Msc. Thesis

Impact of control authority on the drivers' perceived responsibility: a haptic shared control driving study

C. H. Kok

TU Delft

Msc. Thesis

Impact of control authority on the drivers' perceived responsibility: a haptic shared control driving study

by

C. H. Kok

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on April 8^{th}, 2021

Student number:	4253485	
Faculty:	Mechanical, Maritime and Materials Engineering (3mE)	
Project duration:	September 1, 2020 – April 8, 2021	
Thesis committee:	Prof. dr. ir. D. A. Abbink,	TU Delft, supervisor
	Dr. ir. N. W. M. Beckers,	TU Delft, supervisor
	Dr. ir. L. C. Siebert,	TU Delft, external member

TU Delft

Preface

In a world of constant progress and innovation, driving automation has been on the leading edge. This innovative process has always been centered on pure technological advancement. This poses no problem as long as the technology is used as a tool. However, new technology is shifting away from technology as a mere tool, to a setting in which machine and man cooperate. This poses a myriad of new problems, many of which are in the social-psychological field.

When assessing the current literature, I found many reviews about the social-psychological views and opinions of the larger public regarding automated vehicles, but much less inquiries into the viewpoint of the driver of the car itself. In this MSc. thesis I aim to broaden the understanding on the interaction and responsibility of driver and car.

C.H. Kok
Groningen, March 2021

Acknowledgement

I am very grateful for the support and guidance of both my supervisors. The enthusiasm of David Abbink has been an inspiration from the start of my MSc programme, and his sharp remarks and his impeccable analytic capabilities have been of great help. Niek Beckers has been an invaluable source of help and support. You have been a guiding light in an hour of need many times, and did so without ever making me feel like I was a burden on your time. Your dedication to my support has been greatly felt and appreciated.

I also would like to thank my parents and family for sticking with me all these years. In this I include my brothers in Delft who always provided me with a space of relaxation outside of my study, and a possibility for escape often very much needed. Lastly I would like to thank Rebecca, for being the steady ground for my feet, for your advice, and help regardless of time and distance. This report would not exist without you.

C.H. Kok
Groningen, March 2021

Contents

1

Journal Paper

Kok C.H., Jonker C.M., Abbink D.A., Beckers N.W.M. (2021). *Impact of control authority on the drivers' perceived responsibility: a haptic shared control driving study.*

Impact of control authority on the drivers' perceived responsibility: a haptic shared control driving study

C.H. Kok, Prof.dr. C.M. Jonker, Prof.dr.ir. D.A. Abbink, and Dr.ir. N.W.M. Beckers

***Abstract*—More and more vehicles have multiple advanced driver-assistance systems (ADAS), that take over tasks from the human driver, thereby taking the driver out of the loop of control. This might create a discrepancy between the responsibility that the human driver feels and the responsibility that is attributed to them when something goes wrong. Previous studies into perceived responsibility were mostly conducted in traded control systems, in which either the vehicle or the driver was performing the task, and tasks were shifted between them. In haptic shared control systems the automation and human driver cooperate continuously. The Level of Haptic Authority (LoHA) determines how strong the controller enforces its guidance. We examine how this LoHA impacts the driver's own perceived outcome responsibility, as well as that attributed to the automation when the automation makes a mistake. We found that when authority is shifted towards the car, the human driver feels less responsible, and attributes more responsibility to the automation, but only to a certain degree. Our findings correspond with previous research and with our own hypothesis. They add a new perspective to the current literature, as this is the first research-paper to examine responsibility perception in haptic shared driving from the drivers perspective. More research in the human driver's experience is needed to better understand human behaviour whilst driving with driving automation systems.**

I. Introduction

Nowadays, we can hardly imagine what it would be like to drive without any form of assistive technology in our car. A continuously increasing amount of on-road vehicles depend on one or multiple advanced driver-assistance systems (ADAS) . These ADAS are designed to assist with driving tasks, including Lane Keep Assist or Adaptive Cruise Control. In turn, the introduction of such assistive systems also has drawbacks. For example, ADAS can lead to driver deskilling, or yield unforeseen situations [1]. Furthermore, overly optimistic introductions to automated driving features by the automotive industry (e.g. Tesla's autopilot) have demonstrated to create expectations about functionality that are greater than the actual capabilities of these systems possibly leading to overreliance on the automation [2]. The scientific community therefore shows a growing concern that this shift towards automated driving might create a gap in responsibility for negative outcomes in which one or multiple ADAS assisted a human driver [3], [4].

One increasingly popular attempt at reducing this gap in responsibility is the concept of Meaningful Human Control (MHC) [3], which includes human moral values back into the design of ADAS. For a system to be under MHC it should satisfy two necessary conditions: a tracking condition and a tracing condition. The tracking condition states that a system, as a whole (i.e., both human and non-human agents together), is not under MHC unless it responds to human norms and values to act, or refrain from acting [5]. An example of this would be the system prioritizing certain objectives (e.g., safely stop the car at the side of the road because the driver is unresponsive) over others (e.g., continue at the set cruise-control speed between the current lane-boundaries) [5]. The tracing condition, on the other hand, states that human agents (e.g., designers of the car, end-users) should understand the system in terms of its capabilities, effects, and the potential moral consequences its use might hold [6]. That includes a thorough understanding of that person's own abilities, level of skill and their role in the system. This second condition might be violated for example when car manufacturers are aware of the technical limits of ADAS they produce, yet shift the responsibility for accidents caused by said limitations to the end-user by having them accept the terms and conditions [5].

Calvert et al. [4] illustrate this in a qualitative sense by analyzing three well-known accidents with semi-automated vehicles that use traded control (TC). They found that these accidents exhibit discrepancies between the amount of responsibility that is attributed to the driver and the abilities of the driver to intervene and avoid the accident. Others [7], [8] have applied a more quantitative approach by getting the public's opinion using online surveys. They found that in crashes involving semi-automated vehicles, human drivers are generally blamed more than the car(-manufacturer) [8], even when both make mistakes and regardless of whether the driver was in a position to actually interfere [7]. These gaps in responsibility are a direct result of violations of the tracing condition. In none of the above examples was the system under MHC.

To our knowledge, these studies on responsibility perceptions/attributions centered on vehicle automation have solely focused on TC systems [8]–[10], which are prone to violating the tracing condition [4]. One type of ADAS that might lead to more meaningful human control is haptic steering guidance or haptic shared control (HSC). The key idea of shared control is that driver and ADAS share the control task; hence they continuously cooperate. For HSC specifically, the continuous cooperation occurs through haptic forces that facilitate mutual communication between driver and ADAS. As a result, the human driver stays actively involved in the driving task and is kept in the control loop [11], [12] preventing out-of-the-loop problems regularly shown to occur in TC driving [13]–[16]. This could potentially prevent violations of the tracing condition.

C.H. Kok, D.A. Abbink, and N.W.M. Beckers are part of the Department of Mechanical Engineering, Faculty of 3mE, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands,

C.M. Jonker is part of the Department of Interactive Intelligence, Faculty of EWI, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands.

Email: C.H.Kok@student.tudelft.nl

When driving with TC, control is not shared: either the ADAS is in control of the driving task, or the driver. Control over tasks is traded dichotomously between driver and automation by means of take-over requests [17]. With HSC, however, the level of guidance can vary continuously depending on the level of haptic authority (LoHA) [11]. The LoHA determines how strong the controller enforces its guidance by imposing a virtual spring around the controller's desired steering wheel angle; the higher the LoHA, the stronger the controller enforces its guidance [18]. It has the potential to allow for a continuous degree of authority from weakly enforced haptic authority (i.e., a low torsional resistance) to a strong enforced haptic authority (i.e., a strong torsional resistance). The driver may go along with the guidance by relaxing their muscles or overwrite the ADAS by steering against it [11]. This way the human driver and the automation jointly steer the car, yet the driver holds the option to overwrite the automation.

However, one downside of HSC is that this continuous scale of support could lead to confusion about who is in charge [19]. In turn, this confusion might lead to mistakes and potentially to accidents. We do not yet know to what degree the human driver actually feels responsible when accidents occur. Therefore, we need more insights in the driver's experience and change of perspective due to the shift towards automated driving, specifically concerning their perceptions of responsibility [20].

Most of the studies on responsibility and driving using ADAS make use of online surveys and vignettes. However, the ecological validity of such online studies is low and can be difficult to use to assess the driver's own perceived responsibility [21]. The body of literature on studies from a driver's perspective in a simulator is surprisingly scarce. Indeed, the bulk of the available studies examine responsibility attribution from an observer's point of view, while it is widely known that observers attribute responsibility differently than actors [22]. As a result, we lack insight into the driver's actual experience or perspective on how they perceive their responsibility when an accident occurs while using ADAS.

Previous studies on perceptions of responsibility show that the humans perception of control over a robots behaviour is an important element in their evaluation of achieved outcomes [10]. It seems that humans assume more responsibility for outcomes if they felt in control, as opposed to when they felt out of control [23]. Perceived behavioural control (PBC) over steering, therefore, appears a promising antecedent for perceived outcome responsibility in automated driving.

In this paper, we will use a driving simulator to examine how the level of haptic authority impacts' the driver's own perceived outcome responsibility and responsibility attributed to the automated vehicle in case the automation makes a mistake. We will use a virtual driving simulator with a haptic steering wheel which allows for HSC. We investigate the effect of LoHA, valence of outcome and whether the ADAS makes a mistake on how the driver feels responsible for the outcome.

We expect that a higher LoHA will shift the authority towards the car. Authority allows for control [17], so a shift in authority towards the car is expected to decrease the drivers perceived behavioural control over steering the vehicle. This in turn, is expected to decrease the driver's perceived outcome responsibility and increase the outcome responsibility attributed to the vehicle. Furthermore, in line with the well-known self-serving bias [24], we expect that the valence of the outcome might influence the attribution of responsibility for said outcomes. That is, positive outcomes are attributed internally (i.e., higher perceived responsibility towards the human driver, and lower responsibility attributed to the car), and negative outcomes are attributed externally. Additionally, we will look both at cases in which the automation makes a mistake, and in which the automation makes no mistake. On the one hand, we expect responsibility to be attributed externally when the automation makes a mistake and the outcome is negative, as opposed to internally when the outcome is positive. In other words, a negative action by the automation is expected to magnify the above mentioned self-serving bias. On the other hand, we expect a positive action by the automation to have the reversed effect; inhibit the self-serving bias.

II. Method

A. Participants

Twenty-four participants (7 female, 17 male) between 19 and 39 years old ($M = 25.3$, $SD = 4.1$) volunteered for a driving simulator experiment. Participants had their driving license for an average of 6.9 years ($SD = 3.5$). Regarding principal mode of transport, the most frequently selected response category was human powered transportation (e.g. walking, cycling) (12 respondents), followed by private automobile (8 respondents), and public transportation (2 respondents). Two respondends reported driving every day in de past 12 months, 2 participants reported to drive every day, 4 drove 4-6 days a week, 4 drove 1-3 days per week, 6 drove once a month to once a week, 4 drove less than once a month, and 4 never. Regarding mileage in the past 12 months, the most frequently selected response category was 1.001-5.000 km (7 respondents), followed by 1.001-1.000 km (6 respondents), and 0 km (3 respondents). No inquiry was made about the participants' familiarity with ADAS.

B. Apparatus

The experiment was conducted in a fixed-base simulator at the Cognitive Robotics Laboratory at the faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology. This simulator uses the open-source simulation platform called CARLA [25]. The driving simulator set-up contained one TV-Screen (4K Samsung 65") located on a desk, a driving seat (i.e. an office chair without rotational freedom) for the participant, and a SensoDrive SENSO-Wheel (see Figure 1). The steering wheel actuation was done at 500 Hz; steering wheel damping: .4 [Nms/rad], torsional spring stiffness: 1.0 [Nm/rad], with a torque resolution of .03 Nm. Cruise control was set to 80 km/h.

C. Level of Haptic Authority

Apart from the Manual (M) condition, in which only natural self-alignment torques and natural torsional resistances were simulated, the Four Design Choice Architecture (FDCA,

Fig. 1. A participant seated in the fixed-base driving simulator, holding the steering wheel in a 'ten-to-two' position. The final on-road obstruction is visible in the distance.

[18], [26]) was used to provide superimposed haptic guidance torques on the steering wheel. The FDCA controller is designed to follow a Human Compatible Reference (HCR) trajectory. The Level of Haptic Authority (LoHA) determines the amount of effort it takes to counteract the controller. That is, it determines how much torque τ_{LoHA} the steering wheel should generate when the driver's current steering wheel angle θ_{sw} deviates from the optimal angle θ_{op}, as in Eq. 1. The amount of torque per deviation is determined by a gain K_{LoHA}. A more extensive description of the FDCA controller can be found in the Appendix B.

$$\tau_{LoHA} = K_{LoHA}(\theta_{op} - \theta_{sw}) \quad (1)$$

In this paper we look at two different settings for K_{LoHA}: low assistance (L condition; $K_{LoHA,L} = 4$ Nm/rad), and high assistance (H condition; $K_{LoHA,H} = 12$ Nm/rad), in addition to the aforementioned M condition ($K_{LoHA,M} = 0$ Nm/rad). All FDCA parameter settings used in this experiment were determined heuristically and can be found in Table I.

D. Road environment

All participants drove each trial on the same two-lane road (7 m wide and 5.4 km long), using a cruise-control set at 80 km/h. The trajectory used in the experiment contained a variety of curves with radii ranging from 112 m to 266 m with respect to the center-line of the road (aimed at increasing the need for haptic assistance, see Appendix A). The only exception was the first curve with a radius $R = 500$ m. To provoke an active

TABLE I
FDCA PARAMETER SETTINGS USED IN THE EXPERIMENT FOR THE LOHS, SOHF, AND LOHA. THE THREE VALUES FOR K_{LoHA}, IN ASCENDING ORDER, CORRESPOND TO THE M, L_i, AND THE H_i CONDITION RESPECTIVELY.

λ_{LoHS}	1.0 [-]	λ_{SoHF}	1.0 [-]
K_{y-SoHF}	.1 [Nm/m]	K_{LoHA}	0 - 4 - 12 [Nm/rad]
$K_{\psi-SoHF}$	1.5 [Nm/rad]		

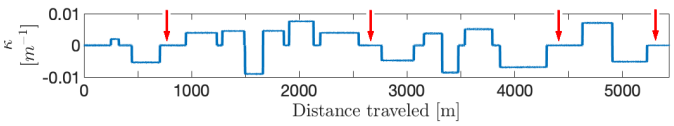

Fig. 2. Curvature ($1/R$) of the trajectory. Red arrows indicate on-road obstructions.

TABLE II
OVERVIEW OF THE TRIAL CONDITIONS USED IN THIS PAPER. EACH PARTICIPANT DROVE FIVE TRIALS.

	LoHA		
Robot action	Manual	Low	High
None	X		
Bad		X	X
Good		X	X

interaction between the human- and the haptic assistance we added four on-road obstructions, as seen in Fig. 2. The first on-road obstruction forced participants to take an evasive action on to the emergency lane. The second and fourth obstruction forced a right-to-left lane switch, and the third obstruction forced two lane switches (right-to-left lane switch followed by a left-to-right lane switch).

E. Manipulation of robot action

Two human compatible reference (HCR) trajectories were created for this experiment, the FDCA controller used one of these depending on which trial was driven. Both of these were designed to evade the first three obstructions and were identical for most of the trajectory. However, HCR_g was designed to evade the final obstruction and counted as a positive action of the robot (in the L_i and H_i conditions this is also denoted by subscript g for good). On the other hand, the HCR_b was designed to navigate right through the obstruction and counted as a negative action of the robot (denoted by subscript b).

As a results, the trajectory was divided in a first part, up to and including the final curve ($\text{HCR}_{overall}$; 5.25 km), and a second part which contained only the final on-road obstruction (HCR_{event}; .15 km). The $\text{HCR}_{overall}$ was identical for all conditions, as opposed to the HCR_{event} which differed between the 'positive action' (HCR_g) and 'negative action' (HCR_b) conditions.

F. Experimental design

We used an unbalanced within-participants design with the robot's action as sub-condition nested under the low- and high LoHA conditions (Table II). Each participant drove five trials; first one trial without haptic guidance (i.e., M condition) to familiarize themselves with the virtual environment, then two trials for each LoHA setting ('good-action' vs. 'bad-action'). The four haptic guidance conditions were counterbalanced.

G. Experiment procedure

Participants read an experiment briefing prior to the experiment explaining the purpose, instructions, and procedures of

the study. Participants were informed that we were interested in how perceived responsibility is impacted by the amount of control the driver experiences over an automated car, but not the underlying hypotheses. Additionally, they were told to keep both hands on the steering wheel in a ten-to-two position with their thumbs on the steering wheel for safety reasons. Participants were also notified that they could feel torques on the steering wheel which may help or hinder them, but that they would always be able to override these torques. Participants were instructed to drive as they normally would except during the first straight section of road; here they were advised to probe the guidance system to get a feel for the system they were driving with.

After reading the experiment briefing the participants read and signed an informed consent form. Before starting the first trial, participants completed a questionnaire regarding their sex and driving experience, adapted from the Driver Behaviour Questionnaire [27]. After each trial, participants were requested to fill out a questionnaire to assess their perceived behavioral control and responsibility for two outcomes; a positive and a negative outcome. After filling in the questionnaire, the participants were asked if they needed a break. The total experiment, including filling out the questionnaires, took approximately 45 min. per participant.

H. Manipulation of outcome

For the positive outcome scenario, participants were told to assume they evaded all of the roadblocks (two trials resulted in a collision, see Fig. 9 in Appendix G). For the negative outcome scenario, participants were told to imagine that they did crash into one of the roadblocks.

I. Dependent measures

Perceived outcome responsibility, was measured with a 1-item scale per actor (i.e., Human vs. Car). For example, "To what extent do you (the driver) feel responsible for this outcome?" was rated on a 7-point Likert scale from 1 = *Not at all* to 7 = *Completely*. The proposed mediator variable, perceived behavioural control, was measured with a 1-item scale "I feel in control steering this car" rated on a 7-point Likert scale from 1 = *Strongly disagree* to 7 = *Strongly agree*. We initially intended to use a 2-item scale for perceived control which included the statement "The car lets me (the driver) have control over steering". However, we realized (post-hoc) that this statement better represents perceived authority and is, therefore, used as a subjective check for manipulation of the LoHA.

To increase the signal-to-noise ratio, we added an attention check statement [28]; "I am paying attention" rated on a reversed 7-point Likert scale from 1 = *Completely* to 7 = *Not at all*. Trial runs in which this attention check was answered with a 4 or higher were excluded from the analyses ($n_{ex} = 8$). The post-trial questionnaire that was used in this paper can be found in Appendix E.

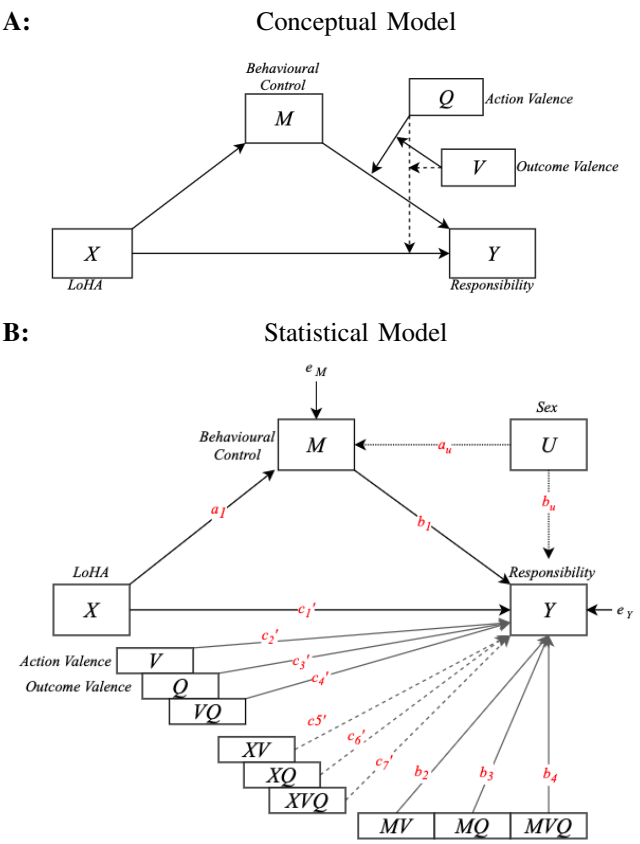

Fig. 3. Representation of the relations between level of haptic authority (LoHA), perceived behavioural control and attributed outcome responsibility. Panel A: Conceptual model, these relations are believed to be moderated by the action of the robot, and the valence of the outcome. Perpendicular arrows indicate moderation. Panel B: Corresponding statistical model, sex (U) is added as a covariate.

J. Statistical Model

The conceptual model in Figure 3, panel A is a second stage and direct effect moderation model [29], [30] with two multiplicative moderators [31]. This model describes the effect that level of haptic authority has on the responsibility attributed to either the car or the human driver; the effects for each actor is analysed individually. Statistically this can be represented in the form of a path diagram as shown in Figure 3, panel B, and in the form of two regression equation as in Eq. 2 and 3.

$$M_{driver} = i_M + a_1 X + a_u U + e_M \tag{2}$$

$$\begin{aligned} Y_{actor} = {} & i_Y + c_2' V + c_3' Q + c_4' VQ + b_u U + e_Y \\ & + (c_1' + c_5' V + c_6' Q + c_7' VQ) X \\ & + (b_1 + b_2 V + b_3 Q + b_4 VQ) M_{driver} \end{aligned} \tag{3}$$

In these regression equations, U is the sex of the participants ($U_{female} = -.5$, $U_{male} = .5$) and used as a covariate (not included in the conceptual diagram). The effect that LoHA (X) has on the outcome responsibility (Y) is believed to be mediated by the perceived behavioural control of the driver (M_{driver}) . The above model allows the effect that M_{driver} has on Y to be moderated by the robot's action V and the valence of the outcome Q. The corresponding contrast coding can be found in Table IV. This model includes multiplicative moderation [31] between both moderators and fixed effect X;

TABLE III
MEANS (M) AND STANDARD DEVIATIONS (SD) OF PERCEIVED AUTHORITY, PERCEIVED BEHAVIORAL CONTROL (1 = *Strongly disagree* - 7 = *Strongly agree*), AND OUTCOME RESPONSIBILITY (1 = *Not at all* - 7 = *Completely*) FOR EACH OF THE CONDITIONS; GOOD/BAD REFERS TO THE ACTION OF THE ROBOT, POSITIVE/NEGATIVE REFERS TO THE VALENCE OF THE OUTCOME.

	Authority	Behavioural Control	Outcome Responsibility			
Conditions			Positive	Negative	Positive	Negative
	M (SD)	M (SD)	M_{Car} (SD)		M_{Human} (SD)	
Manual level	5.38 (2.14)	5.46 (1.67)	2.46 (1.86)	2.17 (1.66)	6.71 (.69)	6.62 (.65)
Low LoHA						
Good	4.29 (1.71)	4.75 (1.73)	4.17 (1.81)	4.42 (1.67)	5.08 (1.59)	4.96 (1.27)
Bad	4.12 (1.70)	4.58 (1.77)	4.29 (1.71)	4.12 (1.96)	4.96 (1.78)	4.71 (1.76)
High LoHA						
Good	3.96 (1.85)	4.33 (1.97)	4.71 (1.81)	4.71 (1.63)	4.71 (1.73)	5.00 (1.47)
Bad	4.12 (1.51)	4.83 (1.55)	4.08 (1.69)	4.88 (1.68)	4.75 (2.15)	4.88 (1.54)

TABLE IV
CONTRAST CODING OF FIXED EFFECT X (I.E., LOHA) AND MODERATORS V (I.E., VALENCE OF ROBOT ACTION) AND Q (I.E., VALENCE OF OUTCOME).

LoHA [Nm/rad]		Moderators		
Manual → Low	**Low → High**	**Action** (V)	**Outcome** (Q)	**Encoding**
0 (M)	4 (L)	bad	negative	-.5
4 (L)	12 (H)	good	positive	.5

i.e. three-way interaction. From Eq. 3 we get the moderated indirect effect ω of **X on Y through M** (Eq. 4), and the moderated direct effect γ of **X on Y** (Eq. 5).

$$\omega = a_1(b_1 + b_2V + b_3Q + b_4VQ) \tag{4}$$
$$\gamma = c_1' + c_5'V + c_6'Q + c_7'VQ \tag{5}$$

We will call a_1b_2, a_1b_3, and a_1b_4 the indices of moderated mediation with respect to V, Q, and VQ respectively. These indices represent a quantification of the effects that V, Q, and the combined effect of VQ have on ω in our model [31], [32].

Regression coefficients were estimated using Linear Mixed-Effect (LME) modelling [33]. This procedure is described in more detail in the Appendix D. We use the bias-corrected and accelerated (BCA) bootstrap method to calculate 95% confidence intervals (CI) for our coefficients and corresponding conditional effects [34]. As a proxy of effectiveness of the LME model we calculate the conditional coefficient of determination R_c^2 for mixed-effect models, as defined in Eq. 6 [35]:

$$R_c^2 = \frac{\sigma_{f^2} + \sigma_{r^2}}{\sigma_{f^2} + \sigma_{r^2} + \sigma_{\epsilon^2}} \tag{6}$$

where σ_{f^2} is the variance of the fixed effect components, σ_{r^2} is the variance of the random effects, and σ_{ϵ^2} is the observation-level variance.

On the one hand, for the objective measure (i.e. μ_τ), we report the results of right-tailed Wilcoxon signed rank tests for $\alpha = .05$. On the other hand, for the moderated mediation analysis we report the regression coefficients, corresponding indices of moderated mediation, and their corresponding 95% bias-corrected CI based on 10,000-bootstrap iterations.

III. RESULTS

Table III provides the estimated coefficients (mean values and SD) corresponding to perceived authority, perceived behavioral control over steering, and outcome responsibility for each experimental condition. First, we will check if the manipulation of the LoHA had the desired result. Then, we will examine if and how a change in the LoHA affects the participants perceptions of responsibility for outcomes. Finally, we will see how the valence of the outcome influenced attributions of outcome responsibility.

A. Manipulation Check

The first 1.25 km of each trial run, containing 3 curves and one road obstruction, were considered as adjustment periods for the driver to adjust to the new system. The corresponding measured data was discarded accordingly.

We compared the steering wheel torque $\tau_{LoHA,overall}$ corresponding to $\text{HCR}_{overall}$ of the M, L, and H conditions to asses if manipulating the LoHA had the desired effect. Data of both the L_i and H_i conditions were pooled across their nested conditions (i.e., good and bad action of the robot). As intended, the $\tau_{LoHA,overall}$ for the L condition were greater than those for the M condition ($M_{LoHA,L} = .25$ Nm vs. $M_{LoHA,M} = 0$ Nm; $W = 253$, $p < .001$). This increase in LoHA was accompanied by a decrease in perceived authority ($M_{authority,L} = 4.21$ vs. $M_{authority,M} = 5.38$, $W = 742$, $p = .0015$). A further increase in LoHA did results in larger torques on the steering wheel ($M_{LoHA,H} = .55$ Nm; $W = 252$, $p < .001$), however, this did not lead to a significant difference in perceived authority ($M_{authority,H} = 4.04$; $W = 461$, $p = .351$). Increasing the LoHA had the intended mechanical effect on the steering wheel. However, it only had the desired effect on authority for the Low condition; a further increase did not significantly change the perceived authority.

We compared the steering wheel torque $\tau_{LoHA,event}$ corresponding to HCR_{event} of the good versus bad conditions to assess manipulation of the robot action. Data of the good action conditions (L_g, H_g) were pooled, as was the data of the bad action conditions (L_b, H_b). As intended, the torques during navigation of the final obstacle were greater for the 'bad action' ($M_{LoHA,b} = 1.77$) than those for 'good action'

TABLE V
LME REGRESSION COEFFICIENTS WITH BCA CONFIDENCE INTERVALS ESTIMATING BEHAVIOURAL CONTROL AND OUTCOME RESPONSIBILITY. LOHA AND BEHAVIOURAL CONTROL ARE CONTRAST CODED AND MEAN CENTERED.

		Behavioural Control (M)			Outcome Responsibility (Y)			
		Manual → Low	Low → High		Manual → Low		Low → High	
					Car	Human	Car	Human
		Coeff. 95% CI	Coeff. 95% CI		Coeff. 95% CI	Coeff. 95% CI	Coeff. 95% CI	Coeff. 95% CI
LoHA (X)	$a_1 \rightarrow$	**-.79 [-1.08, -.56]***	-.08 [-.28, .18]	$c'_1* \rightarrow$	**1.44 [.90, 1.97]***	**-1.41 [-1.75, -1.14]***	.32 [-.09, .69]	-.08 [-.44, .32]
Behavioural Control (M)				$b_1* \rightarrow$	**-.32 [-.50, -.06]***	**.31 [.17, .46]***	**-.41 [-.56, -.22]**	**.49 [.37, .68]**
Action (V)				$c'_2 \rightarrow$			.09 [-.25, .42]	.19 [-.09, .52]
Outcome (Q)				$c'_3 \rightarrow$	.07 [-.29, .45]	.15 [-.13, .46]	-.22 [-.56, .10]	.02 [-.33, .29]
V × Q				$c'_4 \rightarrow$			.20 [-.51, .86]	-.12 [-.78, .52]
X × V				$c'_5 \rightarrow$			-.11 [-.75, .56]	.15 [-.45, .83]
X × Q				$c'_6 \rightarrow$	-.31 [-1.33, .67]	.28 [-.25, .90]	-.37 [-1.09, .31]	-.41 [-1.07, .25]
X × V × Q				$c'_7 \rightarrow$			1.23 [-.09, 2.67]	.07 [-1.23, 1.34]
M × V				$b_2 \rightarrow$			.01 [-.20, .21]	-.10 [-.31, .13]
M × Q				$b_3 \rightarrow$	.06 [-.15, .24]	.21 [-.01, .40]	.05 [-.15, .26]	.18 [-.03, .38]
M × V × Q				$b_4 \rightarrow$			-.10 [-.56, .30]	-.14 [-.55, .27]
Sex (U)	$u_a \rightarrow$	**1.78 [1.53, 2.04]***	**.93 [.71, 1.15]***	$u_b \rightarrow$	**.64 [.09, 1.10]***	**-.13 [-.49, -.19]***	**.59 [.18, .95]***	-.08 [-.47, .32]
Intercept	$i_M \rightarrow$	**-.37 [-.50, -.24]***	**-.19 [-.30, -.08]***	$i_Y \rightarrow$	**3.44 [3.17, 3.66]***	**5.55 [5.42, 5.76]***	**4.31 [4.11, 4.49]***	**4.90 [4.71, 5.09]***
		$R^2_c = .894$	$R^2_c = .846$		$R^2_c = .731$	$R^2_c = .703$	$R^2_c = .708$	$R^2_c = .739$

* $0 \notin 95\%$ CI

condition ($M_{LoHA,g} = 0.33$ Nm; $W = 253$, $p < .001$). In other words, participants had to apply more torque on the steering wheel, when the robot did not steer to evade the final on-road obstruction (i.e., robot mistake), to counteract the guidance torques. Thus, the manipulation of the robot action proved to be successful.

B. Main findings

The proposed mediation model did not converge when comparing the Manual conditions to the Low LoHA conditions. We found no significant difference between observations of the Low and Good action condition when compared to those of the Low and Bad action condition for none of the subjective variables. Thus, we pooled observations of the good and bad robot action trials of the low LoHA condition when comparing these observations to those of the manual condition, therefore, effectively not considering the valence of the robot action as a moderator in this analysis. This simplified model did converge. When comparing the Low LoHA observations to the high LoHA observations we did not exclude the robot action as a potential moderator. The corresponding estimated regression coefficients of both analyses can be found in Table V.

Perceived behavioural control as mediator

Supporting the first hypothesis, at least to a a certain degree, we see that a moderate increase in LoHA (i.e. manual → low) is matched by a decrease in behavioural control over steering ($a_1|_{M\rightarrow L} = -.79$, 95% CI $= [-1.08, -.56]$). However, we found that further increasing the LoHA did not necessarily lead to a change in perceived control over steering ($a_1|_{L\rightarrow H} = .08$, 95% CI $= [-.28, .18]$). Providing support for our second hypothesis, we found for both increases of LoHA a positive correlation between the perceived behavioural control of the driver over the steering of the car and the attributed responsibility to the car ($b_{1,car}|_{M\rightarrow L} = -.32$, 95% CI $= [-.50, -.06]$, $b_{1,car}|_{L\rightarrow H} = -.41$, 95% CI $= [-.56, -.22]$). As expected, we found a negative correlation for the perceived responsibility of the driver ($b_{1,human}|_{M\rightarrow L} = .31$, 95% CI $= [.17, .46]$, $b_{1,human}|_{L\rightarrow H} = .49$, 95% CI $= [.37, .68]$). As a result, perceived control significantly mediated the effect that a moderate increase of haptic authority had on attributions of responsibility; both for attributions towards the car ($a_1 b_{1,car}|_{M\rightarrow L} = .48$, 95% CI $= [.21, .80]$), and for attributions towards the human driver ($a_1 b_{1,human}|_{M\rightarrow L} = -.33$, 95% CI $= [-.50, -.15]$). No significant mediation, and therefore no moderated mediation, was found for a further increase of LoHA. All indices of (moderated) mediation can be found in Appendix G.

Outcome valence as moderator

Surprisingly, the degree to which behavioural control affected outcome responsibility was not explicitly influenced by whether the outcome was positive or negative ($b_{3,car}|_{M\rightarrow L} = .06$, 95% CI $= [-.15, .24]$, $b_{3,human}|_{M\rightarrow L} = .21$, 95% CI $= [-.01, .40]$). Nonetheless, at least for responsibility attributed to the driver, we found a significant moderating effect of outcome valence on the indirect relation between an increase in the level of haptic authority and attributed outcome responsibility ($a_1 b_{3,human}|_{M\rightarrow L} = -.17$, 95% CI $=$

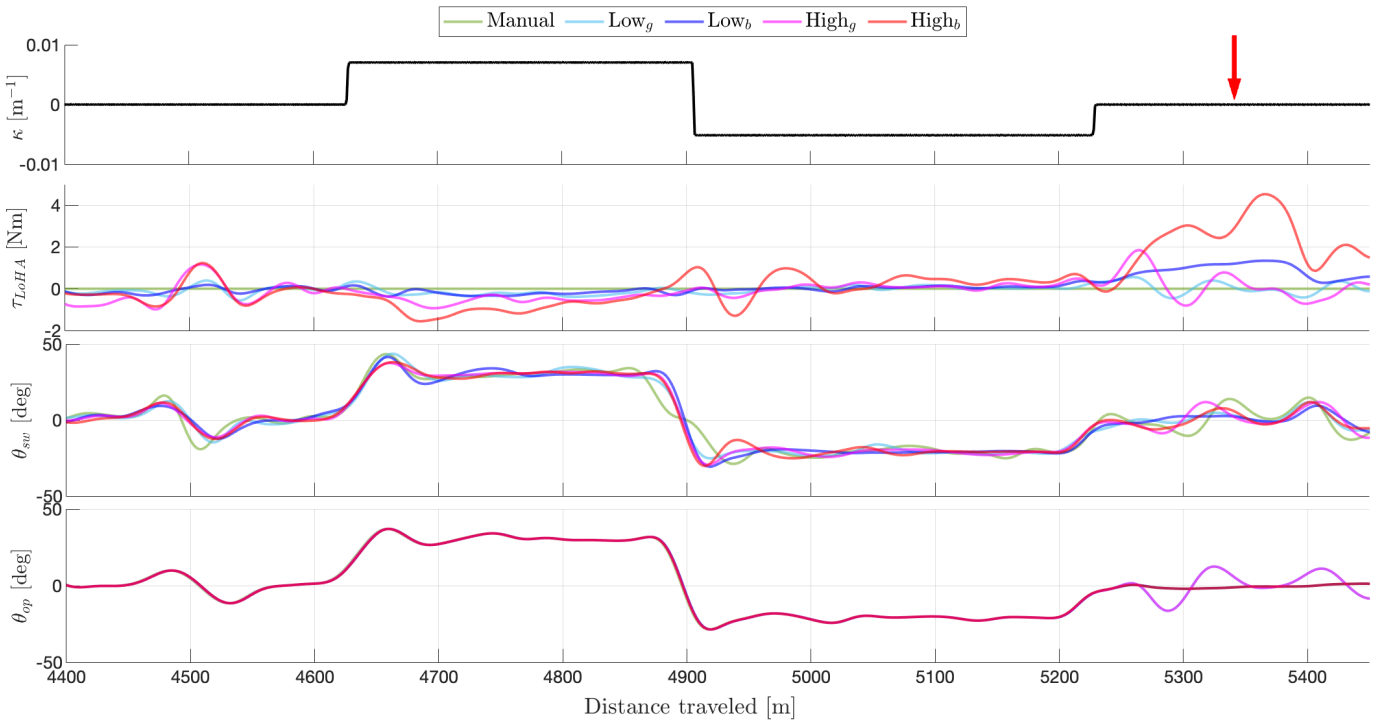

Fig. 4. Raw data for a random participant during the last kilometre of the track including the final two curves and the final on-road obstruction (red arrow). From top to bottom: Road curvature κ, LoHA torque τ_{LoHA}, steering wheel angle θ_{sw}, optimal steering angle θ_{op}.

$[-.37, \ -.01]$).

LoHA via behavioural control is positively correlated to responsibility attributed to the car regardless of the outcome; an increase in LoHA *increases* the responsibility attributed to the car. For the human driver, in case of a positive outcome, this same increase in LoHA leads to a *decrease* of approximately the same magnitude in self-attributed responsibility. However, the driver's perceived responsibility decreases less strongly for negative outcomes.

Direct (non mediated) effect

In contrast, at least for an increase in LoHA from Manual to Low, an increase in authority directly increases responsibility attributed to the car ($c'_{1,car}|_{M\rightarrow L} = 1.44$, 95% CI $= [.90, 1.97]$), and decreases responsibility attributed to the human driver ($c'_{1,human}|_{M\rightarrow L} = . - 1.41$, 95% CI $= [-1.75, -1.14]$) regardless of the outcome valence Q. This direct effect γ is considerably stronger than the indirect effect ω. No significant result was found for the individual coefficients of γ for a further increase of the LoHA. However, when probing the model we did find a significant direct effect for responsibility towards the car in the bad robot action combined with a negative result scenario ($\gamma_{car,bad,neg}|_{L\rightarrow H} = .86$, 95% CI $= [.15, 1.56]$). A complete overview of all probed effects can be found in Table 2 and 3 in Appendix G.

Figure 5 shows the responsibility (Y) as a function of LoHA (X), and the total effect. The total effect ε is defined as the sum of ω_i and γ_i averaged over all participants (see Appendix B). We found significant total

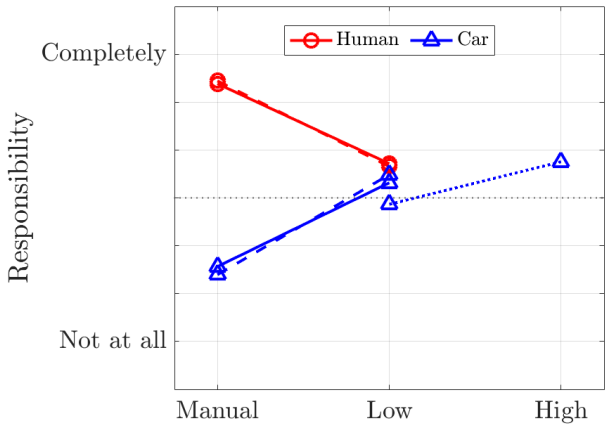

Fig. 5. Responsibility (Y) as a function of LoHA (X), defined by $Y = \varepsilon \ X + i_Y$, for the human (red circle) and the car (blue triangle) for both the positive (line) and the negative (dash) outcome. The dotted line corresponds to the 'bad action-negative outcome' scenario of the $L \rightarrow H$ regression model. Only significant total effects ε are included in the figure.

effects for every outcome of the probed Manual $\rightarrow$ Low LME model. For responsibility towards the driver, we found significant total negative effect in case of a negative outcome ($\varepsilon_{human,neg}|_{M\rightarrow L} = -1.79$, 95% CI $= [-2.17, -1.44]$), and in case of a positive outcome ($\varepsilon_{human,pos}|_{M\rightarrow L} = -1.67$, 95% CI $= [-2.06, -1.28]$). For responsibility towards the car, we found significant total positive effect in case of a negative outcome ($\varepsilon_{car,neg}|_{M\rightarrow L} = 2.09$, 95% CI $= [1.46, 2.70]$), and in case of a positive outcome ($\varepsilon_{car,pos}|_{M\rightarrow L} = 1.75$, 95% CI $= [1.07, 2.43]$). For the probed Low $\rightarrow$ High

LME model we only found a significant result in the bad robot action combined with a negative result scenario ($\varepsilon_{car,bad,neg}|_{L\rightarrow H} = .89$, 95% CI $= [.23, 1.59]$).

On a general note, the results additionally show that sex, as a co-variate, is positively correlated with behavioural control ($u_a|_{M\rightarrow L} = 1.78$, 95% CI $= [1.53, 2.04]$, $u_a|_{L\rightarrow H} = .93$, 95% CI $= [.71, 1.15]$). This implies that women ($U_{female} = -.5$) generally perceive less control over steering than men. A similar effect of sex is observed regarding attributions of outcome responsibility towards the car ($u_{b,car}|_{M\rightarrow L} = .64$, 95% CI $= [.09, 1.10]$, $u_{b,car}|_{L\rightarrow H} = .59$, 95% CI $= [.18, .95]$). The opposite, however, is observed for attributions of outcome responsibility towards the driver ($u_{b,human}|_{M\rightarrow L} = -.13$, 95% CI $= [-.46, -.19]$); women seem to perceive more outcome responsibility than men.

IV. Discussion

This study aimed to examine how the level of haptic authority impacts the driver's own perceived outcome responsibility, as well as the responsibility attributed to the automated vehicle in case the automation makes a mistake. Of particular interest was the potential mediating role that perceived behavioural control might hold in this relation. In addition, the study examined whether the valence of the robot action and of the outcome altered the level of responsibility attribution towards either actor.

Implications

The results support our hypotheses to a certain degree. That is, participants felt less in control over steering while cooperating together with the haptic shared controller, when compared to driving by themselves. As expected, we found a positive correlation between perceived control and responsibility; participants who felt less in control also perceived less responsibility for the achieved outcome. As a result, participants felt less responsible for outcomes achieved in cooperation. This indeed demonstrates that self-attribution of responsibility for achieved outcomes is mediated, to certain extent, by the perceived control that the driver experiences over steering. This finding corroborates the findings of Jörling et al. [10], yet to our knowledge, this study is the first to use a driving simulator to examine perceptions from the driver's perspective. The portion of this attribution explained by perceived behavioural control, however, is small compared to the total effect of the LoHA. Increasing the haptic shared controller's authority decreases the driver's perceived responsibility, regardless of the perceived control over steering.

The opposite appears for responsibility that is attributed to the car(-manufacturer). Our expectation to see an increase in the attributed responsibility to the car as a result of a shift in authority towards the car was confirmed. Surprisingly, we found that the responsibility attributed to the car increased with approximately the same rate as that the driver's own perceived responsibility decreased. This holds for both the indirect effect, via perceived behavioural control, and the direct effect of LoHA on perceived outcome responsibility. This implies that participants' perceptions follow a rule of conservation of responsibility when attributing responsibility to each actor; if they are less responsible, then the car should be more responsible. This observation is conceptually similar to one made in a social psychology study that examined how an observer's attribution of responsibility to multiple actors is influenced by the moral character and the motive of the actors [36]. However, this is observed for the first time in the context of a human cooperating with a driving automation system.

Unfortunately, conservation could work both ways. If the automation is believed by the driver to be more capable than it actually is, then what does that signify for the drivers perception of responsibility if it does indeed follow a rule of conservation [2]? False expectations may lead to the driver not taking full responsibility of his role in driving and overreliance on the automation [37]. As a result, responsibility attributed by the driver to the car(-manufacturer) might be inflated, thereby decreasing the driver's own feelings of outcome responsibility. This potential shift in responsibility from the driver's perspective would, if not accounted for by the car's design, directly violate the tracing condition for meaningful human control. This could lead to serious safety issues like those shown to occur in TC driving [4].

Haptic shared control however, is believed to decrease overreliance on the automation as the driver stays actively involved in the driving task and is kept in the control loop [11], [12]. This might explain why the driver's perceived and attributed responsibility did not change significantly for an increase in LoHA past what we defined as 'low assistance' or the low LoHA condition. Our low LoHA condition was set at $K_{LoHA} = 4$ Nm/rad, which is approximately 2 to 3 times higher than what Zwaan et al. heuristically determined as the upper limit of the LoHA before drivers would report dissatisfaction [18]. In hindsight our definition of 'low assistance' might have already been on the strong side of the LoHA spectrum. HSC is designed so that the driver holds the option to overwrite the automation, thereby taking back the authority [11]. Increasing the LoHA had the intended mechanical effect on the steering wheel. However, the results of the subjective manipulation check suggest that the LoHA had a ceiling effect on the authority shift towards the car (see Appendix G). Our high LoHA condition might, therefore, have been so strong that participants felt the need to hold the authority (by not giving way to, or steering against the automation), and consequently not 'hand over' the responsibility for driving to the car [17]. Below, in the additional analysis section we will offer an alternative possible explanation.

We found a reversed self-serving bias for the portion of self-attributed responsibility that is accounted for by perceived control, much like Jörling et al. found in the use of semi-autonomous service robots [10]. That is, we observed a slightly stronger decrease in self-attributed responsibility to the driver for positive outcomes when compared to negative outcomes. This implies that drivers that use HSC steering feel more responsible when the outcome is negative, as opposed to when that outcome is positive. Jörling et al. identify that negative outcomes obtained by service robots are attributed internally only if the service customer perceives

ownership of the service robot. They hypothesise that "long-term satisfaction with service robots might be higher when customers do not own them". Keeping the driver in-the-loop might not lead to actual 'ownership' of the car, but rather keeps ownership over the driving or steering behaviour of the car with the driver. We believe that this could contribute to a more responsible driving behaviour of the system as a whole (i.e. human and non-human actors together). On the other hand, it might be that the driver felt less responsible for the positive outcome because the car actively steered away from the obstruction and is therefor to be credited for the positive outcome.

The valence of the robot's action had no significant effect. This is not surprising seeing as the robot's action was only included as a potential moderator when we analysed an increase of the LoHA past the 'low assistance' condition. Here we found no significant indirect effect of LoHA on responsibility, let alone moderated by valence of the robot's action. We did find a significant direct effect of responsibility towards the car in the Bad action combined with a negative result scenario. However, the fact that we found a significant effect in this specific case and not on the other scenarios is no definitive evidence of moderation. Or as Hayes [38] put it: "difference in significance does not imply significantly different". Other factors that were not accounted for might have led to this result. It does, however, hold the door open for the valence of the robot action as a potential moderating factor for responsibility attribution in shared control.

Additional analysis

A closer look at the torques involved suggests an opposite explanation to the 'too high LoHA' argument given above. We looked at the mean absolute guidance torques during the main part of the track, (see Figure 1 and 2 in Appendix G), and saw an increase in guidance torques for the high LoHA conditions when compared to the low LoHA conditions. Why then do we not see an effect on perceived control or responsibility for that increase in LoHA? Perhaps the human force perception is not sensitive enough to feel these differences in the guidance torques, due to limitations in the arm's proprioception or due to limitations in their sensory perception. Following the same reasoning, we looked at the mean absolute guidance torques during navigation around the final obstruction. As intended, we saw an increase in mean absolute guidance torque when the automation 'overlooks' the on-road obstruction compared to when the automation guides the car around the obstruction. Surely, this should have led to some effect of the automation's action?

In an attempt to answer this we looked at the driver's mean absolute steering torques and noticed that these barely differed between conditions during the entirety of the track including the final obstruction (see Figure 3, 4, 7, and 8 in Appendix G). It seems that participants did not (have to) exert more physical effort to steer against the automation, in contrast to our expectations. Perhaps the levels of haptic authority weren't sufficiently far-removed to adequately represent a low haptic authority scenario versus a high haptic authority scenario. A thorough statistical analysis is needed to corroborate these hypotheses. All of the above interpretations, however, imply that the high LoHA condition might not have been high enough. This contradicts the idea that the High LoHA condition might have been too high.

Additionally, we looked at the raw car positions and corresponding safety margins like the Time to Line Cross (TLC) [s] and the Time to Collision (TC) [s] (see Fig. $9-11$ in Appendix G). These however did not offer new insights.

Limitations and future research

We investigated the effect of the level of haptic authority on responsibility attributions for outcomes obtained in cooperation between human driver and a haptic guidance system. When looking at our research we concluded that an increase in level of haptic authority was related to a decrease in perceived human responsibility. We also concluded that this responsibility stopped decreasing at some point. Future research could focus on finding this tipping point. We could also speculate that the curve might start in a similar way. This would mean that there could be an increase of LoHA for which the perceived responsibility of the driver does not decrease. A next step would be to investigate if and where the tipping point exists at which perceived responsibility towards the driver starts to decrease for increasing level of haptic authority.

We also found a strong effect of sex, when modeled as a covariate, on both perceived behavioural control and on responsibility attribution. Especially its effect on perceived behavioural control is large compared to the effect that LoHA has. This suggests that sex might play a more central role in how behavioural control and responsibility are perceived. In this paper we modeled sex as a covariate. Future research might be conducted in which sex is better accounted for both theoretically as in terms of participants.

Almost all our participants were from the Delft University of Technology, following a Mechanical Engineering master course and often specialising in robotics or human-machine interfaces. This might confound findings because of the participant's familiarity with automation in general, and with HSC in particular. An interesting conception from the field of political science is that responsibility attribution is often influenced by the level of knowledge an individual has with regards to a subject [39]. That is, people who understand an automation might attribute responsibility in systematically different ways than a person who does not. Future research might broaden the participant pool to better reflect the general population. Similarly, almost all our participants were relative young; i.e. younger than thirty years of age. As a drivers licence in the Netherlands cannot be obtained before 17 years of age, driving experience was by default limited. Subsequent research might examine the influence of driving experience on perceived responsibility.

V. Conclusion

This research is one of the first investigations that make use of a virtual simulator to examine responsibility attribution

in shared control driving from the driver's perspective. We believe that developers should take factors that could influence the driver's perceived responsibility into account to prevent gaps in responsibility when designing for cooperative driving automation systems. We conclude that a shift in authority towards the car is accompanied by a shift in outcome responsibility from the driver towards the car. This shift in responsibility is only partially explained by the decrease in the driver's perceived control over the steering behaviour of the car. Our findings should be seen as motivation for further research into responsibility attribution in shared control driving.

REFERENCES

[1] D. D. Heikoop, M. Hagenzieker, G. Mecacci, S. Calvert, F. Santoni De Sio, and B. van Arem, "Human behaviour with automated driving systems: a quantitative framework for meaningful human control," *Theoretical Issues in Ergonomics Science*, vol. 20, no. 6, pp. 711–730, 11 2019. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1463922X.2019.1574931

[2] M. A. Nees, "Drivers' Perceptions of Functionality Implied by Terms Used to Describe Automation in Vehicles," *Proceedings of the Human Factors and Ergonomics Society*, 2018.

[3] F. Santoni De Sio, "Ethics and Self-driving Cars: A White Paper on Responsible Innovation in Automated Driving Systems," Tech. Rep., 2016.

[4] S. Calvert, G. Mecacci, B. Van Arem, F. Santoni De Sio, S. C. Calvert, B. Van Arem, D. D. Heikoop, and M. Hagenzieker, "Gaps in the Control of Automated Vehicles on Roads," *IEEE Intelligent Transportation Systems Magazine*, 2019. [Online]. Available: https://www.researchgate.net/publication/333566108

[5] G. Mecacci and F. Santoni de Sio, "Meaningful human control as reason-responsiveness: the case of dual-mode vehicles," *Ethics and Information Technology*, pp. 1–13, 12 2019.

[6] F. Santoni de Sio and J. van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account," *Frontiers in Robotics and AI*, vol. 5, no. FEB, p. 15, 2 2018. [Online]. Available: http://journal.frontiersin.org/article/10.3389/frobt.2018.00015/full

[7] E. Awad, S. Levine, M. Kleiman-Weiner, S. Dsouza, J. B. Tenenbaum, A. Shariff, J. F. Bonnefon, and I. Rahwan, "Drivers are blamed more than their automated cars when both make mistakes," *Nature Human Behaviour*, 2019.

[8] E. Pöllänen, G. J. M. Read, B. R. Lane, J. Thompson, and P. M. Salmon, "Who is to blame for crashes involving autonomous vehicles? Exploring blame attribution across the road transport system," *Ergonomics*, 2020. [Online]. Available: https://doi.org/10.1080/00140139.2020.1744064

[9] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of Experimental Social Psychology*, vol. 52, pp. 113–117, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jesp.2014.01.005

[10] M. Jörling, R. Böhm, and S. Paluch, "Service Robots: Drivers of Perceived Responsibility for Service Outcomes," *Journal of Service Research*, vol. 22, no. 4, pp. 404–420, 2019. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1177/1094670519842334

[11] D. A. Abbink, M. Mulder, and E. R. Boer, "Haptic shared control: Smoothly shifting control authority?" *Cognition, Technology and Work*, vol. 14, no. 1, pp. 19–28, 2012.

[12] D. Abbink, T. Carlson, M. Mulder, J. de Winter, T. L. Gibo, F. Aminravan, and E. R. Boer, "A Topology of Shared Control Systems – Finding Common Ground in Diversity," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, 2018.

[13] F. Flemisch, J. Kelsch, C. Löper, A. Schieben, and J. Schindler, "Automation spectrum , inner/outer compatibility and other potentially useful human factors concepts for assistance and automation," *Human Factors for Assistance and Automation*, no. 2008, pp. 1–16, 2008.

[14] F. Mars, M. Deroo, and J. M. Hoc, "Analysis of human-machine cooperation when driving with different degrees of haptic shared control," *IEEE Transactions on Haptics*, vol. 7, no. 3, pp. 324–333, 2014.

[15] P. G. Griffiths and R. B. Gillespie, "Sharing control between humans and automation using haptic interface: Primary and secondary task performance benefits," *Human Factors*, vol. 47, no. 3, pp. 574–590, 9 2005. [Online]. Available: http://journals.sagepub.com/doi/10.1518/001872005774859944

[16] S. M. Petermeijer, D. A. Abbink, M. Mulder, and J. C. De Winter, "The Effect of Haptic Support Systems on Driver Performance: A Literature Survey," *IEEE Transactions on Haptics*, vol. 8, no. 4, pp. 467–479, 2015.

[17] F. Flemisch, M. Heesen, T. Hesse, J. Kelsch, A. Schieben, and J. Beller, "Towards a dynamic balance between humans and automation: Authority, ability, responsibility and control in shared and cooperative control situations," *Cognition, Technology and Work*, vol. 14, no. 1, pp. 3–18, 2012.

[18] H. Zwaan, B. Petermeijer, and D. A. Abbink, "Haptic shared steering control with an adaptive level of authority based on time-to-line crossing," *IFAC PepersOnline*, vol. 52, no. 19, pp. 49–54, 2019. [Online]. Available: www.sciencedirect.com

[19] J. C. De Winter and D. Dodou, "Preparing drivers for dangerous situations: A critical reflection on continuous shared control," *Conference*

Proceedings - IEEE International Conference on Systems, Man and Cybernetics, pp. 1050–1056, 2011.

[20] M. Coeckelbergh, "Responsibility and the Moral Phenomenology of Using Self-Driving Cars," *Applied Artificial Intelligence*, vol. 30, no. 8, pp. 748–757, 2016. [Online]. Available: https://www.tandfonline.com/action/journalInformation?journalCode=uaai20

[21] P. A. Hancock and T. B. Sheridan, "The future of driving simulation," in *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. CRC Press, 1 2011, pp. 4–1.

[22] H. H. Kelley and J. L. Michela, "Attribution Theory and Research," *Annual Review Psychology*, vol. 31, pp. 457–501, 1980. [Online]. Available: www.annualreviews.org

[23] Y. Moon, M. Hall, and C. Nass, "Are computers scapegoats? Attributions of responsibility in human-computer interaction," Tech. Rep., 1998.

[24] H. H. Kelley, "Attribution theory in social psychology." *Nebraska Symposium on Motivation*, vol. 15, pp. 192–238, 1967.

[25] A. Dosovitskiy, G. Ros, F. Codevilla, A. López, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[26] M. M. Van Paassen, R. P. Boink, D. Abbink, M. Mulder, and E. Mc, "Four design choices for haptic shared control Simulator motion cueing View project Three-Dimensional Airborne Separation Assistance Displays View project," *Advances in Aviation Psychology*, vol. 2, pp. 237 – 254, 2017. [Online]. Available: https://www.researchgate.net/publication/318323318

[27] J. C. de Winter and D. Dodou, "National correlates of self-reported traffic violations across 41 countries," *Personality and Individual Differences*, vol. 98, pp. 145–152, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.paid.2016.03.091

[28] D. M. Oppenheimer, T. Meyvis, and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of Experimental Social Psychology*, vol. 45, no. 4, pp. 867–872, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.jesp.2009.03.009

[29] K. J. Preacher, D. D. Rucker, A. F. Hayes, K. J. Preacher, D. D. Rucker, A. F. Hayes, D. D. Rucker, and A. F. Hayes, "Addressing Moderated Mediation Hypotheses : Theory , Methods , and Prescriptions," *Multivariate behavioral*, vol. 42, no. 2, pp. 185–227, 2007. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00273170701341316

[30] J. R. Edwards and L. S. Lambert, "Methods for Integrating Moderation and Mediation: A General Analytical Framework Using Moderated Path Analysis," 2007.

[31] A. K. Montoya, "Moderation analysis in two-instance repeated measures designs: Probing methods and multiple moderator models," *Behavior Research Methods*, vol. 51, no. 1, pp. 61–82, 2019.

[32] A. F. Hayes, *Introduction to mediation, moderation, and conditional process analysis : a regression-based approach*, 2nd ed. New York: The Guilford Press, 2013.

[33] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, 2015.

[34] A. F. Hayes, K. J. Preacher, P. Andrew, F. Hayes, A. K. Montoya, J. P. Selig, E. Page-Gould, and A. Sharples, "ADVANCES IN WITHIN-SUBJECT MEDIATION ANALYSIS," *SPSP Annual Convention*, 2016.

[35] S. Nakagawa, P. C. Johnson, and H. Schielzeth, "The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded," *Journal of the Royal Society Interface*, vol. 14, no. 134, 9 2017.

[36] J. Nadler and M.-H. Mcdonnell, "Moral Character, Motive, and the Psychology of Blame," *Cornell Law Review*, vol. 97, no. 2, pp. 256–304, 2012. [Online]. Available: http://scholarship.law.cornell.edu/clrhttp://scholarship.law.cornell.edu/clr/vol97/iss2/3

[37] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," pp. 50–80, 1 2004. [Online]. Available: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50_30392

[38] A. F. Hayes, "An Index and Test of Linear Moderated Mediation," *Multivariate Behavioral Research*, vol. 50, no. 1, 2015.

[39] B. T. Gomez and J. M. Wilson, "Political Sophistication and Economic Voting in the American Electorate: A Theory of Heterogeneous Attribution," *American Journal of Political Science*, vol. 45, no. 4, p. 899, 2001.

[40] M. F. Ibrahim, M. Y. Misro, A. Ramli, and J. M. Ali, "Maximum safe speed estimation using planar quintic Bezier curve with C2 continuity," in *AIP Conference Proceedings*, vol. 1870. American Institute of Physics Inc., 8 2017.

[41] C. Ugrinowitsch, G. W. Fellingham, and M. D. Ricard, "Limitations of ordinary least squares models in analyzing repeated measures data," *Medicine and Science in Sports and Exercise*, vol. 36, no. 12, pp. 2144–2148, 2004.

[42] R. A. Armstrong, "Recommendations for analysis of repeated-measures designs: testing and correcting for sphericity and use of ¡scp¿manova¡/scp¿ and mixed model analysis," *Ophthalmic and Physiological Optics*, vol. 37, no. 5, pp. 585–593, 9 2017. [Online]. Available: http://doi.wiley.com/10.1111/opo.12399

[43] K. Vanbrabant, Y. Boddez, P. Verduyn, M. Mestdagh, D. Hermans, and F. Raes, "A new approach for modeling generalization gradients: a case for hierarchical models," *Frontiers in Psychology*, vol. 6, 5 2015.

[44] R. G. O'Brien and M. K. Kaiser, "MANOVA for analyzing repeated measurement design: An extensive primer," *Psychological Bulletin*, vol. 92, no. 2, pp. 316–333, 1985.

[45] T. A. Snijders and R. J. Bosker, *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*, 2nd ed. SAGE Publications, 2012.

[46] G. Molenberghs and G. Verbeke, *Linear Mixed Models for Longitudinal Data*, ser. Springer Series in Statistics. New York, NY: Springer New York, 2000. [Online]. Available: http://link.springer.com/10.1007/978-1-4419-0300-6

[47] W. Van Winsum, K. A. Brookhuis, and D. De Waard, "A comparison of different ways to approximate time-to-line crossing (TLC) during car driving," *Accident Analysis and Prevention*, vol. 32, no. 1, pp. 47–56, 1 2000.

A

Road and guidance design

Road environment

The radius R for which driving is comfortable and safe at a given design speed V_d is dependent on both the pavement super-elevation rate e, and the tire-pavement side friction factor f, and is calculated as:

$$R = \frac{V_d^2}{127 \cdot (e + f)}$$

Given our cruise-control speed ($V_d = 80$ km/h), a safe side friction factor ($f = 0.16$) [40], and for simplicity no super-elevation ($e = 0$), we found a comfortable and safe curve radius of $R_c = 300$ m.

Haptic guidance: Four Design Choice Architecture

The FDCA controller used in this experiment is based on that of [18] and consists of the following four components (see Figure A.1):

- *Human Compatible Reference (HCR)*:
  In the current research two pre-recorded reference trajectories for the controller to follow. Each trajectory consists of four variables; the x- and y-positions of the car as a reference for the lateral error, the cars heading as a reference for the heading error, and the steering wheel angles used to create this reference trajectory.
- *Level of Haptic Support (LoHS)*:
  The feedforward fraction of haptic support that the system contributes to the control effort to follow the HCR. The total feedforward torque (τ_{hcr}) required to follow the HCR, when no disturbances are present, is the product of the reference steering wheel angle (θ_{HCR}) and the inverse steering dynamics (H_{sw}^{-1}). The actual feedforward contribution (τ_{FF}) to the haptic guidance torque (τ_{HSC}) is a product of the percentage of LoHS (λ_{LoHS}) and τ_{hcr}:

$$\begin{aligned} \tau_{FF} &= \lambda_{LoHS} \cdot \tau_{hcr} \\ &= \lambda_{LoHS} \cdot H_{sw}^{-1} \cdot \theta_{HCR} \end{aligned}$$

- *Strength of Haptic Feedback (SoHF)*:
  The feedback gains needed to correct for deviations from the HCR. The desired feedback steering wheel angle θ_{FB} is essentially calculated using a proportional controller that penalizes the lateral error (Δ_y) and the heading error (Δ_ψ) with their respective gains (K_y and K_ψ) multiplied by a pre-factor (λ_{SoHF}). The feedback torques required to generate these angles are then calculated by multiplying H_{sw}^{-1} with θ_{FB} as in:

$$\begin{aligned} \theta_{FB} &= \lambda_{SoHF} \cdot (K_y \Delta_y + K_\psi \Delta_\psi) \\ \tau_{FB} &= H_{sw}^{-1} \cdot \theta_{FB} \end{aligned}$$

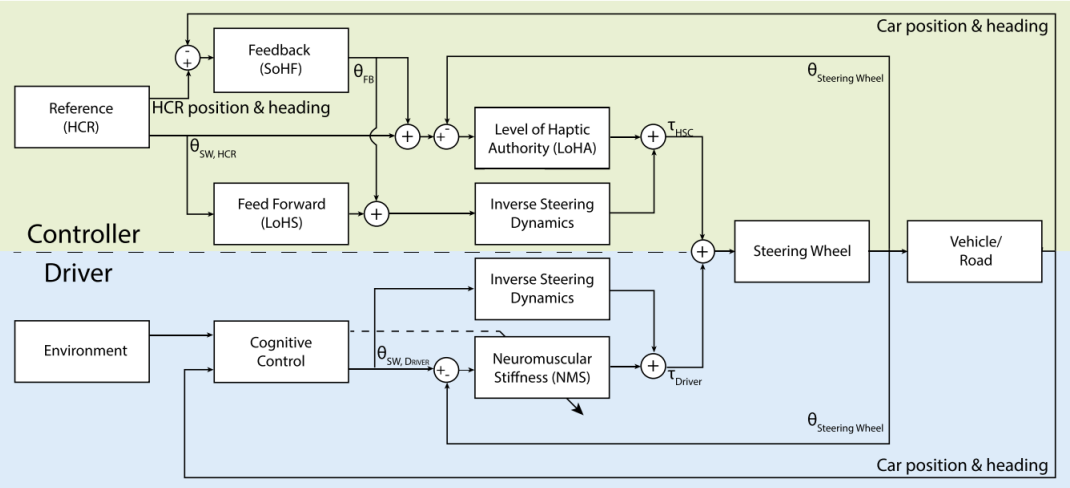

Figure A.1: A block scheme of the FDCA-controller (top, green) and the driver (bottom, blue) in a haptic shared steering control system, adapted from [18, Fig. 1]. Note that this implementation uses fixed LoHA settings as opposed to the Time to Lane Crossing dependent LoHA of the original author.

- *Level of Haptic Authority (LoHA)*:
  A virtual spring around the controller's desired steering wheel angle. In a sense it determines how strong the controller enforces the feedback and feedforward torques. It functions as an added torsional resistance that determines the amount of effort it takes to counteract the controller. The Level of Haptic Authority torques (τ_{LoHA}) are calculated by multiplying the difference in steering wheel angles (Δ_θ, defined below) by the LoHA gain (K_{LoHA}) as in:

$$\Delta_\theta = \theta_{FB} + \theta_{FF} - \theta_{sw}$$
$$\tau_{LoHA} = K_{LoHA} \cdot \Delta_\theta$$

The total amount of torque τ_{HSC} that the controller actually generates on the steering wheel as haptic guidance is the sum of these torques:

$$\tau_{HSC} = \tau_{FF} + \tau_{FB} + \tau_{LoHA}$$

B

Linear mixed-effect model for within-subject design

To estimate the regression coefficients in Eq. 2 and 3 we use LME modeling as opposed to regular regression analysis methods like Ordinary Least Squares (OLS) models. We decided to use LME models for two reasons.

Firstly, because OLS models have been proven to inflate probabilities of Type I errors in repeated measures analyses when the variance/covariance structure of the data set is not compound symmetric (CS) [41], or does not satisfy the assumption of sphericity. That is, when not all response variables have the same variance, and not all pairs of response variables share a common correlation [42]. This assumptions rarely holds for within-subject design [43, 44]. However, with LME models violation of the assumption of sphericity is of no concern, because the variance and covariance are explicitly included in the model [45].

Secondly, because the LME framework handles within-participant clustering of data by using levels, i.e. repeated measurements (level-1) for every subject (level-2). Where regular regression analyses only allow regression parameters common to all subjects (i.e., fixed effects), LME models also allow parameters that model these subject-specific deviations (i.e., random effects) [43]. The clustering of the data can thus be accounted for by these random effects [46]. This is especially relevant in mediation analyses when not taking random slopes into account inevitably leads to biased indirect and total effects [34]. Not taking into account random slopes refers to calculating the average of paths a and b before calculating the indirect effect ω (i.e. $\omega = ab$).

For statistical validation of our results we apply the BCA bootstrap method using 10,000 repetitions with stratified re-sampling for which participant ID is used as the grouping factor. Foremost, because it does not assume normally distributed random effects [34] , but also because it is widely accepted within the field of statistical mediation analyses [38].

Random slopes for subject-specific differences

Results of LME models for within-participant mediation analyses can be biased if you don't account for random effects between participants. In the simple mediation case, the degree to which the indirect and total effects are biased when not taking into account random effects depends on how much the slopes a and b covary [34] , see Fig. B.1. For our moderated mediation we argue that the same holds if we replace the b-path by the moderated b-path.

In our model we used participant ID as a grouping factor by expanding the a_1, the b_j (for $j = 1-4$), and the c_1' coefficients in Eq. 2 and 3

$$\omega = \frac{1}{N}\sum_{i}^{N}\omega_i$$

For the total effect ε we first calculate the total effects for each participant and then take the average over all participants. The subject-specific total effects are defined as the sum of ω_i and γ_i:

$$\varepsilon = \frac{1}{N}\sum_{i}^{N}(\omega_i + \gamma_i)$$

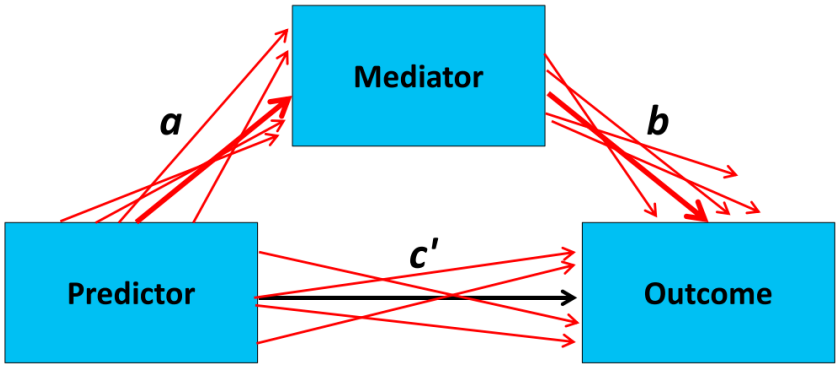

Figure B.1: Simple mediation model taken from [34]. Narrow arrows indicate per subject random slopes, bold arrows indicate fixed slopes. The indirect effect that the predictor, via the mediator, has on the outcome is given by the product of path a and path b. The direct effect of predictor on outcome is given by path c'.

Contrast coding

One of the advantages of LME models is that they allow for simple contrast analyses by contrast coding fixed effects [42]. First, we do a regression analysis on the observations of the M condition and the L condition to examine to the effect that adding haptic guidance has to the perception of responsibility. Then, we examine how a further increase in LoHA affects this perception by doing a regression analysis on the observations of the L condition and the H condition.

We contrast-coded these dichotomous cases to differ exactly a single unit (i.e., analysis 1: Manual = $-.5$ and Low = .5, analysis 2: Low = $-.5$ and High = .5), see Table IV. Coefficients of interacting variables are sensitive to the magnitude of said variables; larger magnitudes increase the product of these variables and, as a result, decrease the coefficients that correspond to these interacting variables.

For this same reason we mean centered the mediating variable M in Eq. 2 and 3 to prevent unnecessary inflation/deflation of the (moderated) mediation indices $a_1 b_j$. This has no effect on the value of ω and γ but makes the individual regression coefficients interpretable within the range of the data [38].

Additionally, we contrast-coded our dichotomous moderators to also differ exactly a single unit (i.e., V, and $Q = -.5$ or .5 see Table IV). In doing so, their respective indices of moderated mediation equate the differences between two corresponding conditional indirect effects. For example, $a_1 b_3$ represents the effect that changing from a negative outcome ($Q_{negative} = -.5$) to a positive outcome ($Q_{positive} = .5$) has on ω.

C

Informed Consent Form

Consent Form

Impact of control authority on perceived responsibility and blame: a haptic shared control driving study

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information dated [../../….], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
I understand that taking part in the study involves a survey questionnaire completed by me (participant).	☐	☐
Use of the information in the study		
I understand that information I provide will be used for a master thesis report and potentially for further studies within the same area of interest.	☐	☐
I understand that personal information collected about me that can identify me, such as [e.g. my name, age, e-mail adres], will not be shared beyond the study team.	☐	☐
Future use and reuse of the information by others		
I give permission for the survey database that I provide to be archived in 4TU so it can be used for future research and learning.	☐	☐

_____	_____	_____
Name of participant	Signature	Date

I have accurately read or handed out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

__Cedric H. Kok___	_____	_____
Researcher name	Signature	Date

Study contact details for further information:

Cedric H. Kok
06-83598754
c.h.kok@student.tudelft.nl

D

Experiment Briefing

Experiment Briefing

Impact of control authority on perceived responsibility and blame: a haptic shared control driving study

Thank you for participating in this experiment. This experiment, conducted at the Cognitive Robotics Laboratory (CoR-Lab), examines the human drivers' user experience when interacting with two types of Advanced Driver Assistance Systems (ADAS). Most ADAS are designed so that the human driver and automation trade control over a dynamic driving task (DDT); either the driver is in control or the automation. In Haptics Shared Control (HSC) steering both the driver and automation jointly influence the control over a DDT by applying torques on the steering wheel. In other words, the automation guides the driver by applying a guidance torque on the steering wheel. One of the main issues facing the development and release of intelligent vehicles is how people will think about its responsibility, and whether it should be held accountable for what happens with it on the road.

This briefing will describe what this experiment entails and what is expected of you as a participant.

Experiment goal

The goal of this experiment is to examine how the use of a HSC steering wheel in semi-automated driving influences the driving experience of the human driver. Specifically, we are interested in your experience of how the HSC control setting impacts your perceived sense of responsibility, and how much you feel in control. Knowledge on how your perceived responsibility is impacted by the amount of control you experience over an automated car is imperative for the design of future automated vehicles.

Experiment task

In this task, you will drive a car with cruise control (80 km/h) on a two-lane street without oncoming traffic. It is your goal to drive on the road and interact with the traffic scenarios as if you were driving on a real road. Your task is to steer the vehicle using the steering wheel, the cruise control will take care of the car's velocity. You will be asked to drive as you normally would, while keeping your hands at 10 to 2 o'clock positions. You are expected to drive in the right lane of the road unless the traffic situation requires otherwise. After each 4-min drive you will be asked to fill in a task-related questionnaire.

Sometimes you will feel torques on the steering wheel which may help or hinder you (or the car might steer itself). You are in principle able to override these torque. Although the car may be helping you with steering, you always have final control so you can intervene at any time. The first part of the track consists of a straight road. Feel free to gently probe the guidance system to get a feel for the system you are driving with.

The experimental setup consists of a SensoDrive force feedback steering wheel positioned in front of a 65-inch 4K screen. On the screen you will see part of the interior of the car and the surrounding landscape as if you were sitting in the driver's seat. The speed of the car will be shown on the front window.

Experiment procedure

The experiment consists of two parts: a fixed-base simulator driving task, and a task-related questionnaire. During this experiment, you will drive the same road for a total of seven runs. You start with a manual driving run so that you can get accustomed to the steering wheel. After each run, you will be asked to fill out a questionnaire about your driving experience.

Each run lasts 4 minutes. The researcher will ask you if you need a break between runs after filling in the questionnaire. The experiment will take approximately ~45 minutes.

For each driving task, the subsequent procedure will be followed:

1. The researcher applies the settings for the next trial.
2. The researcher asks whether the participant is ready to proceed (i.e., any signs of discomfort) and initiates the run after a countdown from 3 (3-2-1-go).
3. The participant performs the driving trial.
4. The participant is asked to fill out a questionnaire to assess their driving experience.

Other instructions

It is possible that some participants may develop nausea (simulator sickness) during the tests. In case you experience any signs of discomfort, you are asked to **inform the researcher** immediately and the experiment will be stopped.

For safety reasons you are asked to **keep your thumbs on** the steering wheel. This prevents your thumbs from being in the way in the unlikely event that the steering wheel turns unexpectedly.

- You can withdraw from the study at any moment without giving any reason
- If you want to withdraw from the study and have your data removed after the experiment, contact Cedric

Additional information regarding COVID-19

To prevent the spread of the coronavirus (in compliance with the university's policy), researchers and participants in the study:

- have to be younger than 70 years
- do not have any underlying ailments that could be seen as a risk-factor for a COVID-19 infection
- do not have any complaints or symptoms that could be indicative of a COVID-19 infection
- have not been in contact with a COVID-19 patient at least 14 days before participation in the study
- take suitable protective measures if a minimum distance of 1.5 meters is not viable
- are enabled to travel outside of rush hours to and from the research location

Also any objects or surfaces researchers and participants come into contact will be disinfected prior and after use.

Contact information researcher:	Contact information research supervisor
Cedric H. Kok	Prof. dr. ir. D.A. Abbink
c.h.kok@student.tudelft.nl	d.a.abbink@tudelft.nl
+31 683 59 87 54	+31 15 27 82077

Thank you for participating!

E

Post Trial Questionnaire

Trial1

For the following statements, please fill in to what extent you agree with these.

I feel in control steering this car.

Strongly disagree			Neither agree nor disagree			Strongly agree
○	○	○	○	○	○	○

The car lets me (the driver) have control over steering.

Strongly agree			Neither agree nor disagree			Strongly disagree
○	○	○	○	○	○	○

The car follows a trajectory on the road that I would normally drive myself.

Strongly disagree			Neither agree nor disagree			Strongly agree
○	○	○	○	○	○	○

For the following questions, you will indicate how much each actor is responsible for the given outcome. There are no right or wrong answers. Just give your honest opinion on how much you feel that each actor is responsible.

Assume that you and the car evaded all of the roadblocks. To what extent do **you** (the driver) feel **responsible** for this outcome?

Not at all			A moderate amount			Completely
○	○	○	○	○	○	○

Assume that you and the car evaded all of the roadblocks. To what extent do you feel that the car, specifically **the car manufacturer,** is **responsible** for this outcome?

Not at all			A moderate amount			Completely
○	○	○	○	○	○	○

I am paying attention.

Completely			A moderate amount			Not at all
○	○	○	○	○	○	○

Now imagine that you and the car did crash into one of the roadblocks. To what extent do **you** (the driver) feel **responsible** for this outcome?

Not at all			A moderate amount			Completely
○	○	○	○	○	○	○

Now imagine that you and the car did crash into one of the roadblocks. To what extent do you feel that the car, specifically **the car manufacturer,** is **responsible** for this outcome?

Not at all			A moderate amount			Completely
○	○	○	○	○	○	○

F

Generic Characteristics Questionnaire

Generic Characteristics

What is your age?

What is your gender?

- Female
- Male
- I prefer not to answer
- Other, please specify

What is your primary mode of transport?

- Private automobile
- Private motorcycle
- Public transport
- Human powered transportation (e.g. walking, cycling)
- I prefer not to answer
- Other, please specify

At which age did you obtain your first driver's license?

On average, how often did you drive a vehicle in the last months?

- Every day
- 4 to 6 days a week
- 1 to 3 days a week
- Once a month to once a week
- Less than once a week
- Less than once a month
- Never
- I prefer not to answer

Roughly how many kilometers did you drive in the last 12 months?

- 0
- 1-1.000
- 1.001-5.000
- 5.001-10.000
- 10.001-15.000
- 15.001-20.000
- 20.001-25.000
- 25.001-35.000
- 35.001-50.000
- 50.001-100.000
- More than 100.000
- I prefer not to answer

G

Extensive results and additional analysis

This appendix shows the results and figures that couldn't be included in the paper. It contains the analysis of the following topics:

- Probed results of the moderated mediation analyses
- Steering torques (guidance torques and driver torques)
- Raw car positions
- Safety margins (in terms of TLC and TC)
- Questionnaire results
- Within subject responsibility shifts

Moderated mediation analysis

Table G.1: Indices of (moderated) mediation for both regression analyses. Significant results are given in bold.

	Manual → Low				Low → High			
	Car		Human		Car		Human	
	Coeff.	95% CI	Coeff.	95% CI	Coeff.	95% CI	Coeff.	95% CI
$a_1 b_1$ →	**.48**	**[21, .80]***	**-.33**	**[-.50, -.15]***	.03	[-.17, .26]	-.03	[-.24, .18]
$a_1 b_2$ →					-.00	[-.04, .02]	.01	[-.01, .07]
$a_1 b_3$ →	-.02	[-.18, .14]	**-.17**	**[-.37, -.01]***	-.00	[-.06, .01]	-.02	[-.08, .02]
$a_1 b_4$ →					.01	[-.02, .13]	.01	[-.02, .13]

* $0 \notin$ 95% CI

Direct, indirect and total effects for the Manual versus Low LoHA model

Table G.2: Probed direct (γ), indirect (ω) and total ε effects for the Manual → Low LoHA model. Significant effects are given in bold. Robot action was not included in the model.

	Direct effect (γ)		Indirect effect (ω)		Total effect ε	
	Coeff.	95% CI	Coeff.	95% CI	Coeff.	95% CI
Human						
positive	**-1.27**	**[-1.71, -.86]***	**-.42**	**[-.62, -.20]***	**-1.67**	**[-2.06, -1.28]***
negative	**-1.55**	**[-2.04, -1.21]***	**-.24**	**[-.42, -.06]***	**-1.79**	**[-2.17, -1.44]***
Car						
positive	**1.28**	**[.52, 2.06]***	**.47**	**[.19, .79]***	**1.75**	**[1.07, 2.43]***
negative	**1.59**	**[.90, 2.33]***	**.49**	**[.21, .82]***	**2.09**	**[1.46, 2.70]***

* $0 \notin$ 95% CI

Direct, indirect and total effects for the Low versus High LoHA model

Table G.3: Probed direct (γ), indirect (ω) and total ε effects for the Low → High LoHA model. Significant effects are given in bold.

	Direct effect (γ)		Indirect effect (ω)		Total effect ε	
	Coeff.	95% CI	Coeff.	95% CI	Coeff.	95% CI
Human						
good action						
positive	-.19	[-.86, .56]	-.03	[-.24, .18]	-.22	[-.87, .46]
negative	.18	[-.39, .86]	-.02	[-.22, .18]	.16	[-.39, .73]
bad action						
positive	-.37	[-1.10, .38]	-.05	[-.28, .19]	-.42	[-1.15, .28]
negative	.07	[-.57, .75]	-.03	[-.23, .18]	.04	[-.59, .68]
Car						
good						
positive	.39	[-.270, 1.00]	.03	[-.17, .26]	.42	[-.21, 1.00]
negative	.14	[-.495, .71]	.03	[-.17, .26]	.17	[-.40, .72]
bad						
positive	-.12	[-1.06, .70]	.02	[-.17, .27]	-.10	[-1.01, .75]
negative	**.86***	**[.15, 1.56]***	.03	[-.17, .27]	**.89**	**[.23, 1.59]***

* $0 \notin 95\%$ CI

Steering torques

Steering torques [Nm] during main part of track that were superimposed on the steering wheel as guidance

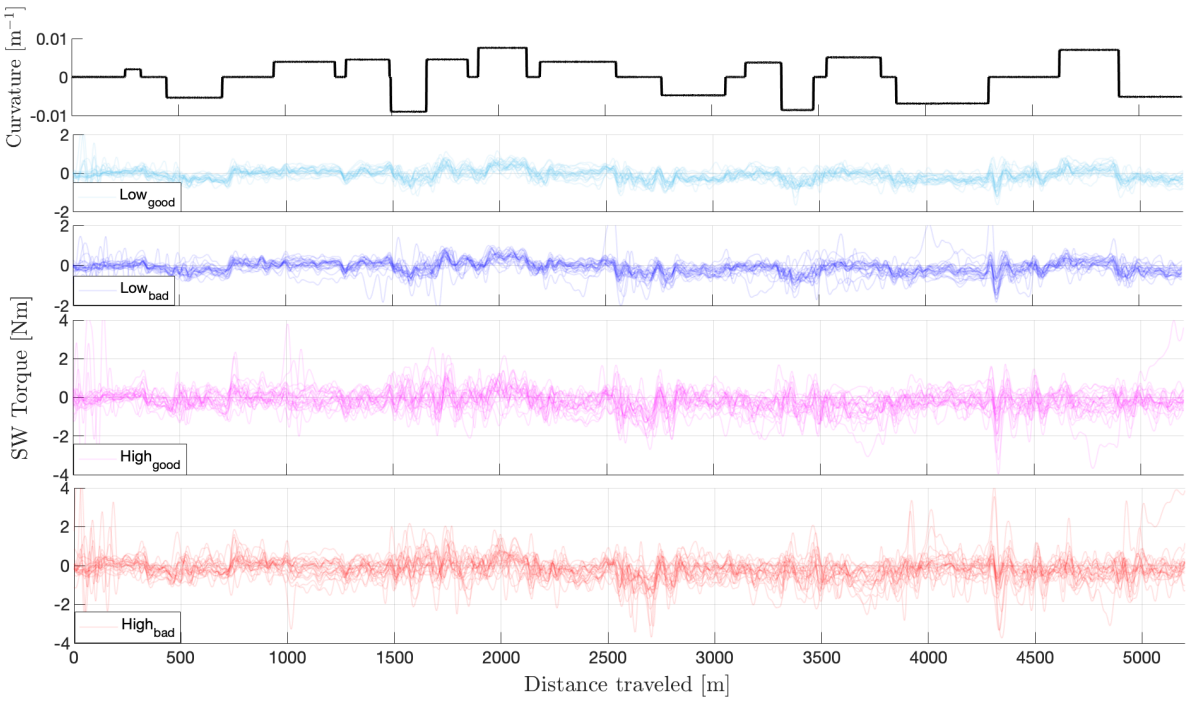

Figure G.1: Raw guidance torques [Nm] for all participants during the main part of the track for the low and high LoHA conditions. Manual condition had no guidance torques superimposed on the steering wheel. From top to bottom: Road curvature, Low$_{good}$, Low$_{bad}$, High$_{good}$, High$_{bad}$.

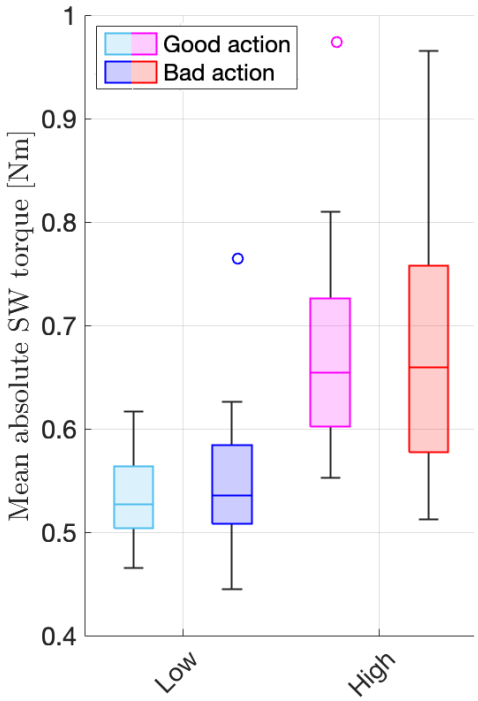

Figure G.2: Mean absolute guidance torques [Nm] for all participants during the main part of the track for the low and high LoHA conditions. Manual condition had no guidance torques superimposed on the steering wheel.

Steering torques [Nm] during main part of track that were generated by the driver

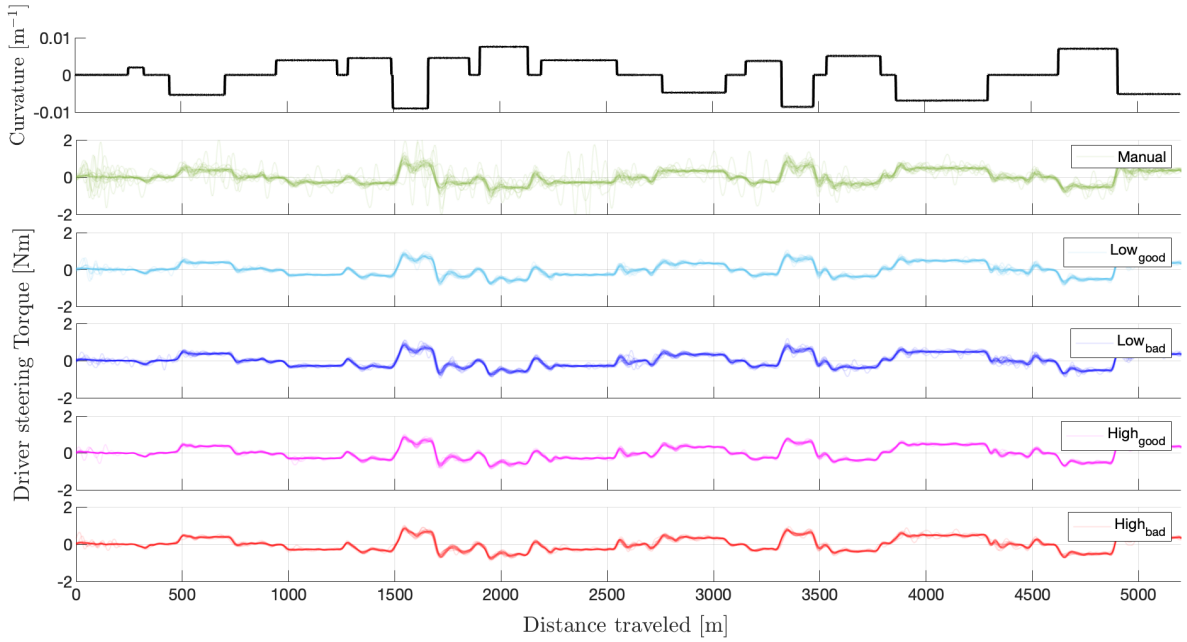

Figure G.3: Raw driver torques [Nm] for all participants during the main part of the track. From top to bottom: Road curvature, Manual, Low_{good}, Low_{bad}, High_{good}, High_{bad}.

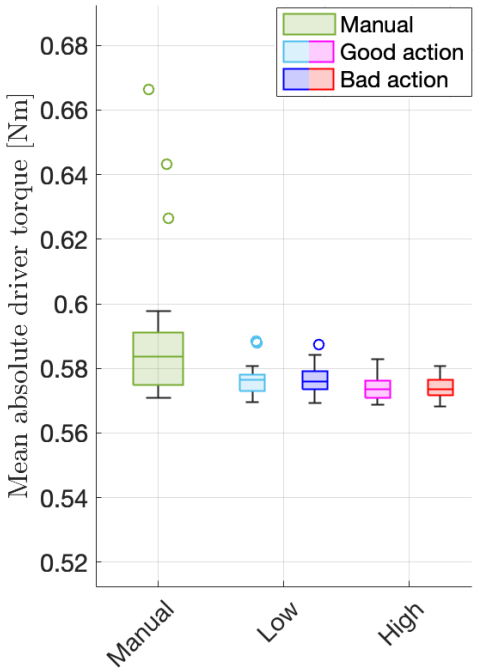

Figure G.4: Mean absolute driver torques [Nm] during the main part of the track.

Steering torques [Nm] during navigation around obstruction that were superimposed on the steering wheel as guidance

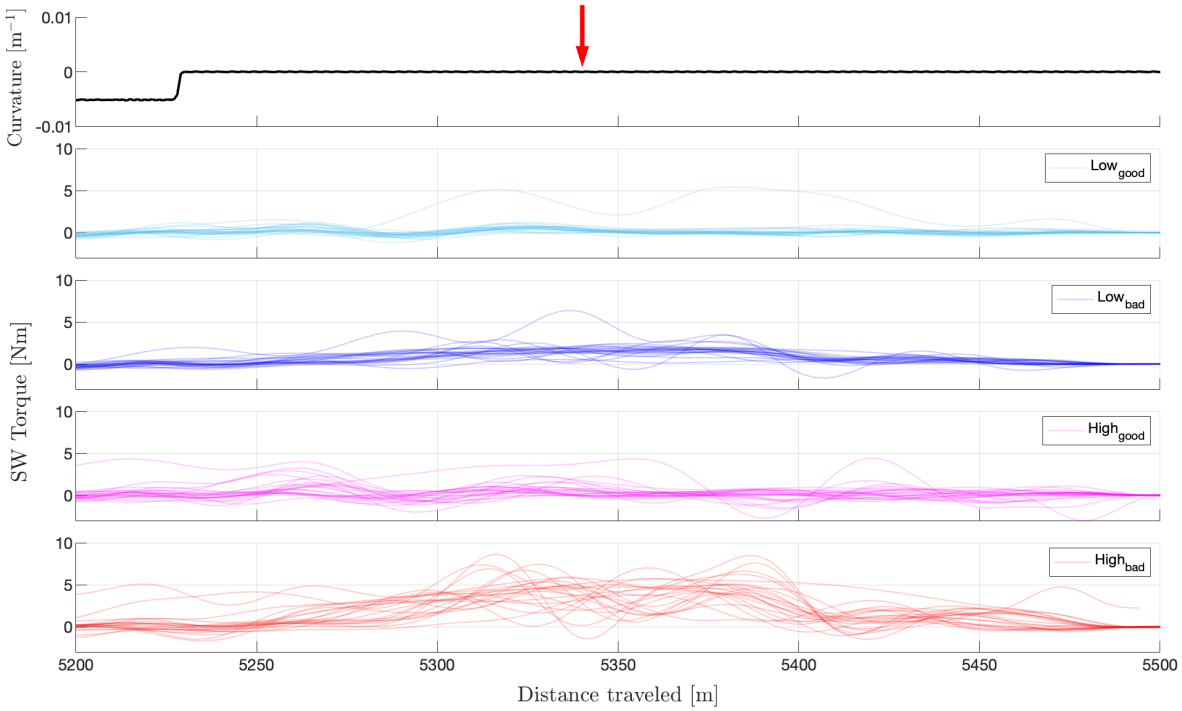

Figure G.5: Raw guidance torques [Nm] for all participants during navigation around final obstruction (red arrow) for the low and high LoHA conditions. Manual condition had no guidance torques superimposed on the steering wheel. From top to bottom: Road curvature, Low$_{good}$, Low$_{bad}$, High$_{good}$, High$_{bad}$.

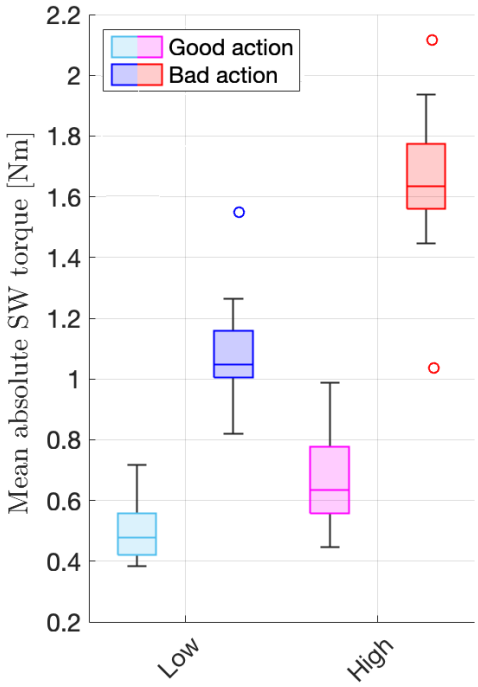

Figure G.6: Mean absolute guidance torques [Nm] during navigation around final obstruction for the low and high LoHA conditions. Manual condition had no guidance torques superimposed on the steering wheel.

Steering torques [Nm] during navigation around obstruction that were generated by the driver

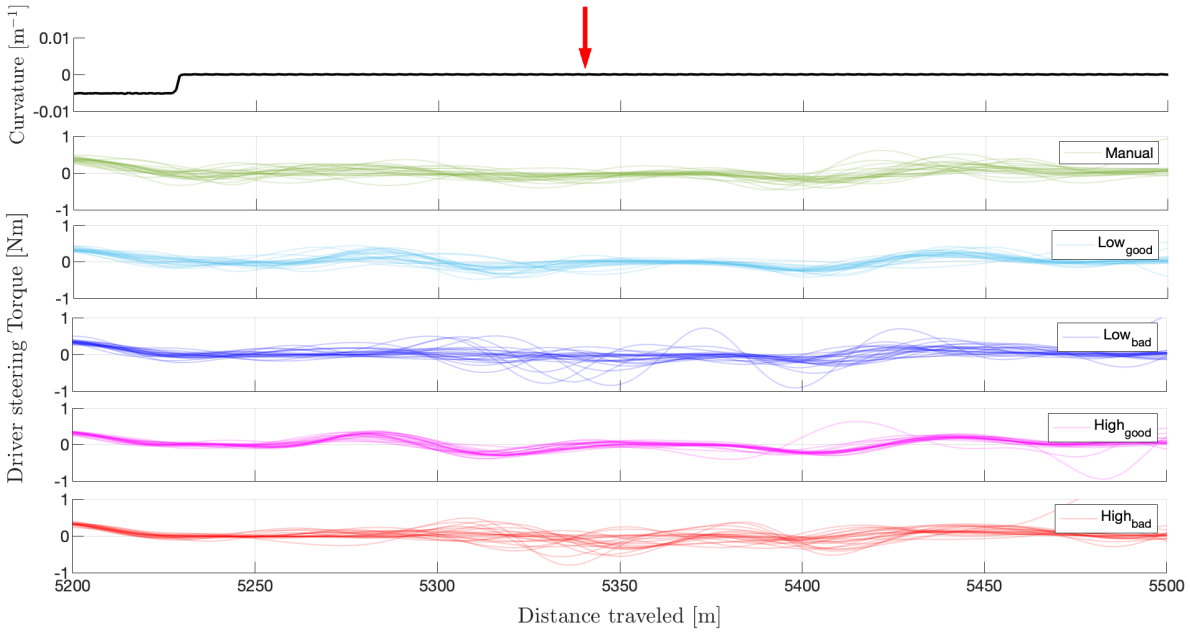

Figure G.7: Raw driver torques [Nm] for all participants during navigation around final obstruction (red arrow). From top to bottom: Road curvature, Manual, Low_{good}, Low_{bad}, High_{good}, High_{bad}.

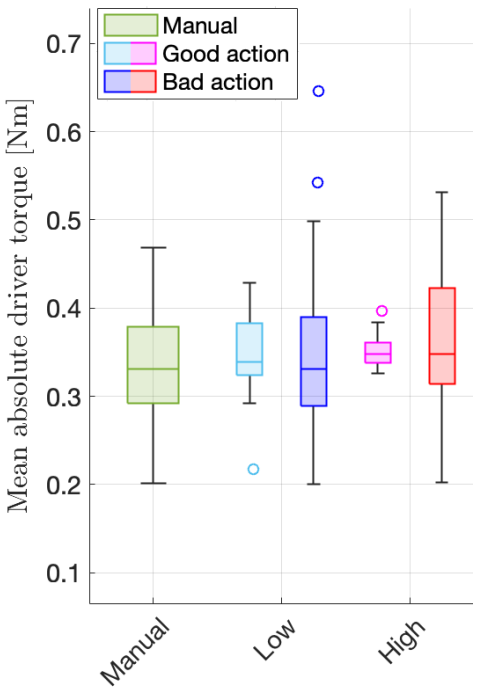

Figure G.8: Mean absolute driver torques [Nm] during navigation around final obstruction.

Car positions during last obstruction

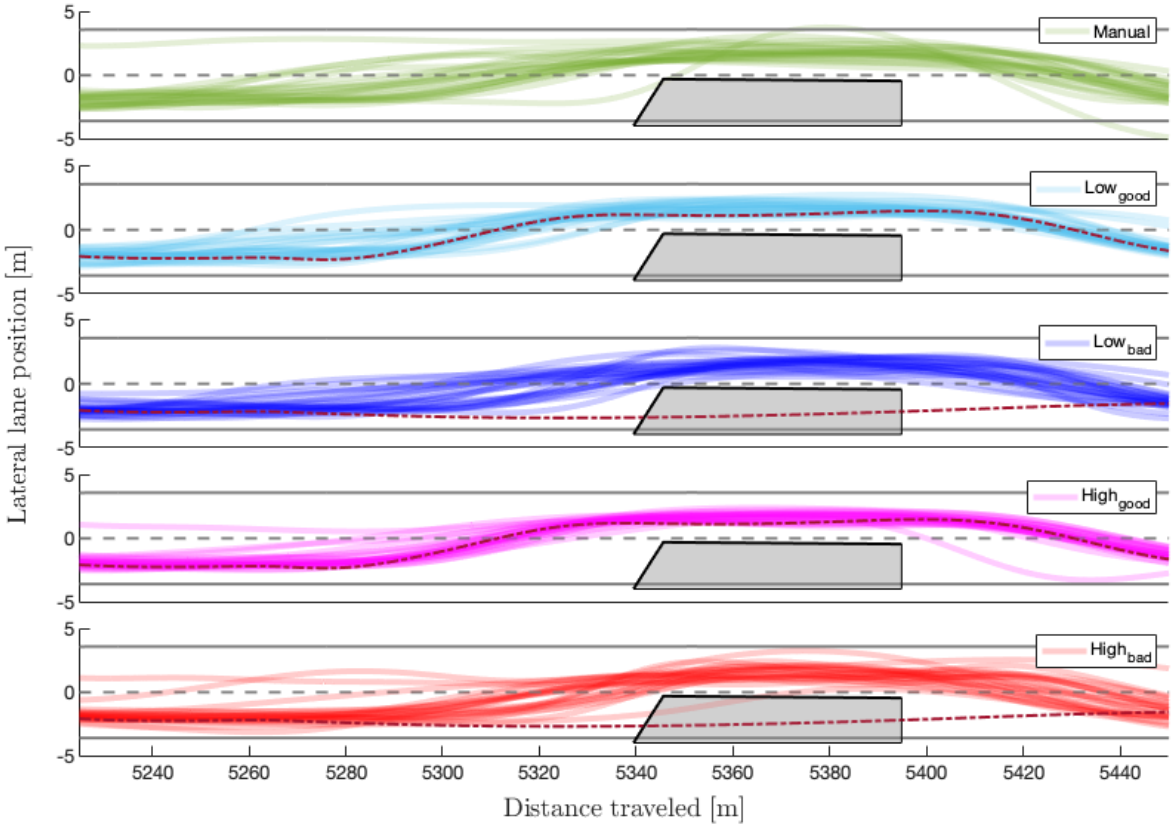

Figure G.9: Lateral lane position [m] of individual participants and trial specific HCR (dash-dotted line) per condition during navigation of final obstruction (gray area). From top to bottom: Manual, Low_{good}, Low_{bad}, High_{good}, High_{bad}.

Safety margins results (TLC and TTC)

Time to Line Cross (TLC) with respect to left and right lane boundaries

The Time to Line Cross (TLC) with respect to left and right lane boundaries was approximated using the car's lateral speed and acceleration [47]. Results are based on the final 250 meters of the track.

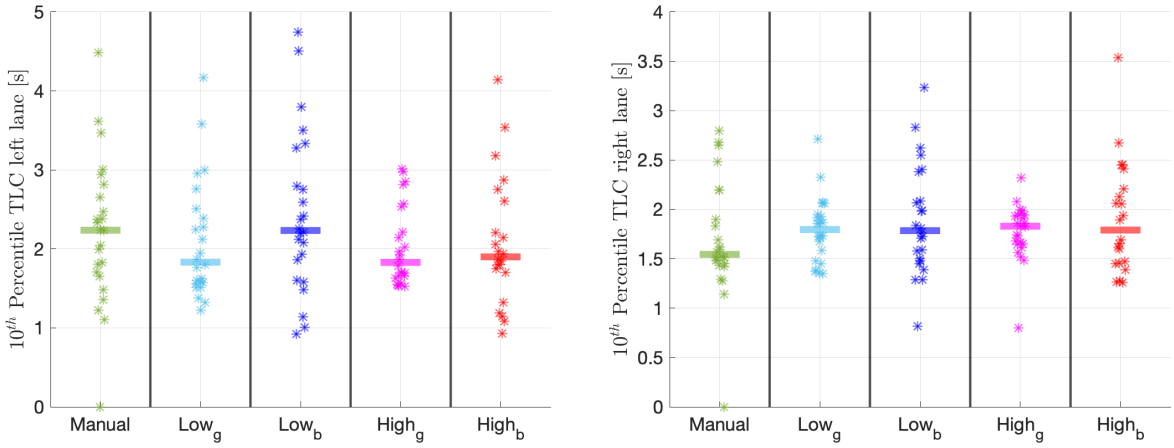

Figure G.10: 10^{th} percentile TLC [s] of lane boundaries for individual participants (stars) and median TLC [s] across all participants (horizontal line) per condition during navigation of the final road obstruction. Left figure: with respect to left lane. Right figure: with respect to right lane.

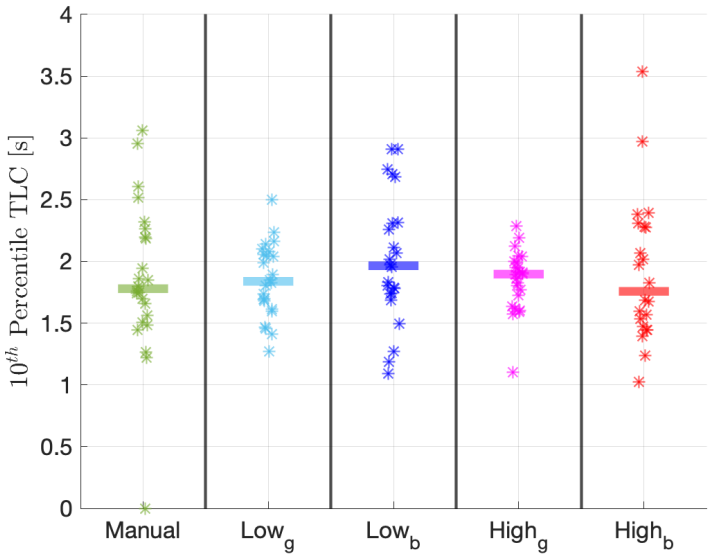

Figure G.11: 10^{th} percentile TLC [s] with respect to both lane boundaries for individual participants (stars) and median TLC [s] across all participants (horizontal line) per condition during navigation of the final road obstruction.

Time to Collision (TC) with respect to final road obstruction

The Time to Collision (TC) with respect to the road obstruction was approximated using the car's longitudinal speed. Results are based on the final 250 meters of the track.

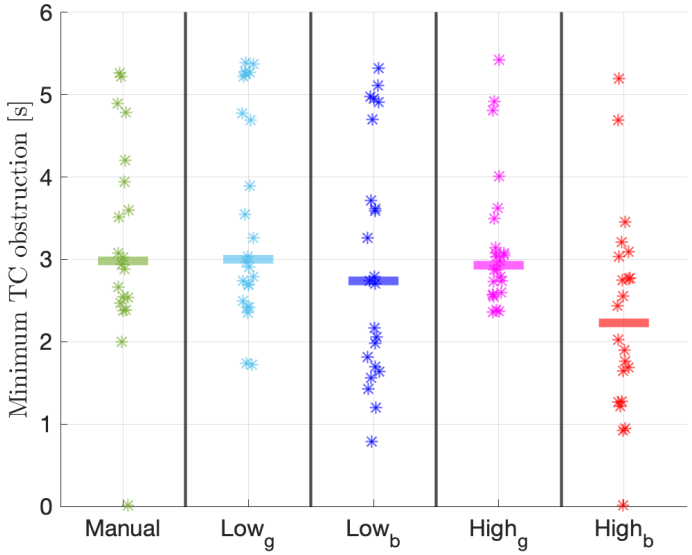

Figure G.12: Minimum TC [s] of individual participants (stars) and median TLC [s] across all participants (horizontal line) per condition during navigation of the final road obstruction.

Questionnaire results

Questionnaire results for perceived authority and control

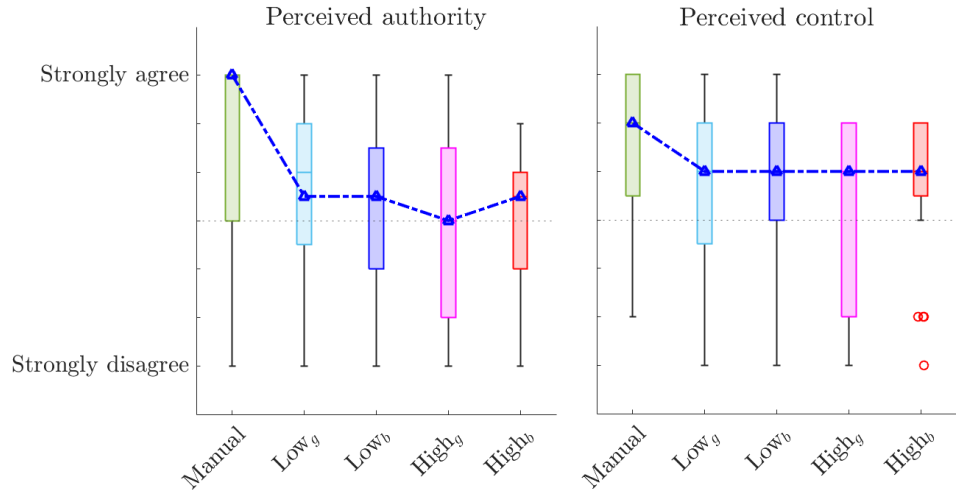

Figure G.13: Boxplot of the questionnaire results and corresponding median (blue triangles and line) for perceived authority (left) and control (right) for all conditions.

Questionnaire results for attributed outcome responsibility

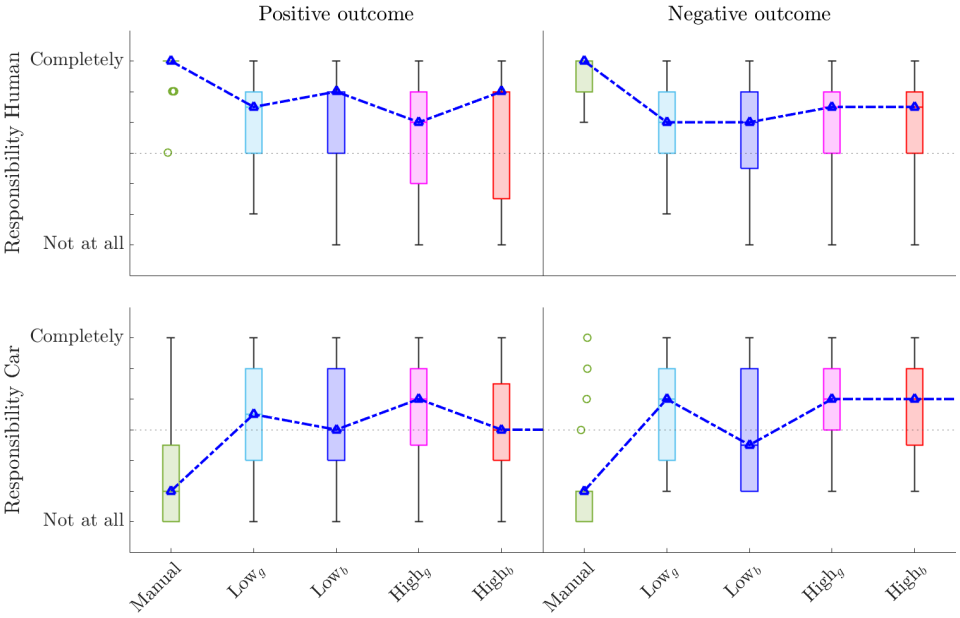

Figure G.14: Boxplot of the questionnaire results and corresponding median (blue triangle and line) for responsibility towards the driver (top) and towards the car (bottom) for both positive outcomes (left) and negative outcomes (right) for all conditions.

Within-subject responsibility shifts

Responsibility shift for positive outcomes

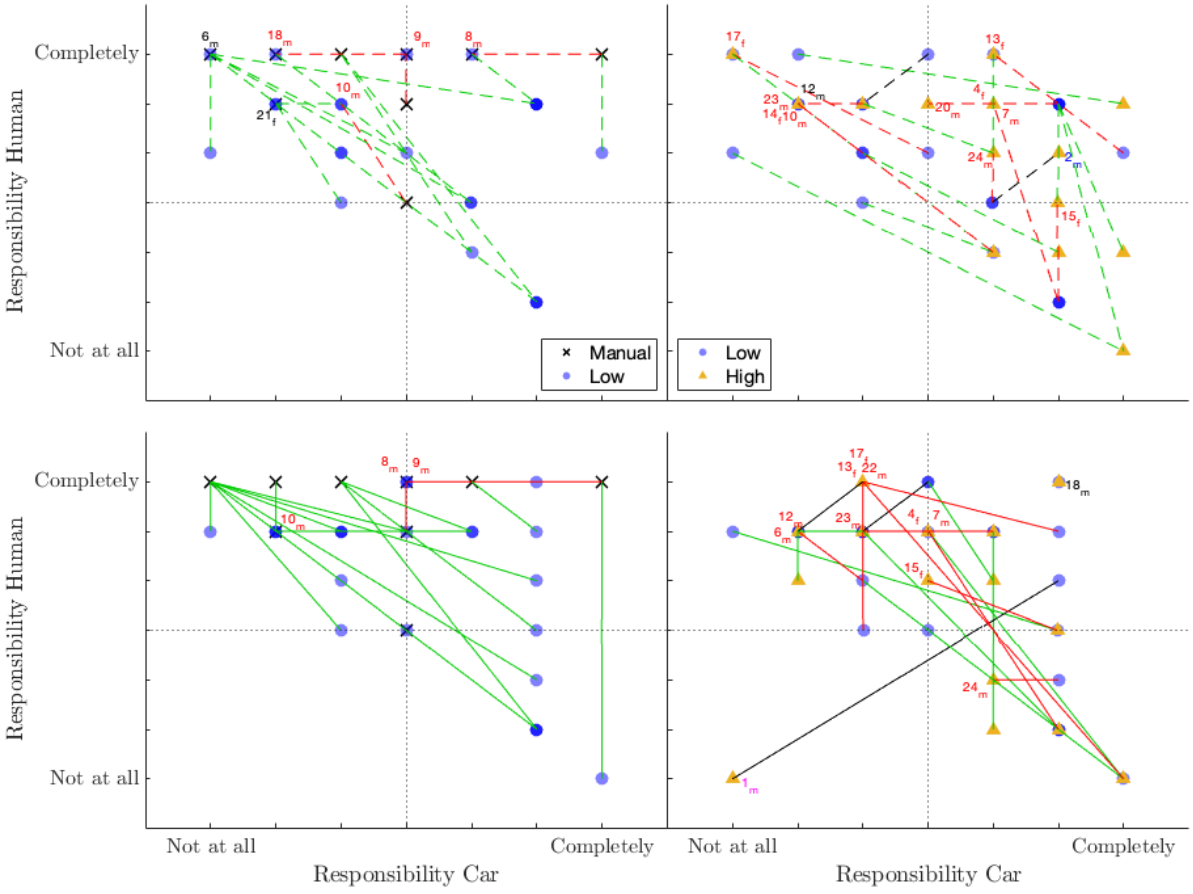

Figure G.15: Within subject responsibility shift per participant for positive outcomes. Unexpected shifts are labeled with their corresponding participant id. Green indicates a shift toward the car and away from the human, red indicates an opposite shift. The black lines indicate a shift either towards both car and human (blue label id) or away from both car and human (pink label id). No shift in responsibility is indicated by a black label id. From top to bottom: bad robot action (dashed), good robot action (line)

Responsibility shift for negative outcomes

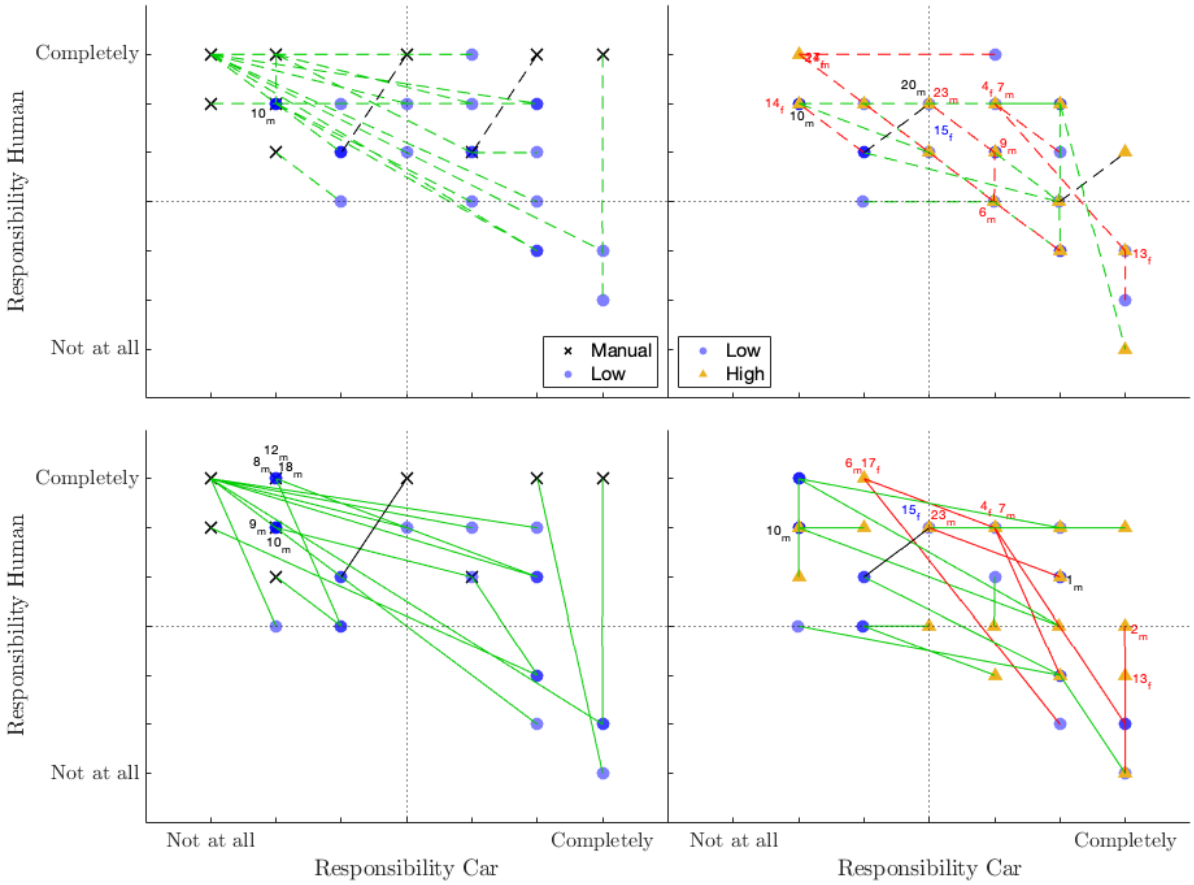

Figure G.16: Within subject responsibility shift per participant for negative outcomes. Unexpected shifts are labeled with their corresponding participant id. Green indicates a shift toward the car and away from the human, red indicates an opposite shift. The black lines indicate a shift either towards both car and human (blue label id) or away from both car and human (pink label id). No shift in responsibility is indicated by a black label id. From top to bottom: bad robot action (dashed), good robot action (line).