheel."

There was however no clear difference between crashers and non-crashers regarding reported difficulties with the supervising role.

3.3.4. Education:

The participants in experiment 3 were asked to rate to which extent the driver education prepared them for the drive ranging between 1 (not at all) and 7 (completely). The average rating was 5.63 (SD=1.45, N=16) for crashers and 6.23 (SD=1.10, N=44) for non-crashers. The participants were satisfied with the education and there was no clear

difference in rating that could potentially explain crash outcome.

3.4. Driver actions and experiences during the drift event (E2)

In experiment 2, 38% (n=6/16) of the drivers did not intervene when the car drifted over into the adjacent lane. Four of these six participants did also crash with the garbage bag in the conflict event. One participant stated that she did not trust the car in the conflict event because of the earlier drift event. All participants stated post drive that they noticed the drift event during the drive.

The most common themes for non-intervening participants (note that some had statements in several themes, so those below are not mutually exclusive) were that they thought that it was a part of the test (3/6), that they would have acted sooner on a public road (2/6), that they looked ahead for obstacles instead of intervening (2/6), that they observed and got ready to intervene (2/6) and in one case the driver was afraid of aborting the automation and thus the test.

3.5. Driver actions and expectations during the conflict situation

This section presents subjective data acquired post-drive specifically related to the conflict situation.

3.5.1. Driver actions:

The subjective data results regarding driver actions are presented in table 2.

Table 2 *Themes – Driver actions E2+E3*

Theme	% (n)	% Crashers (n)	% Non-crashers (n)
I intervened	37% (28)	0% (0)	51% (28)
I intervened late	41% (31)	19% (4)	49% (27)
I intervened too late	7% (5)	24% (5)	0% (0)
I did not intervene	16% (12)	57% (12)	0% (0)

Table 2 shows that a majority of the drivers intervened. All of the non-crashers intervened or intervened late, while a majority of the crashers did not intervene.

3.5.2. Realization of the need to act:

The subjective data results regarding realization of the need to act are presented in table 3.

Table 3 *Themes – Realization of the need to act E2+E3*

Theme	% (n)	% Crashers (n)	% Non-crashers (n)
I realized the need	38% (29)	10% (2)	49% (27)
I realized the need late	30% (23)	14% (3)	36% (20)
I realized the need too late	13% (10)	48% (10)	0% (0)
I did not realize the need	18% (14)	29% (6)	15% (8)

As shown in table 3, a majority of the drivers realized the need to act. Some non-crashers expressed that they did not realize the need to act but instead acted instinctively. A majority of the crashers realized the need to act too late or not at all. Explanations to the five other crashers are presented in section 3.5.6.

3.5.3. Conflict object perception:

The subjective data results regarding perception of the conflict object are presented in table 4.

Table 4 *Themes – Perception of the object E2+E3*

Theme	% (n)	% Crashers (n)	% Non-crashers (n)
I perceived the object early	68% (52)	43% (9)	78% (43)
I perceived the object early, uncertain if in lane	8% (6)	10% (2)	7% (4)
I perceived the object late	24% (18)	48% (10)	15% (8)

Table 4 shows that all drivers perceived the object at some point in time. A majority perceived the object early, of which a few were uncertain about the lateral position of the object. Among the crashers, 48% (n=10/21) did not notice the object until it was fully revealed after the LV cut-out manoeuvre, compared to 15% (n=8/55) of the non-crashers. All of the 10 crashers that did not notice the object early did however have their eyes on-path towards the object when it was visible in the curve/crest prior to the lead vehicle cut-out manoeuvre [21].

3.5.4. Expectations on the automation:

The subjective data results regarding expectations on the automation being able to intervene in the conflict situation are presented in table 5.

Table 5 *Themes – Expectations on the automation E2+E3*

Theme	% (n)	% Crashers (n)	% Non-crashers (n)
I expected an intervention	54% (41)	95% (20)	38% (21)
I was uncertain about an intervention	28% (21)	5% (1)	36% (20)
I did not expect an intervention	18% (14)	0% (0)	25% (14)

As shown in table 5, a majority of the drivers expected the automation to intervene in the situation. All crashers except one expected the automation to intervene. Only 18% (n=14/76) of the drivers did not expect the automation to intervene.

Participants were also asked if they felt that they received enough information from the vehicle during the drive. Half of the participants (n=38/76) expressed that they wanted the car to warn them of the target and/or that they needed to intervene manually.

3.5.5. Trust in the automation

The subjective data results regarding trust in the automation vehicle being able to avoid a crash are presented in table 6.

Table 6 *Categories – Extent of trust in the automated vehicle E2+E3*

Category	% (n)	% Crashers (n)	% Non-crashers (n)
I had high trust	61% (46)	100% (21)	45% (25)
I was neutral	5% (4)	0% (0)	7% (4)
I had low trust	34% (26)	0% (0)	47% (26)

Table 6 shows that the non-crashers were close to equally represented in the high and low trust category, while all the crashers were categorized as high trusters.

In figure 4 the frequency of trust ratings separated by crash outcome are shown.

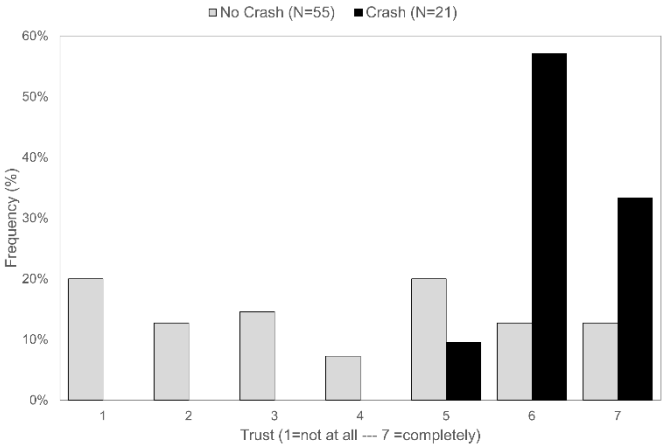

Figure 4 - *Rated trust in the automation being able to handle the situation for crashers (black bars) and non-crashers (grey bars).*

The drivers rated trust in the automation being able to handle the conflict on a medium or neutral level on average (M = 4.50, SD = 2.10, N=76). The crashers reported higher trust in the automation (M = 6.24, SD = 0.62, N=21) compared to the non-crashers (M = 3.84, SD = 2.09, N=55); t(71) = 7.68, p=0.000. There were no clear differences in trust ratings dependent on experiment, object type, or hands on wheel condition.

3.5.6. Visualization of actions and mental process

To sum up the driver actions and expectations in the conflict situation, the following figures visualize the different themes and the links between them. The charts are read from right to left, starting with the outcome which is then linked to different themes. The numbers in the boxes and on the links show the frequency of occurrence in the aggregated chart. The most common themes are marked in grey which together with the black links show the most common patterns.

Figure 5 visualizes role model performance (the "optimal process") of a supervising driver, found in the data of 11 participants (all non-crashers).

Figure 6 visualizes the process of all 21 crashers. There are five special cases that need to be explained. Two of the four crashers that were categorized as *I intervened late* did not notice that there was any contact between the TV and the conflict object and therefore described their realization and action as late rather than too late. The third crasher was aware that a crash occurred but expressed that she did not apply enough steering power rather than intervening too late. The fourth crasher is a participant that expressed that she realized the need to act but were uncertain how to act since it was a test situation. The final special case is a crasher that expressed that she realized the need to act but chose to give the automated vehicle a chance to solve it since it was a test situation.

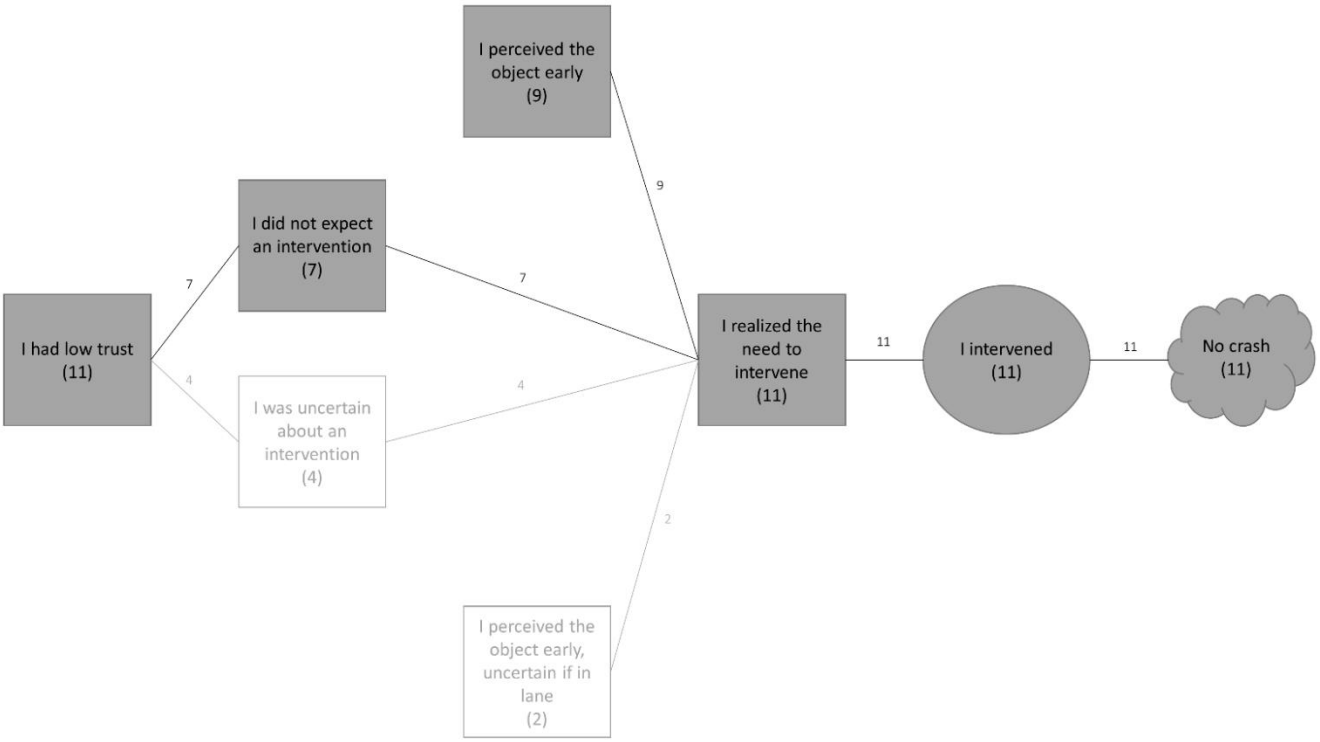

Figure 5 – *Visualization of outcome and themes for role model drivers (drivers with an optimal response in the conflict situation, n=11).*

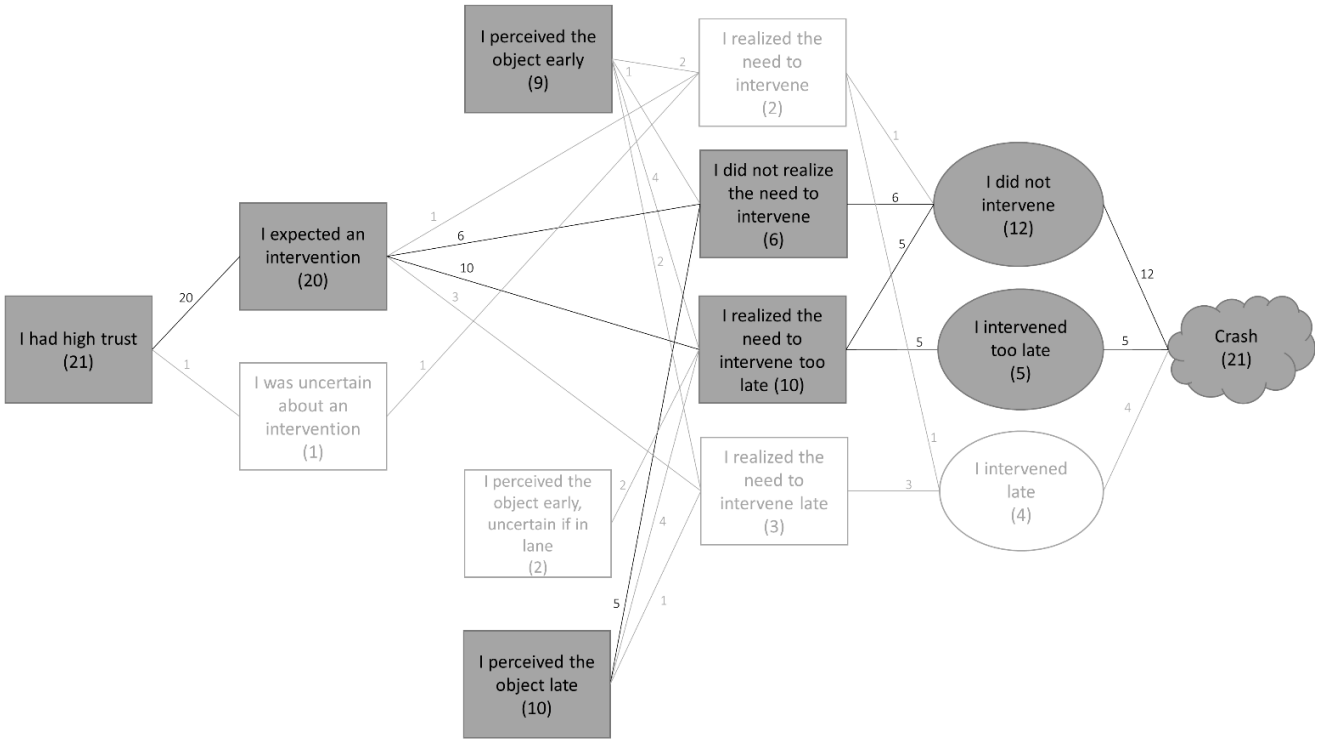

Figure 6 - *Visualization of outcome and themes for all drivers that crashed in the conflict situation (n=21).*

3.6. Overall factors affecting trust in the automation

This section presents participants' explanations to their ratings of trust in the automation being able to avoid a crash. The final themes are based on, but not limited to, the explanations of trust given when the participants rated their trust. This was supplemented by taking other parts of the interview into account, since factors affecting trust were found spread over the entire interview.

The results are presented mainly based on trust ratings independent of crash outcomes. To clarify, some participants had statements in several themes, providing themes that are not mutually exclusive. Additionally, some participants did not have any statements that were included in the final themes. To clarify the theme *Expected vehicle able to detect object and intervene*, this included explicit statements of trusting the automated vehicle based on expectations on it being able to detect the specific conflict object and/or handle the specific conflict situation as it unfolded. This theme is thus more explanatory than the more general *I expected an intervention* theme in section 3.5.4. The theme *Uncertain driver responsibility* is related to the general issues with the supervisor role found in section 3.3.3, but is strictly connected to trust and responsibility in the conflict situation (i.e. uncertainty regarding who was responsible to intervene since they believed both were able to).

The most common themes among the drivers that rated their trust as high (5-7) were that they expected the vehicle to be able to detect the object and intervene (52%, n=24/46), that they based their trust on the vehicle's good driving performance (28%, n=13/46), that they felt safe during the drive (15%, n=7/46), and that they felt uncertain about their responsibility in the conflict situation (15%, n=7/46). Note that uncertainty about responsibility in the conflict situation was expressed only by crashers; none of the non-crashers expressed the same uncertainty (see figure 7).

Comparing E2 to E3, only 13% (n=1/8) of the high trust participants in E2 based their trust on the vehicle's good driving performance, compared to 32% (n=12/38) in E3. Also, 25% (n=2/8) expressed uncertainties regarding their responsibilities in E2 compared to 13% (n=5/38) in E3.

The four drivers that rated their trust as neutral (4) did not elaborate their rating further. Looking into expectations however, three out of four were uncertain about the car intervening in the conflict, while one did not expect an intervention.

The most common themes among the drivers that rated their trust as low (1-3) were instead that they had insufficient knowledge about the capabilities of the automation (27%, n=7/26), that the driver is responsible to handle events like this (15%, n=4/26), that they felt that they wanted to handle the situation themselves (15%, n=4/26), and that they did not expect the vehicle to be able to detect the object (15%, n=4/26). Comparing the low trust participants in E2 and E3, 33% (n=2/6) stated that they did not have enough knowledge of the capabilities of the automation in E2 while 25% (n=5/20) stated this in E3. No participant with low trust in E2 expressed that it was the driver's responsibility to handle situations like this compared to 20% (n=4/20) in E3.

The most common theme related to high trust was expecting the car to be able to detect the object and intervene. One reoccurring explanation to this expectation was the perceived good driving performance of the automation. Thus, there is a relationship between driving performance, trust and expectations. In contrast, the themes of the low trust drivers were in line with the content of the education given to drivers in E3. In figure 7 the frequencies for each theme among drivers that rated their trust as high (separated by crash outcome) and low (all non-crashers) are presented.

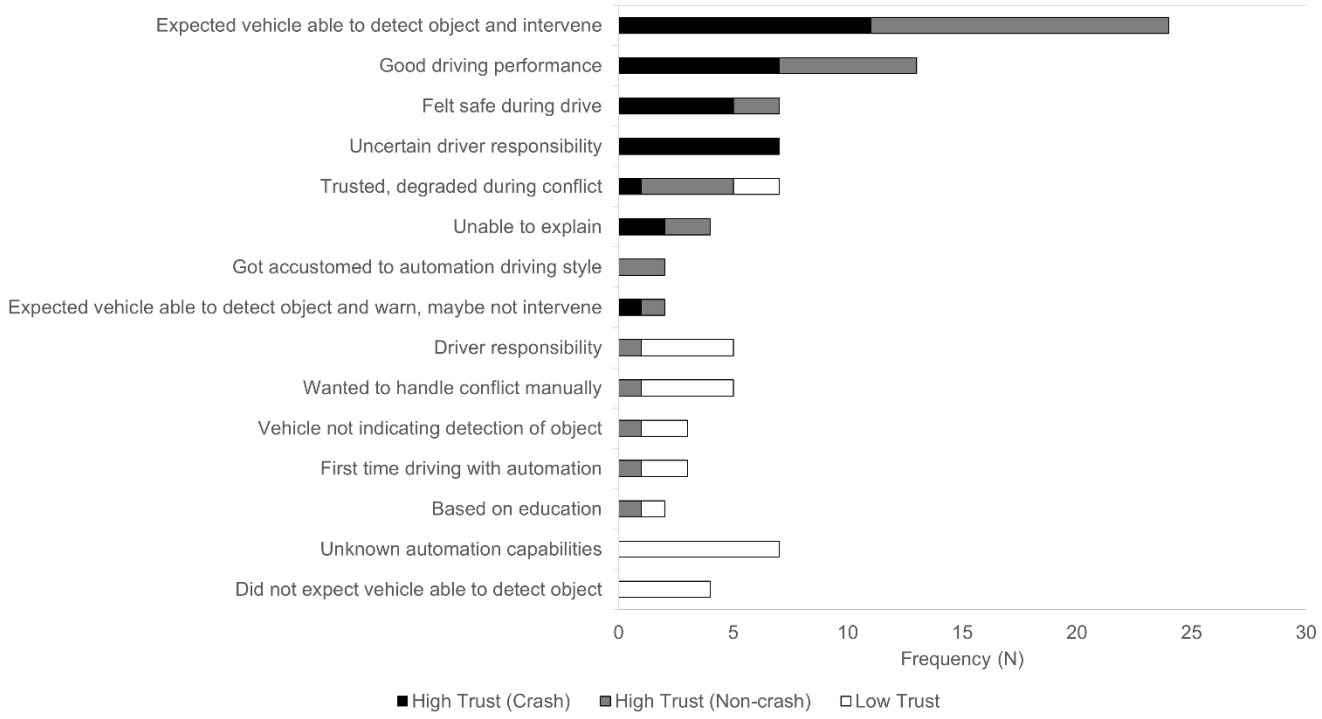

Figure 7 - *Frequency of trust themes in participants that reported high trust and crashed (black), high trust and did not crash (grey), and low trust (white).*

4. Discussion and conclusions

Explanations of the mental processes associated with the conflict event in crashers and non-crashers were highly informative. There are several interesting findings in the subjective data that warrant further discussion.

First, the behaviours and themes often brought up as concerns in the discussion of supervised automation (i.e. sleepiness, inattention, problems with the supervising role as such and pre-drive education) were not found to be highly explanatory in distinguishing between crashers and non-crashers in the subjective data from these experiments; they were quite equally expressed among participants independent of crash outcome.

Second, it certainly seems possible to engineer a supervision reminder that is both effective in terms of keeping eyes on road and hands on wheel and has high user acceptance. However, it is also evident that having eyes on road and hands on wheel does not equate to being sufficiently in the loop to act on imminent conflicts in supervised automation.

Third, interesting patterns regarding trust and expectations emerged.

To start with, a majority of the participants did realize the need to act and managed to avoid a crash. Most of these drivers reported seeing the conflict object early on, did not have very high levels of trust in the automation and were not necessarily expecting the vehicle to act. Taken together, these subjective reports correspond well with their conflict response, i.e. to actively intervene to avoid the conflict object. Interestingly though, there were only 11 role model drivers (14%) who acted and explained their actions in full accordance with the content of the instructions.

The crashers on the other hand generally reported high levels of trust in the automation, detected the conflict object later than non-crashers and expected the vehicle to deal with the conflict object on its own. In fact, almost half of the crashers did not perceive the object until it was fully revealed despite having their gaze in the direction of the object when it was visible earlier on. There was a clear expectation mismatch as crashers expected the automation to detect the object and intervene (see table 5 and figures 6 and 7). This is a subjective data pattern that intuitively correlates well with crash outcome. Also, exposing drivers to a drift event in experiment 2 did not reduce reported trust in the conflict event 15 minutes later, similar to the findings in [16].

However, there is also a group of non-crashers who do not conform to this pattern. These drivers reported high levels of trust in the automation and were expecting the vehicle to intervene, just like the crashers did. The major difference between these high trusting non-crashers and crashers is that they did not crash. Perhaps they thereby could be considered as near-crashers, but near-crashes are typically defined from physical rather than mental closeness to crash. Further investigation of this is required.

The obvious question to ask here is of course why did this sub-group not crash? What is it that distinguishes this sub-group of non-crashers from the crashers, in spite of their very similar subjective data?

The only finding in the present analysis that could provide some answers is found in the trust explanations in section 3.6, namely the uncertainty about driver responsibility in the conflict situation that a third of the crashers expressed. Since this uncertainty was not present in the data of the non-crashers, it seems like the mental models on task allocation could differ between these high trusting crashers and non-crashers.

Another possibility is that the crash outcome represents an artificial dichotomy imposed on an underlying response time continuum. Simply put, while all these drivers may have realized the need to intervene at some point in time, these who did not crash were the faster ones to do so. It is possible that combining the subjective reports with other recorded driver and vehicle data may shed further light on this issue.

Yet another possibility is that the difference stems from some deeper, underlying trait not captured in the subjective data analysed here. It is possible that the predictive processing framework [27] can provide answers and this should be further investigated. For example, the most common theme among drivers who expressed a high level of trust in the automation was expecting the vehicle to be able to detect the conflict object and intervene, despite prior training on system limitations. The most common reason for this expectation was the perceived good driving performance of the vehicle during the 30 minutes prior to the conflict. Another possible interpretation here is therefore that crashers represent drivers who are more susceptible to dynamic learned trust [17]. For them, those 30 minutes of uneventful driving in a highly reliable automated vehicle with good driving performance was enough to generate first failure effects, i.e. incorrect predictions of avoidance capability. This naturally leads to the question if more of the high trusting non-crashers would have crashed if the drive had been longer in time, or repeated on another day. That is a very interesting topic for future research.

In sum, expectation mismatch is clearly evident in subjective explanations of mental processes in drivers who crash during supervised automated driving.

5. Acknowledgments

This research was supported by the Swedish FFI project ADEST, grant number 2014-06012, within the Drive Me programme. This research complied with the tenets of the Declaration of Helsinki and was approved by the Regionala Etikprövningsnämnden i Göteborg, Dnr:369-16.

6. References

[1] B. Seppelt and T. Victor, “Potential solutions to Human Factors Challenges in Road Vehicle Automation,” *Road Vehicle Automation*, vol. 3, pp. 131-148, 2016.

[2] T. Victor, M. Rothoff, E. Coelingh,et al., “When Autonomous Vehicles Are Introduced on a Larger Scale in the Road Transport System: The Drive Me Project,” in *Automated Driving: Safer and More Efficient Future Driving*, D. Watzenig and M. Horn, Eds., Cham, Springer International Publishing, 2017, pp. 541-546.

[3] M. Lindman, I. Isaksson-Hellman and J. Strandroth, “Basic numbers needed to understand the traffic safety effect of Automated Cars,” in *IRC-17-40 IRCOBI Conference 2017*, 2017.

[4] L. Onnasch, C. D. Wickens, H. Li, et al., "Human Performance Consequences of Stages and Levels of Automation An Integrated Meta-Analysis," *Human Factors: The Journal of the Human Factors and Ergonomics Society, 0018720813501549,* 2013.

[5] R. Parasuraman, T. Sheridan and C. Wickens, "Situation Awareness, Mental Workload, and Trust in Automation, Empirically Supported Cognitive Engineering Constructs," *Journal of Cognitive Engineering and Decision Making,* vol. 2, no. 2, pp. 140-160, 2008.

[6] R. Parasuraman and D. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors: The Journal of the Human Factors and Ergonomics Society,* vol. 52, no. 3, pp. 381-410, 2010.

[7] N. Strand, J. Nilsson, I. Karlsson, et al., "Semi-automated versus highly automated driving in critical situations caused by automation failures," *Transportation Research Part F: Traffic Psychology and Behaviour,* vol. 27, pp. 218-228, 2014.

[8] SAE, "SAE J3016 Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicle.," 2016.

[9] N. Merat, B. Seppelt, T. Louw, et al., "The "Out-of-the-Loop" Concept in Automated Driving: Proposed definition, measures and implications.," *Cognition, Technology and Work.,* (in prep).

[10] E. Miller and J. Cohen, "An Integrative Theory of Prefrontal Cortex Function," *Annual Reviews of Neuroscience,* vol. 24, pp. 167-202, 2001.

[11] M. Corbetta and G. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Reviews Neuroscience,* vol. 3, pp. 201-215, 2002.

[12] L. Bainbridge, "Ironies of automation," *Automatica,* vol. 19, pp. 775-779, 1983.

[13] C. Wickens, B. Hooey, B. Gore, et al., "Identifying Black Swans in NextGen: Predicting Human Performance in Off-Nominal Conditions," *Human Factors,* vol. 51, pp. 638-651, 2009.

[14] R. Molloy and R. Parasuraman, "Monitoring an automated system for a single failure: Vigilance and task complexity effects," *Human Factors,* vol. 38, no. 2, pp. 311-322, 1996.

[15] U. Metzger and R. Parasuraman, "Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload," *Human Factors,* vol. 47, pp. 35-49, 2005.

[16] S. Hergeth, "Automation Trust in Conditional Automated Driving Systems: Approaches to Operationalization and Design," 10.13140/RG.2.2.35823.02724. , 2016.

[17] K. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors,* vol. 57, pp. 407-434, 2015.

[18] D. Kaber and M. Endsley, "The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task," *Theoretical Issues in Ergonomics Science,* vol. 5, no. 2, pp. 113-153, 2004.

[19] E. de Visser and R. Parasuraman, "Adaptive Aiding of Human-Robot Teaming," *Journal of Cognitive Engineering and Decision Making: Effects of Imperfect Automation on Performance, Trust, and Workload,* vol. 5, no. 2, pp. 209-231, 2011.

[20] Y. Seong and A. Bisantz, "The impact of cognitive feedback on judgment performance and trust with decision aids," *International Journal of Industrial Ergonomics,* vol. 38, no. 7-8, pp. 608-625, 2008.

[21] T. W. Victor, E. Tivesten, P. Gustavsson, et al., "Automation Expectation Mismatch: Incorrect Prediction Despite Eyes on Threat and Hands on Wheel," *Human Factors,* 2018, (accepted).

[22] T. W. Victor, Keeping Eye and Mind on the road. Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 9, Uppsala: Acta Universitatis Upsaliensis, 2005.

[23] J. Lee, J. Moeckli, T. Brown, et al., "Distraction detection and mitigation through driver feedback," National Highway Traffic Safety Administration, Report DOT HS 811 547A, 2013.

[24] K. Kaida, M. Takahashi, T. Åkerstedt, et al., "Validation of the Karolinska sleepiness scale against performance and EEG variables," *Clinical Neurophysiology,* vol. 117, no. 7, pp. 1574-1581, 2006.

[25] S. Elo and H. Kyngäs, "The qualitative content analysis process," *Journal of Advamced Nursing,* vol. 62, no. 1, pp. 107-115, 2008.

[26] M. Ljung Aust, A. Habibovic, E. Tivesten, et al., "Manual for DREAM version 3.2, Driving Reliability and Error Analysis Method," Chalmers University of Technology, Gothenburg, Sweden, 2012.

[27] J. Engström, J. Bärgman, D. Nilsson, et al., "Great expectations: a predictive processing account of automobile driving," *Theoretical Issues in Ergonomics Science,* vol. 19, no. 2, pp. 156-194, 2017.