Intelligent Human Systems Integration (IHSI), Vol. 200, 2026, 713–723
https://doi.org/10.54941/ahfe1007130

Interactive Driving Turing Test With Think-Aloud Protocol: How Realistic are Behavioural Driver Models?

Tianyu Tang[1], Tobias Zillmann[1], Johan Olstam[2,3], Christer Ahlström[2,4], Fredrik Johansson[2], Wouter Schakel[5], Klaus Bengler[1]

[1]Technical University of Munich, Chair of Ergonomics, Munich, 85748, Germany
[2]Swedish National Road and Transport Research Institute (VTI), Linköping, 581 95, Sweden
[3]Linköping University, Department of Science and Technology, Norrköping, 601 74, Sweden
[4]Linköping University, Department of Biomedical Engineering, Linköping, 581 85, Sweden
[5]Delft University of Technology, Department of Transport and Planning, Delft, 2628 CN, Netherlands

ABSTRACT

Driver models are essential for virtual safety assessments of automated vehicles. This study evaluates the realism of the i4Driving model, a behavioural driver model designed to emulate human-like driving, using an interactive driving Turing test combined with a think-aloud protocol in connected driving simulators. Thirty participants interacted with either a human-controlled vehicle or the i4Driving model across three motorway scenarios and rated realism, predictability, safety, and aggressiveness. Results showed that the i4Driving model was perceived as less realistic and predictable than human drivers ($p < 0.001$), yet participants could not reliably distinguish between the two (classification accuracy = 0.673). Think-aloud analysis revealed specific shortcomings, such as unrealistic merging and speed adaptation, alongside instances of naturalistic behaviour. These findings highlight the need for improvements in tactical decision-making and demonstrate the value of combining subjective ratings with qualitative insights for refining driver models.

Keywords: Driving turing test, Driving behaviour, Human-likeness, Subjective rating, Connected driving simulators, Driver model

INTRODUCTION

The development of automated vehicles raises the question how automated vehicles should be tested and certified as competent drivers. Current safety approval concepts rely on distance-based and statistical approaches, requiring billions of miles to demonstrate performance better than human drivers (Karla et al., 2016). Such requirements have motivated the use of virtual assessment methods to speed up approval testing. A key challenge is developing credible models that capture the diversity and complexity of human driving behaviour. These heterogenous models could then be used to simulate a mixed driving environment in which the automated vehicles can be tested. The EU Horizon-funded project i4Driving addresses this by creating driver models that represent heterogeneous human driving styles.

Received September 25, 2025; Revised November 10, 2025; Accepted November 25, 2025; Available online February 1, 2026

In traffic modelling, mature methods already exist to objectively validate models through mathematical measures that compare model outputs in terms of trajectories or relevant traffic performance metrics with real world measurements. However, few studies have explored whether simulated driver behaviour is perceived as realistic from a subjective perspective. Our study addresses this gap by proposing an approach for evaluating driver behaviour models in terms of perceived human-likeness, with the hope that this evaluation can pinpoint specific development needs of the simulation model.

Inspired by Alan Turing's test (Turing, 1950), the original idea was to evaluate whether a human can distinguish a machine from a human through conversation. Previous studies have adapted the driving Turing test to assess whether humans can distinguish between human- and machine-controlled driving behaviour. Two main approaches have been commonly applied. The first approach involves the rater as an observer, who assesses the behaviour of the designated vehicle either from a third-person perspective through video recordings or from a passenger's perspective in on-road studies (Stanton et al., 2020; Bhattacharyya et al., 2022; Cascetta et al., 2022; Li et al., 2024). However, because this approach does not allow direct interaction with the vehicle, it lacks immersion and often prevents raters from perceiving substantial differences between human-driven and machine-controlled driving behaviours. The second approach treats the rater as an interactor, allowing raters to actively engage with the vehicle. Most existing studies adopting this approach used driving simulators (Peng et al., 2024; Woo et al., 2024; Zhang et al., 2022) to avoid potential safety risks. These studies found that participants could perceive subtle differences, but they relied mainly on post-experiment questionnaires without revealing the reasoning behind participants' judgments. To address this limitation, the present study integrated a think-aloud protocol, enabling participants to verbalise their observations and explain their evaluations in real time.

To this end, our study aims to evaluate how realistically the i4Driving model reproduces human driving behaviour in interactive traffic situations. By comparing participant perceptions of the model and a human driver, the goal is to identify key differences in perceived human-likeness and provide insights for improving behavioural driver models.

METHOD

Experimental Design and Driving Scenarios

We designed a driving simulator experiment using a 3 × 2 (scenario × agent type) within-subjects repeated-measures design, resulting in six experimental trials per participant. As shown in Figure 1, participants drove in a driving simulator and interacted with another agent that was either controlled by a human driver (confederate) or by the i4Driving model. However, the participants were, instead, informed that all agents were controlled by various computer models / automated driving systems (ADS), and their task was to rate and describe the behaviour of these agents. Three motorway scenarios were included, and each scenario was repeated three times, allowing participants to try different ways of interacting with the agents and observe the agent's responses:

A. In the onramp scenario, the participant drove on the motorway and the agent vehicle (pink) merged onto the motorway from an on-ramp. The agent's vehicle (pink) was set up to enter the merging area slightly ahead of the participant's vehicle (white) at a time headway of 0.5 s. The pink vehicle had 240 m of ramp distance available for merging.
B. In the catch-up scenario, the participant followed the pink agent vehicle with a 3 s headway as both approached a slower truck. The agent vehicle attempted to overtake the truck, creating an interaction in which the participant could respond to its manoeuvre, e.g. by applying social pressure.
C. In the weaving scenario, the participant and the pink vehicle needed to switch lane with each other. The participant had to change lane to take the right highway while the agent had to take the left one. The pink vehicle started 300 m before the splitpoint, requiring coordination as both vehicles' paths crossed each other.

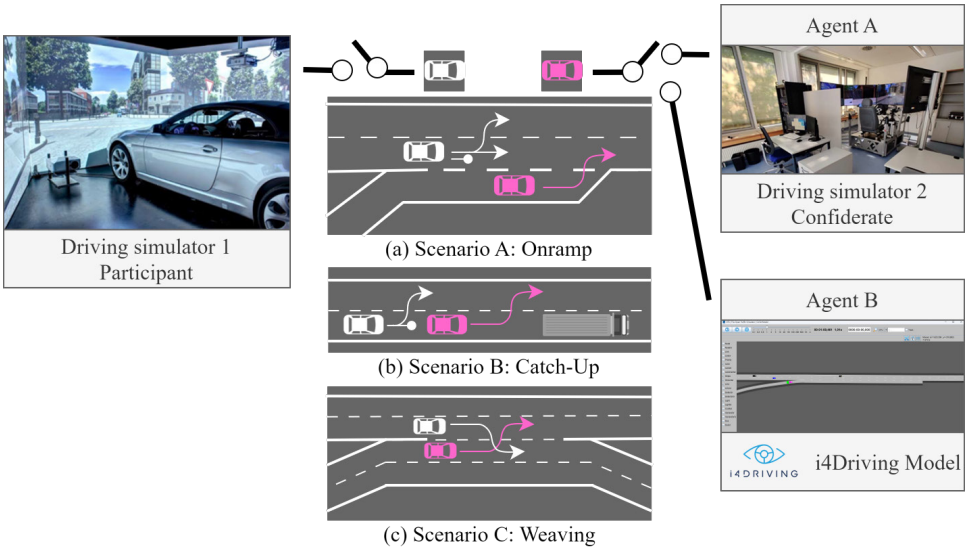

Figure 1: Scenarios and experimental setup.

Additionally, we use a think-aloud protocol to capture participants' ongoing reasoning and perceptions while driving. This method reveals not only whether they believe the agent is human-like, but also why they hold such beliefs. By analysing these verbal reports, we aim to identify the behavioural features that participants consider important for realism and to pinpoint areas where models diverge from human expectations.

Apparatus: Connected Driving Simulators

Two driving simulators were networked locally to synchronise vehicle position and driving data in real time, enabling an interactive implementation of the driving Turing test (see Figure 1). The simulators were placed in separate rooms, with an additional control centre allowing the experimenter to monitor both drivers. Each simulator provided a 180° horizontal field of view and realistic auditory feedback through an integrated sound system.

The participant simulator consisted of a BMW E64 mock-up equipped with a high-fidelity six-channel projection system (three projectors for the front and three for the rear mirrors displayed on separate canvases). The confederate simulator comprised three 55-inch Ultra-HD displays providing the forward and mirror views, with additional screens for the side mirrors. Both simulators operated within the same virtual environment using the SILAB 7.1 driving simulation software at a refresh rate of 60 Hz.

The i4Driving Model

The i4Driving model is based on the Lane change Model with Relaxation and Synchronization (LMRS) (Schakel et al., 2012). The LMRS is capable of simulating highway traffic with realistic macroscopic traffic flow phenomena (e.g. capacity, queue discharge rate) and certain mesoscopic traffic flow phenomena (e.g. distribution of traffic over lanes, average speeds on different lanes) resulting from a collision free controller that is mechanical in nature. The model is extended with the concept of social interactions (Schakel et al., 2020) in which social pressure affects both desired speed and lane change desire. This makes the model responsive to certain social cues and enhances realism in mesoscopic traffic flow phenomena (e.g., headway distribution, platoon sizes, and number of lane changes). Finally, the model is nudged out of perfect control by feeding it imperfectly perceived information such as distances and speed differences. This is achieved in the perception layer that is based on the work of van Lint et al. (2018) and Calvert et al. (2020). In this work, various tasks form an endogenously determined level of mental task demand, which may lead to situations of task oversaturation. Then, perception errors and perception delays become relevant. The perception model furthermore includes constant-speed anticipation which has been shown to be relevant in partially counteracting imperfect perception (Treiber et al., 2006). Finally, the perception model is extended by using different perception channels that roughly correlate to the perception of different areas around the driver. The tasks operate within a perception channel, resulting in task demand in each channel. The steady-state of a Markov chain is used to determine how drivers distribute their attention over the perception channels, resulting in different perception delays for different channels.

Dependent Measurements

The study primarily focused on participants' subjective perceptions of the i4Driving model. During each drive, participants followed a think-aloud protocol in which their verbal reflections on the pink vehicle's behaviour, human-likeness, and responsiveness were recorded as audio data. After every set of three scenario repetitions, participants completed a short questionnaire consisting of four seven-point Likert-scale items. They rated how predictable the pink vehicle's behavior was (1 = not at all predictable, 7 = very predictable), how realistic it appeared compared to a human driver (1 = not at all realistic, 7 = very realistic), how safe it felt relative to a competent human driver (1 = very unsafe, 7 = very safe), and how aggressive its driving style was (1 = very cautious, 7 = very aggressive).

Upon completing all drives, participants answered several open-ended questions exploring how their impressions of the pink vehicle evolved over time. In addition to subjective data, objective measures and synchronised eye-tracking data were logged for future analysis, including the longitudinal and lateral positions, speeds, and accelerations of all vehicles.

Procedure

The study was conducted at the Chair of Ergonomics of Technical University of Munich, and consisted of three parts: a pre-questionnaire, a simulator drive, and a post-questionnaire (see Figure 2). Upon arrival, participants were welcomed, provided written informed consent, and completed a demographic questionnaire including age, gender, and driver's license information. All data was collected under a pseudonymized code. Participants then received standardised instructions and completed a 5-minute practice drive to familiarise themselves with the simulator.

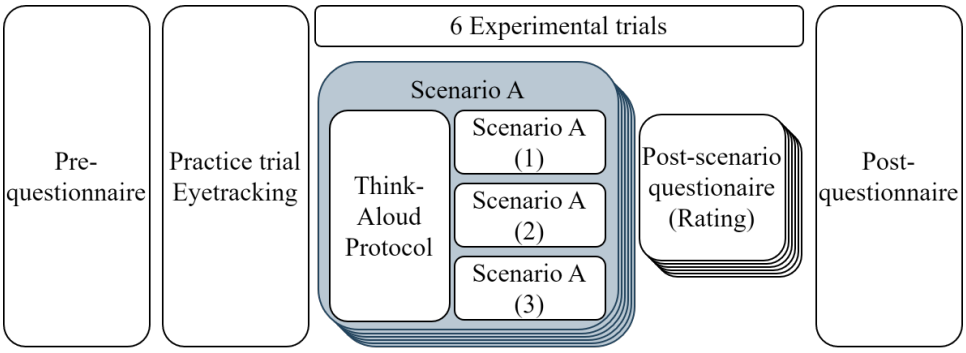

Figure 2: Procedure.

During the experiment, the participants were asked to drive as they would in real traffic and to follow traffic regulations. Each participant completed six experimental drives presented in a counterbalanced order to avoid sequence effects. In each drive, the same agent (human or the i4Driving model) controlled the pink vehicle that appeared three times. Participants were encouraged to interact with the pink vehicle in various ways to explore its behavioural responses. While driving, participants were encouraged to verbally report their observations and impressions of the pink vehicle's behaviour. They were informed that only the pink vehicle was relevant to the task, while the surrounding black vehicles were only contextual. After each drive, participants completed the rating questionnaire on a tablet mounted in the simulator and could take breaks as needed.

Participants were unaware that, in some conditions, the pink vehicle was controlled by a human driver (confederate). This information was disclosed only after all drives and questionnaires were completed, during a final debriefing session in which they were shown the second simulator room where the confederate operated the vehicle.

Sample

A valid driver's license and a minimum age of 18 were required to participate in this study. Thirty-two licensed drivers (22 male, 10 female) participated in the study. Data from two participants were excluded from the analysis due to motion sickness during the experiment, resulting in a final sample of 30 participants. The mean age was 28.43 years (SD = 11.87, Min = 18, Max = 86). On average, participants had held a driver's license for 10.09 years (SD = 11.60) and reported driving 11,328 km per year (SD = 5,350). Self-assessments on a 5-point Likert scale indicated that participants considered themselves good drivers (M = 3.66, SD = 0.75). The experiment lasted approximately 90 minutes. All participants provided informed consent, and the study was approved by the ethics committee of the Technical University of Munich (approval code: 2025-123-NM-BA).

Data Processing and Analysis

Subjective data from the rating questionnaires and think-aloud recordings were analysed to address the research questions. Statistical analyses were performed in Python and R with a significance level of $\alpha = 0.05$. Mauchly's test of sphericity and the Shapiro–Wilk test for normality both indicated assumption violations ($p < 0.05$), precluding a classical ANOVA. Given the multifactorial (3×2) design, a Friedman test was also unsuitable. Therefore, a non-parametric Aligned Rank Transform (ART) ANOVA (Wobbrock et al., 2011) was applied, with pairwise post-hoc tests adjusted using the Holm–Bonferroni method (Elkin et al., 2021).

Think-aloud data were manually transcribed. Following McLellan et al. (2003), only comments concerning the pink vehicle's behaviour were categorised to identify key qualitative themes and grouped by similarity. This analysis examined participants' real-time reasoning and perceptions, identified unrealistic or undesirable behaviours of the i4Driving model, and informed methodological considerations for future interactive driving Turing test studies.

To evaluate whether the Turing test criterion was met, responses to the realism item ("How realistic was the pink vehicle's behaviour compared to a human driver?"; 1 = not at all realistic, 7 = very realistic) were classified as i4Driving model (ratings 1–3) or human driver (ratings 5–7); ratings of 4 were excluded. Based on these classifications, accuracy, precision, recall, and F1-scores were computed as standard metrics of classification performance.

RESULT AND DISCUSSION

The results of the ART-ANOVA of the rating questionnaire, including four items, are presented in Table 1. A significant main effect of agent type was observed for realism, predictability, safety, and aggressiveness, while scenario had no significant effect on realism and predictability. No significant interaction was detected between agent type and scenario on realism, predictability and safety. As shown in Figure 3, post hoc pairwise comparisons revealed that the i4Driving model was rated significantly lower than the confederate

in realism and predictability across all scenarios. For safety, significant differences appeared in scenarios A and C, whereas for aggressiveness, the difference was significant only in scenario C.

Table 1: Results of ART-ANOVA regarding realism.

	Realism		Predictability		Safety		Aggressiveness	
	F	Pr(>F)	F	Pr(>F)	F	Pr(>F)	F	Pr(>F)
Scenario	0.188	0.829	2.673	0.072	5.173	**0.007****	0.704	0.4962
Agent	41.000	**<0.001*****	50.764	**<0.001*****	37.306	**<0.001*****	26.252	**<0.001*****
Scenario: Agent	1.434	0.242	0.963	0.384	1.540	0.218	4.070	0.0189*

The frequency of occurrence of the verbal comments from the think-aloud protocol is presented in the following tables. It should be noted that each theme could be mentioned up to a maximum of 90 times, as all 30 participants drove through each scenario three consecutive times. It is also counted how many distinct participants mentioned the comment, allowing a maximum of 30.

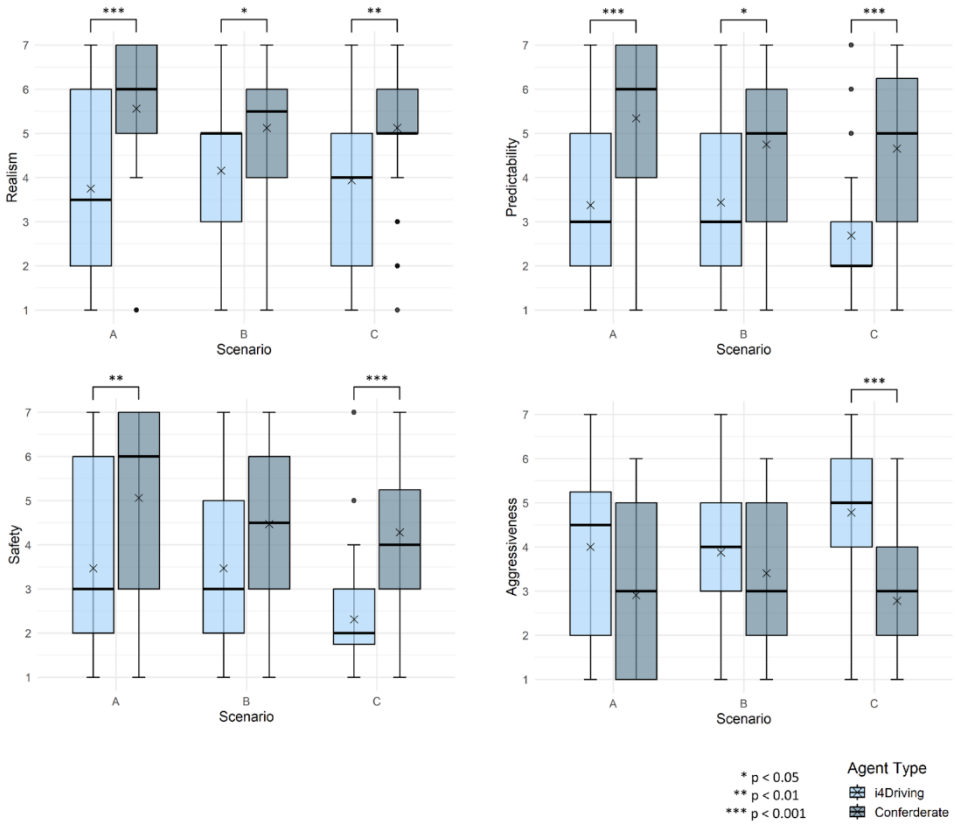

Figure 3: Distribution of rating and results of post-hoc test.

Table 2: Theme rating i4Driving model in the think-aloud protocol in scenario A.

ID	Theme	Count	Distinct
A1	The vehicle failed to merge and drove onto the shoulder or disappeared.	21	15
A2	Too high speed in the on-ramp lane.	15	12
A3	Merging was successful.	8	7
A4	The behavior appeared human-like / natural.	6	4
A5	Braking on the motorway before an exit.	5	5
A6	The vehicle had a limited field of view or failed to consider the surrounding traffic.	5	4
A7	Insufficient acceleration after entering the motorway.	5	5
A8	Driving behaviour is perceived as aggressive.	5	5

As shown in Table 2, the most frequently mentioned comment (ID = A1) concerned the situation in which the pink vehicle attempted to merge while the participant accelerated to overtake it, preventing the vehicle from cutting in. In these cases, the pink vehicle either continued driving on the shoulder or disappeared after some time at the end of the on-ramp lane. Participants found this behaviour strange and unrealistic. The reason for this issue is twofold. First, the LMRS model applies a simple 'follow leader in adjacent lane' heuristic to find and align with a gap to merge. Vehicles behind are not considered, and the merging situation is not comprehensively anticipated. Second, the disappearance is due to technical issues encountered, as the model does not account for not being able to merge. Comment A2 referred to an artefact of the experimental setup. To ensure that the two vehicles met at the designated timing, the entry speed of the pink vehicle on the on-ramp was manipulated to be higher than usual. This aspect can be improved in future experimental designs. Participants also noted that when the pink vehicle successfully completed the merging manoeuvre, its behaviour appeared more natural, as indicated by comments A3 and A4.

In Scenario B, two highlighted comments (B1 and B8) were also related to the experimental setup. Comment B8 referred to a delay in data transmission between the i4Driving model software and the driving simulator. Comment B1 resulted from the requirement that the pink vehicle had to catch up with the truck within a limited time window. This aspect clearly needs to be reconsidered in future experimental designs to provide a more natural interaction. Comment B2 can be caused by the randomness of model parameters, where i4Driving might have a desired speed only slightly above the speed of the trucks, while also not being sensitive to social pressure. As shown in Figure 3, scenario B received the highest realism ratings (M = 4.03, SD = 1.47 on a 7-point Likert scale), indicating that participants generally perceived the pink vehicles behaviour as quite realistic. Table 3 further shows that comments B4–B7 were largely subjective and related to individual driving style or model parameter settings.

Table 3: Theme rating i4Driving model in the think-aloud protocol in scenario B.

ID	Theme	Count	Distinct
B1	Strong acceleration and exceeding the speed limit.	37	26
B2	Slow acceleration during overtaking the truck	25	17
B3	The behaviour appeared human-like / natural.	14	8
B4	Lane change executed without using the turn signal.	9	8
B5	Lane change initiated very early.	9	6
B6	Exit was taken late.	8	6
B7	Reduced speed before performing a lane change for overtaking.	8	6
B8	Off-centre lane keeping.	7	5
B9	The behaviour appeared unpredictable at times.	7	7

Table 4: Theme rating i4Driving model in the think-aloud protocol in scenario C.

ID	Theme	Count	Distinct
C1	The vehicle cut me off quite sharply.	22	16
C2	We drove parallel for too long.	12	9
C3	The vehicle did not react to the turn signal.	8	8
C4	Cooperative braking behaviour during the lane change	7	6
C5	Abrupt braking during the interaction.	6	6
C6	Cooperative acceleration during the lane change.	5	3
C7	Off-centre lane keeping.	5	5
C8	Turn signal used inappropriately (too late or no usage).	5	4

In Scenario C, the most frequently mentioned comment (C1) in Table 3 was likely influenced by the choice of model parameters. As described above, we selected the default parameters representing an average driver profile, which may have appeared too aggressive to participants. Furthermore, perception delays may have caused too positive judgement of the situation by the merging vehicle. Comments C2 and C3 were probably outside the current model scope, as the present version does not include these cues within the tactical-level decision-making, cooperation with adjacent-lane vehicles, or anticipating other vehicles' lane-change intentions. It is also important to note that indicator light signals are not transmitted to the i4Driving model. Comment C5 was most likely related to the speed limitation of 80 km/h within the weaving section of the road network used in this scenario.

Table 5: Response to realistic rating from the driving turing test results.

Confusion Matrix		Who Was Controlling the Car (Truth)	
		i4Driving Model	Confederate
Perceived Realism (Response)	i4Driving Model	37 TP	40 FP
	Confederate	11 FN	68 TN
Accuracy: 0.673	Precision: 0.861	Recall: 0.630	F1 score: 0.727
TN: true negatives; TP: true positives; FN: false negatives; FP: false positives			

Both the ART-ANOVA results and the think-aloud findings revealed some differences in perceived realism between the confederate and the i4Driving model. Nevertheless, when considering the core concept of the Turing test, the quantitative evaluation of participants' classification performance (in Table 5) suggests that participants could not reliably distinguish between human and model behaviour. The overall classification accuracy was only 0.673. There were 40 false positives, meaning that participants frequently judged the confederate's behaviour as not realistic. In contrast, the model's behaviour was perceived as unrealistic in only 37 cases. The overall F1-score for participants' judgments was 0.727, suggesting moderate consistency in their evaluations.

REFLECTION AND CONCLUSION

In this connected driving simulator study, we implemented the concept of a driving Turing test and explored whether humans could recognise differences in driving behaviour between the i4Driving model and a human driver. Based on the think-aloud protocol analysis, we obtained meaningful insights that can contribute to improving both the i4Driving model and the driving Turing test methodology. On the model development side, the findings indicate that humans are highly sensitive to speed adaptation during interaction, particularly to braking behaviour on the motorway. This suggests that tactical-level decision-making should be further developed in the model. The relevance of the comments highly depends on the type of application. Although the comments are relevant for simulating surrounding vehicles in driving simulators the relevance of e.g. modelling of communication of turning indicators is less relevant in applications where all vehicles are controlled by the i4Driving model. On the experimental side, several limitations were identified. Some unexpected behaviours of the relevant vehicle were caused by the experimental setup itself, which made it difficult to interpret certain results. Future studies should develop more refined manipulation strategies to ensure that interactions between vehicles occur in a natural and consistent way, and these manipulations should be thoroughly tested in advance.

ACKNOWLEDGMENT

The authors thank Alireza Kamalidehgan and Arzu Alsan for their support. This work was funded by the European Union's i4Driving project (No. 101076165).

REFERENCES

Bhattacharyya, R., Jung, S., Kruse, L.A., Senanayake, R. and Kochenderfer, M.J. (2022) 'A hybrid rule-based and data-driven approach to driver modeling through particle filtering', *IEEE Transactions on Intelligent Transportation Systems*, 23(8), 13055–13068.

Calvert, S.C., Schakel, W.J. and Van Lint, J.W.C. (2020) 'A generic multi-scale framework for microscopic traffic simulation part II – anticipation reliance as compensation mechanism for potential task overload', *Transportation Research Part B: Methodological*, 140, 42–63.

Cascetta, E., Cartenì, A. and Di Francesco, L. (2022) 'Do autonomous vehicles drive like humans? A Turing approach and an application to SAE automation Level 2 cars', *Transportation Research Part C: Emerging Technologies*.

Elkin, L.A., Kay, M., Higgins, J.J. and Wobbrock, J.O. (2021) 'An aligned rank transform procedure for multifactor contrast tests', *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST'21)*, 754–768.

Kalra, N. and Paddock, S.M. (2016) 'Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?', *Transportation Research Part A: Policy and Practice*.

McLellan, E., MacQueen, K.M. and Neidig, J.L. (2003) 'Beyond the qualitative interview: Data preparation and transcription', *Field Methods*, 15(1), 63–84.

Peng, C., Merat, N., Romano, R., Hajiseyedjavadi, F., Paschalidis, E., Wei, C., Radhakrishnan, V., Solernou, A., Forster, D. and Boer, E. (2024) 'Drivers' evaluation of different automated driving styles: Is it both comfortable and natural?', *Human Factors*, 66(3), 787–806.

Schakel, W., Knoop, V., Keyvan-Ekbatani, M. and Van Lint, H. (2023) 'Social interactions on multi-lane motorways: Towar*ds a theory of impacts', Engine*ering.

Schakel, W.J., Knoop, V.L. and Van Arem, B. (2012) 'Integrated lane change model with relaxation and synchronization', Transportation Research Record, 231*6(1), 47–57*.

Stanton, N.A., Eriksson, A., Banks, V.A. and Hancock, P.A. (2020) 'Turing in the driver's seat: Can people distinguish between automated and manually driven vehicles?'.

Treiber, M., Kesting, A. and Helbing, D. (2006) 'Delays, inaccuracies and anticipation in microscopic traffic models', *Physica A: Statistical Mechanics and its Applications*, 360(1), 71–88.

Turing, A.M. (1950) 'Computing machinery and intelligence', *Mind*, 59(236), 433–460.

van Lint, J.W.C. and Calvert, S.C. (2018) 'A generic multi-level framework for microscopic traffic simulation—theory and an example case in modelling driver distraction', *Transportation Research Part B: Methodological*, 117, 63–86.

Wobbrock, J.O., Findlater, L., Gergle, D. and Higgins, J.J. (2011) 'The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 143–146.

Woo, J. and Ahn, C. (2024) 'Towards human-like autonomous vehicles: A qualitative evaluation and design perspective', *IEEE Intelligent Vehicles Symposium (IV)*, 403–408.

Zhang, Y., Hang, P., Huang, C. and Lv, C. (2022) 'Human-like interactive behavior generation for autonomous vehicles: A Bayesian game-theoretic approach with Turing test', *Advanced Intelligent Systems*, 4(5).