



Human Performance in Critical Scenarios as a Benchmark for Highly Automated Vehicles

Laura Quante¹ · Meng Zhang¹ · Katharina Preuk¹ · Caroline Schießl¹

Received: 6 November 2020 / Accepted: 28 May 2021
© The Author(s) 2021

Abstract

Before highly automated vehicles (HAVs) become part of everyday traffic, their safety has to be proven. The use of human performance as a benchmark represents a promising approach, but appropriate methods to quantify and compare human and HAV performance are rare. By adapting the method of constant stimuli, a scenario-based approach to quantify the limit of (human) performance is developed. The method is applied to a driving simulator study, in which participants are repeatedly confronted with a cut-in manoeuvre on a highway. By systematically manipulating the criticality of the manoeuvre in terms of time to collision, humans' collision avoidance performance is measured. The limit of human performance is then identified by means of logistic regression. The calculated regression curve and its inflection point can be used for direct comparison of human and HAV performance. Accordingly, the presented approach represents one means by which HAVs' safety performance could be proven.

Keywords Highly automated vehicles · Automated driving · Proof of safety · Human performance · Driving performance

Abbreviations

ANOVA	Analysis of variance
FOT	Field operational test
HAV	Highly automated vehicle
NDS	Naturalistic driving study
TTC	Time to collision

1 Introduction

Autonomous driving is associated with many potential advantages, e.g., increased traffic efficiency and safety benefits [1–3]. Highly automated vehicles (HAVs; SAE level 3 or higher [4]) allow the human driver to attend to a task other than the driving task. But before HAVs become part of everyday traffic, their safety performance—especially when confronted with our constantly changing environment—has to be tested and verified [5, 6]. Following the request that HAVs should only be released if they are statistically safer than human drivers made by the German Federal Ministry of

Transport and Digital Infrastructure [7], human performance should be used as a benchmark for HAV performance (at least in Germany). However, appropriate methods to quantify and compare human and HAV performance are rare. To facilitate HAVs' safety assessment, this work presents a scenario-based method to quantify the limit of human performance, which was developed as part of the project *PEGASUS* [6].

This article is structured as follows: in Sect. 2 human performance is defined, followed by a short summary of existing approaches to measure human performance (2.1 and 2.2) and an introduction of the approach presented in this article (2.3). The driving simulator study is described in Sect. 3 and human performance is quantified in Sect. 4. Section 5 discusses the results and the presented approach before drawing conclusions in Sect. 6.

2 Measuring Human Performance

In general, human performance refers to the potential of a person to successfully perform a task. In case of driving, the task-capability interface model [8] considers human performance as a function of the driver's capability and the demands of the driving task. If capability excels task demands, the driver controls the situation. If not, control is

✉ Laura Quante
Laura.Quante@dlr.de

¹ Institute of Transportation Systems, German Aerospace Center (DLR e.V.), Lilienthalplatz 7, 38108 Braunschweig, Germany

lost and an accident is highly possible. The driver's capability results from skills and abilities, and is influenced by human factors (e.g., emotions, inattention or stress). Task demands are composed of environmental factors (e.g., weather and road conditions), the vehicle and the driving task itself.

The driving task consists of navigation, guidance and stabilisation according to the three-level model of the vehicle driving task [9, 10]. Navigation refers to the planning of an appropriate route, whereas guidance involves adapting driving behaviour to the course of the road, traffic conditions and traffic rules. Stabilisation requires drivers to keep the vehicle in a steady state by, for example, braking and steering. The different subtasks demand different competences. Depending on the task, drivers regulate their behaviour based on knowledge, rules or skills [11]. For example, to stabilise a vehicle, experienced drivers will mostly regulate their driving behaviour based on highly automated skills. If driving a vehicle for the first time, stabilisation might be more regulated based on knowledge or rules. In summary, driving covers various subtasks which require specific abilities (see Fig. 1). To measure human performance, task demands have to be defined and human capability has to be determined.

Human performance in traffic can be studied by analysing accident statistics or driving behaviour in real or simulated traffic environments. In previous research, measures like collision probability, driving errors and response times have been used to describe and quantify human safety performance [12–17].

2.1 Data Acquisition Methods

Accident statistics (e.g., German In-Depth Accident Study [18]) provide information about the number, severity and circumstances of traffic accidents and have already been used to compare human and HAV performance [13, 14]. It was shown that HAVs would have to drive hundreds of millions of miles without an accident to prove their superiority in terms of safety—a requirement that seems impossible prior to release [14].

If accident data are evaluated with regard to the presented task-capability-interface model [8], it becomes apparent that these data represent cases in which task demands exceeded human capability. In order to reduce the amount of data needed, it might be useful to first find the limit of human performance, i.e. situations in which task demands slightly exceed human capability, and to then analyse HAVs' performance in these situations.

As previously mentioned, driving performance can also be studied in real traffic or in simulated environments. For example, naturalistic driving studies (NDS) and field

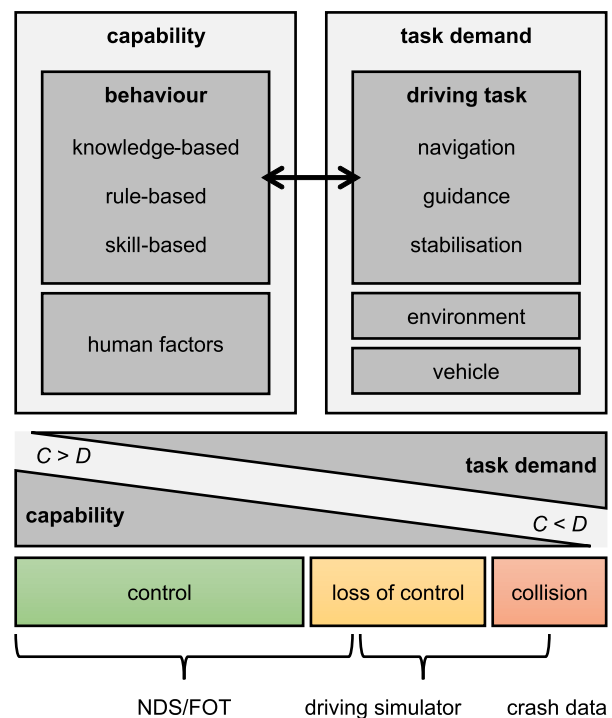


Fig. 1 Simplified model of human performance in the project PEGASUS (top) and relationship between capability (C) and task demands (D) across different methods (bottom)

operational tests (FOT) represent field studies, in which everyday driving behaviour is observed over a long period of time (usually several months, e.g. Refs. [19–22]). Consequently, FOT and NDS data contain a broad range of driving behaviour in various traffic situations and thus have high external validity. Accidents and near misses, however, are very rare. Accordingly, FOT and NDS data mainly include cases in which capability exceeds task demands.

Simulated environments, which are used in driving simulators or virtual reality studies, on the other hand, allow to precisely define and control every aspect within a traffic situation and to repeatedly assess a specific scenario while driving behaviour is observed. Driving simulator studies can thus be used to analyse the underlying causal relationships of accidents and also to systematically approach and assess the limit of human performance by creating an environment in which task demands are systematically increased. The use of the driving simulator, therefore, serves to close the gap between the recording and description of human performance in uncritical and critical scenarios to the point of accidents ([23]; see Fig. 1 bottom).

2.2 Human Performance Measures

Besides choosing an appropriate method, performance has to be adequately operationalised. Because in many critical traffic situations it is essential to react fast to prevent an accident [10, 24], response time, i.e. the time from stimulus onset to response initiation, is used as a measure of performance [15]. Human response time is influenced by several factors such as driver expectation, age, cognitive load, and urgency [25]. Furthermore, Pöwelleit and Vollrath [15] showed that response time is highly dependent on situational factors like road type, driving speed, and stimulus saliency. If the stimulus or the response is not clearly identifiable, response time cannot be calculated or only imprecisely. This might be the case if anticipatory driving behaviour is shown. Response time, therefore, might not always be a sufficient performance indicator and, for example, the type of behaviour, the strength of the response or the outcome of a behaviour should be considered as well [15].

Another measure of human performance could be driving errors. Different types of human driving errors have been defined [26–28]. Graab et al. [29], for example, differentiated between information (access, uptake, processing), goal and action errors, and showed that human drivers mainly commit information related errors, with errors of information uptake being most prominent (see also [13, 30]). Common causes of information uptake-related errors include inattention, distraction, drug consumption or fatigue [29, 31], i.e., purely human factors. It has been shown that in nowadays traffic the majority of accidents are caused by human errors, whereas technical errors and environmental influences (e.g., road or weather conditions) only count for a small percentage of accidents [17]. HAVs might therefore significantly increase traffic safety [2], but could also show other or new types of errors. What is more important: not every type of error might be equally critical when driving and not every error, even errors of the same type, might result in an accident (and might therefore go unnoticed). Consequently, a certain error rate does not necessarily correspond to the same level of performance.

Since both response times and driving errors might not always be clearly linked to performance, collision probability, as the ultimate and decisive outcome of a situation, is used as a measure of human performance in this study. In the following paragraph, it will be described how the limit of human performance could be identified based on collision probability in a driving simulator study.

2.3 The Present Approach

The PEGASUS project focused on the highway chauffeur. The highway chauffeur, as defined in the PEGASUS project, is a conditional automated driving function (SAE level 3 [4])

which performs the longitudinal and lateral driving task on highways within a speed range of 0 to 130 km/h (including lane changes, stop and go traffic jams, and emergency braking/collision avoidance) and has not to be continuously monitored by the human driver. Because HAVs could be exposed to similar traffic conditions as human drivers, in a first step safety-critical scenarios in nowadays traffic were identified by analysing accident data. Results were then complemented by analysing NDS and FOT data to define the upper and lower performance limits (i.e. cases in which capability clearly exceeds vs. fails task demands).

Accident data was searched for accidents that might be prevented by the highway chauffeur and revealed accidents due to lane-change manoeuvres and rear-end collisions to be most relevant. Especially cut-in manoeuvres might be critical for human drivers and HAVs if the lane change is unexpected or abrupt. In line with that, most lane change accidents were characterized by a time to collision (TTC; [32]) of about two seconds or less. In contrast, very few cases with a TTC of less than two seconds were found in the NDS/FOT data and none of them ended in an accident [23, 33].

Based on the findings of the accident and NDS/FOT data analyses, a driving simulator study was conducted in the second step to find the limit of human performance when confronted with a cut-in manoeuvre. To measure human performance in this scenario, task demands were manipulated by systematically varying the criticality of the manoeuvre. To assess the limit of human performance, the method of constant stimuli was adapted [34]. The method of constant stimuli was implemented by many classical psychophysical experiments to determine thresholds of human sensation (e.g., [35, 36]) and represents a stimulus–response model: Participants are repeatedly exposed to stimuli, such as tones, that vary constantly in their intensity, such as volume. Participants are asked to respond when they have perceived the stimulus. The threshold is then defined as the volume at which the probability of having heard it is 50%. In terms of the present driving simulator experiment, the stimulus was the cut-in manoeuvre. Stimulus “intensity” was represented by the manoeuvre’s criticality, which was operationalised by TTC. Drivers responded by avoiding a collision or not. In line with the method of constant stimuli, the “threshold” of human performance was then defined by the criticality, i.e., TTC, at which 50% of drivers failed to avoid a collision. By letting HAVs drive the identical set of scenarios, either within simulation, on a proving ground or during real-world drives, their performance can be estimated in the same manner, so that, finally, the limits of HAV and human performance can be compared.

3 Driving Simulator Study

3.1 Participants

Fifty-two volunteers (30 males, 22 females) aged from 20 to 76 years (*mean* (M) = 44.36 years, *standard deviation* (SD) = 19.1 years) participated in the study. Participants were recruited from the participant database of the Institute of Transportation Systems at the German Aerospace Center. All participants were required to hold a valid driving license. The majority of participants reported driving daily ($n=30$, 57.7%), whereas the minority reported to drive on workdays ($n=5$, 9.6%), once or twice a week ($n=5$, 9.6%), once or twice a month ($n=7$, 13.5%) or less than once a month ($n=5$, 9.6%). The annual mileage was reported to be low (less than 9,000 km/year) by 53.8% ($n=28$) of the participants, 28.8% ($n=15$) reported driving between 9,000 and 20,000 km/year, 15.4% ($n=8$) between 20,000 and 30,000 km/year, and only 1.9% ($n=1$) reported driving more than 30,000 km/year. With regard to age, gender and annual mileage, the sample roughly corresponds to the population of German car drivers. The study protocol was conducted in accordance with the ethical standards of the Declaration of Helsinki.

3.2 Simulator Set-Up

The study was accomplished in the dynamic driving simulator of the Institute of Transportation Systems at the German Aerospace Center in Braunschweig, Germany, in combination with the SimCar. The environment was simulated using a 270°-back-projection visualization with a resolution of 1400 × 2100 for each 30° projection angle (18 projectors in total). The SimCar is a converted passenger car with a real operational interface (including steering wheel, gas and brake pedals). Instrument panel and side mirrors were replaced by displays. To enable observation of rear traffic, a LCD-display was placed on the back seats, which was visible through the rear mirror. The cars' speakers transmitted the engine and traffic sounds. Motion simulation was inactive in the present study.

3.3 Driving Scenario

The scenario was a cut-in manoeuvre on a straight two-lane highway with a speed limit of 130 km/h (lane width = 3.56 m; see Fig. 2 for a simplified visualization of the scenario). In every trial, participants' ego vehicle (blue vehicle in Fig. 2) was placed on the left lane (ego lane) with a speed of 130 km/h ("flying start"). On the right lane, a platoon of passenger cars and trucks was driving with a constant speed of 80 km/h and distances of 10 to 50 m between

vehicles. In every trial, one of the platoon's passenger cars (target vehicle; orange vehicle in Fig. 2) abruptly changed from the right to the ego lane and cut in in front of the ego vehicle. The dynamics of the target vehicle were identical in every trial (averaged lateral speed = 2.31 m/s). To prevent participants from anticipating the target vehicle, it was randomly placed in one of five positions within the platoon. Depending on where the target vehicle was placed in the platoon, the lane change occurred approximately 20, 33, 45, 52 or 72 s (435, 732, 987, 1153, 1601 m) after the beginning of a trial. Every target vehicle was closely (approx. 10 m) followed by a truck (but not after every truck was a target vehicle) and was therefore only visible when the distance between ego and target vehicle was short enough. A trial ended, when a collision occurred or ten seconds after the target vehicle started the lane change. In both cases, the screen turned black for a short period of time before the next trial started. Trial duration ranged from 18 to 125 s ($M=53$ s). For scenario design, operation and visualization, VIRES Virtual Test Drive (VIRES Simulationstechnologie GmbH, Germany) was used.

3.4 Experimental Design

The experimental design contained one independent variable, namely the criticality of the cut-in manoeuvre. It was varied as a within-subjects factor by means of the time to collision (TTC). TTC represents the time it will take for two vehicles to collide if they continue driving on the same path with the same speed [32, 37, 38]. In the present experiment, the TTC was defined as the time distance between the ego and the target vehicle at that moment, when the target vehicle was just in the middle between the two lanes (see Fig. 2). Assuming a speed difference of 50 km/h between ego and target vehicle, TTC was manipulated at six levels: 0.5, 0.7, 0.9, 1.1, 1.3, and 1.5 s. Each participant underwent ten trials of every TTC level, leading to a total of 60 cut-in manoeuvres. The 60 experimental trials were presented in randomized order. The selection of levels and the randomization of trials were in accordance with the method of constant stimuli. TTC values were based on the findings of the FOT/NDS data analyses, which suggested that a TTC of 1.7 s is the bottom line of the criticality in everyday driving behaviour [23]. By randomizing trial order and target vehicle location, the target vehicle's lane change was expected but not entirely predictable.

3.5 Experimental Procedure

After arrival, participants were informed about the experiment (including the critical encounters) and provided written informed consent to take part in the study. Additionally, they were asked to fill out a questionnaire about personal

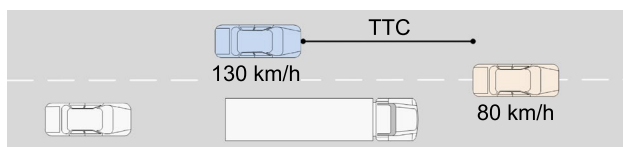


Fig. 2 The cut-in scenario: a target vehicle (orange) abruptly changed from the left to the right lane and cut in in front of the ego vehicle (blue)

information (e.g., gender, age, and driving experience). Before the experiment, every participant drove a training session for at least five minutes to familiarize themselves with the driving simulator (participants were not supposed to shift gears; see Fig. 3). In the actual experiment, participants were told to stay on the left lane and keep a constant speed of 130 km/h until they face a critical event. In this case, they were asked to avoid an accident.

3.6 Measurements

Driving simulator data, which included kinematic driving parameters of the ego vehicle and environment parameters, were recorded with a sampling rate of 25 Hz. The following variables were calculated for further analysis: (1) Due to the variance in ego speed, the actual TTC might slightly deviate from the aspired TTC. The *Actual TTC* was therefore used for further analysis and is defined as the actual time distance (in seconds) between the ego vehicle and the target vehicle at that moment, when the target vehicle is just in the middle between the two lanes. (2) The binary variable *Collision* indicates if there was a collision in a trial or not (0 = no collision, 1 = collision; dummy coded). (3) The *Collision probability* provides the ratio of the number of trials with collision to the number of all experimental trials. (4) *Response time* is defined as the time (in milliseconds) between the onset of the event (here defined as the moment, when the target vehicle touches the road marking between the two lanes) and the first response of the participant (here defined as the earliest point in time within five seconds before and after event onset when the normalized brake pedal position

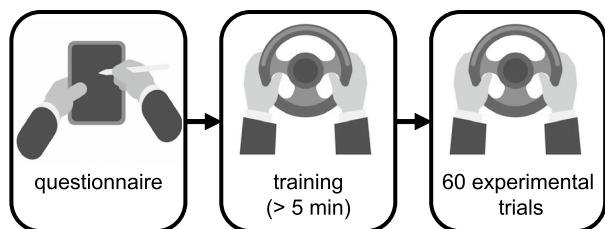


Fig. 3 Experimental procedure

is greater than zero; range = 0–1). The event onset was chosen because the target vehicle was definitely visible at this point in time and a lane change was inevitable. It has to be noted that the target vehicle's lane change might have been visible before the defined event onset (please refer to Supplement A for alternative response times). However, response time was primarily calculated to get a rough estimation of whether participants were attentive.

3.7 Analysis

As described in the introduction, the present study adapted the method of constant stimuli. According to this method, the discriminant threshold corresponds to the stimulus intensity (continuous variable) for which the response (binary variable) is random (i.e., both response types have a probability of 50%). Binary logistic regression is a means to indicate the relationship between a binary variable and a continuous variable in a sample of data and is described by

$$p(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}} \quad (1)$$

In this study, binary logistic regression was therefore implemented to relate the binary variable *Collision* (y) to the continuous variable *Actual TTC* (x_1), $p(y)$ represents the collision probability. The inflection point of the logistic regression, which was supposed to correspond to a collision probability of 50%, was (in line with the method of constant stimuli) defined as the threshold of human drivers' collision avoidance performance. Iteratively reweighted least squares were used to estimate coefficients β_0 and β_1 of the regression models. The analysis was done in RStudio (version 1.2.5019) using the function *glm* of the package *stats* (version 3.6.1).

4 Limit(s) of Human Performance

4.1 Data Pre-processing

In the first step, trials with negative response times (i.e., response time ≤ 0 ms; $N = 24$) were excluded from further analysis. The average response time of the remaining trials ($N = 3096$) was 270.53 ms (SD = 139.37 ms, range = 40–1480 ms).

In the second step, it was verified that the experimental manipulation of criticality was successful. Table 1 shows that the more critical the cut-in manoeuvre (i.e., the shorter the TTC), the more likely a collision. This finding was supported by a chi-squared test [$\chi^2(5, N = 3096) = 2113.40, p < 0.001$].

Table 1 Absolute number and percentage of trials with and without collision across the aspired TTC levels (s)

TTC	0.5	0.7	0.9	1.1	1.3	1.5
NCo	3 0.6%	97 19.3%	368 72.7%	483 94.2%	503 98.6%	546 98.6%
Co	506 99.4%	407 80.7%	138 27.3%	30 5.8%	7 1.4%	8 1.4%

NCo, No collision; Co, Collision

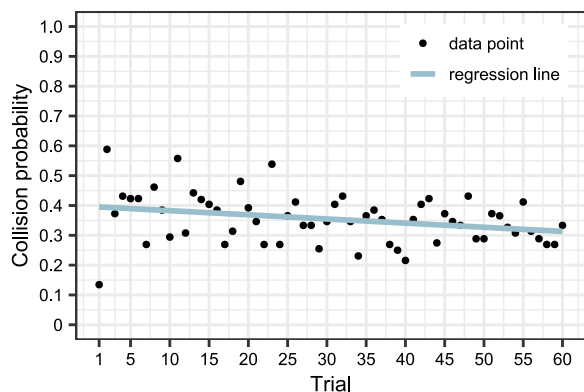


Fig. 4 Relationship between trial number and collision probability. The regression line is displayed in blue

In the third step, it was tested whether collision probability differed across the experiment or with respect to age or driving experience. In order to test the effect of time on collision probability, which could be caused by, for example, fatigue or learning effects, the correlation between trial number and collision probability was tested. A *t*-test indicated that the collision probability decreased over time [$t(58) = -2.32, p = 0.024, r = -0.29$] (see Fig. 4). To assess possible age effects, participants were divided into three age groups: 18–24 years ($n = 11$), 25–44 years ($n = 16$), and above 45 years ($n = 25$). An univariate analysis of variance (ANOVA) showed that age had no significant effect on collision probability [$F(2,49) = 0.90, p = 0.415, \eta^2 = 0.04$] (see Fig. 5). Driving experience was measured by annual mileage: < 9,000 km/year ($n = 28$), 9,000–20,000 km/year ($n = 15$), and > 20,000 km/year ($n = 9$). The corresponding univariate ANOVA revealed that collision probability did not differ significantly between the three levels of driving experience [$F(2,49) = 1.49, p = 0.236, \eta^2 = 0.06$] (see Fig. 6). Due to the absence of differences in collision probability with regard to age and driving experience, all trials were considered for regression analysis.

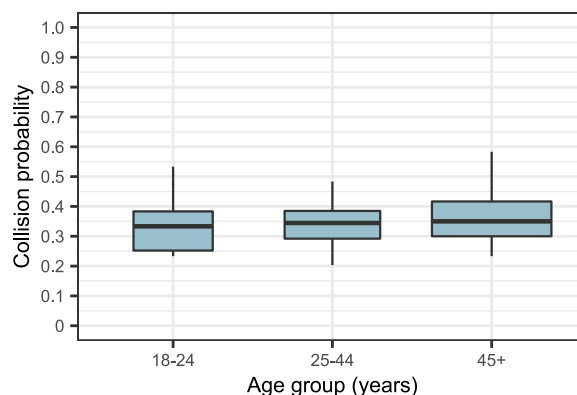


Fig. 5 Distributions of collision probability within the three age groups

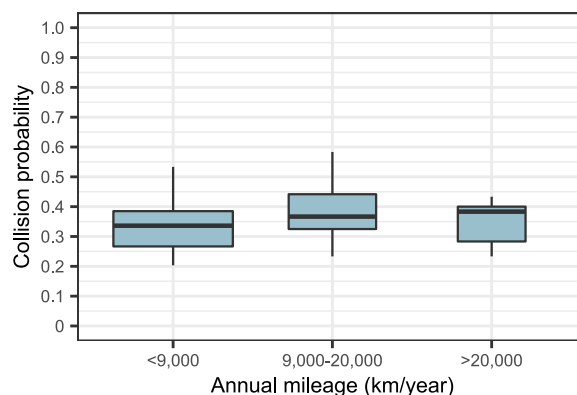


Fig. 6 Boxplots show the distributions of collision probability depending on annual mileage

4.2 Logistic Regression

To compare different performance groups with respect to collision probability, the top 10% ($n = 7$), the median 50% ($n = 26$), and the last 10% ($n = 6$) of the sample were grouped. Mean collision probability and standard error of each group and the full sample are illustrated in Fig. 7.

Logistic regression was calculated for the full sample and for every performance group to determine the relationship between collision probability and the actual TTC.

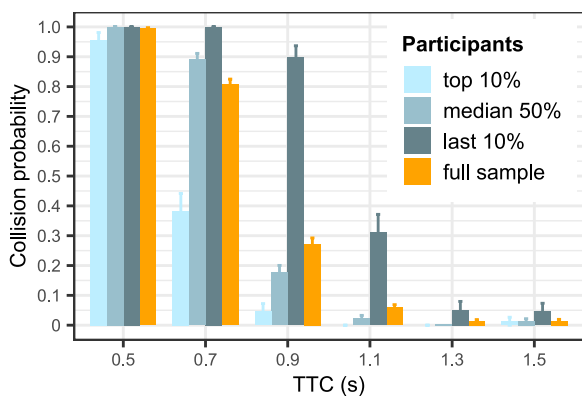


Fig. 7 Mean collision probability and standard error of three performance groups and the full sample across the aspired TTC levels

Regression curves and inflection points of all four regression models are displayed in Fig. 8.

It was revealed that the actual TTC was a significant factor in all four regression models (see Table 2). The inflection point, i.e., the actual TTC at a collision probability of 50%, ranged from 0.72 to 1.09 depending on the group of participants.

5 Discussion

The release of HAVs requires proof of safety. Exactly how HAVs’ safety can be proven, however, is still under discussion. The use of human performance as a benchmark seems reasonable, but needs appropriate methods. By adapting the method of constant stimuli, a scenario-based approach to quantify the limit of human performance was developed

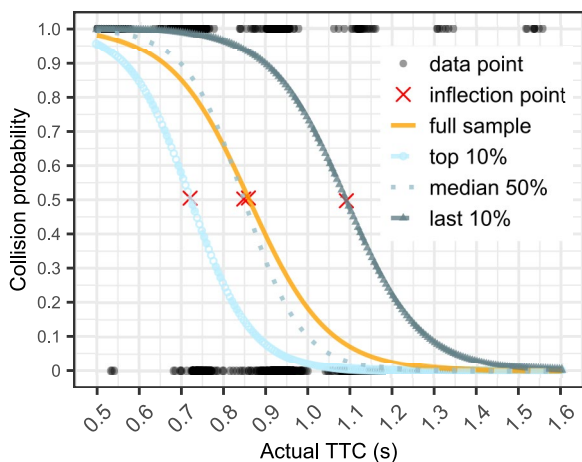


Fig. 8 The relationship of actual TTC and collision probability in the form of logistic regression curves. Inflection points are displayed in red, original data points are displayed in black

Table 2 Results of logistic regression of full sample and three performance groups

Model	<i>n</i>	<i>C</i>	<i>z</i>	<i>p</i>	<i>IP</i>
Full sample	52	-10.74	-26.57	<.001	0.86
Top 10%	7	-13.85	-8.14	<.001	0.72
Median 50%	26	-14.44	-17.00	<.001	0.85
Last 10%	6	-11.35	-9.03	<.001	1.09

C, Coefficient; *IP*, Inflection point

and humans’ collision avoidance performance in a cut-in manoeuvre was identified. The first part of the discussion will focus on the method itself, followed by a discussion of the actual results.

As has been described in the introduction, current approaches to determine HAVs’ safety performance mainly rely on an immense amount of data obtained in real traffic [13, 14]. A scenario-based approach, in contrast, offers the advantage of reducing the amount of data needed by a systematic and structured selection of test cases [6]. These test cases can then be implemented in real traffic, on proving grounds or in simulation. Simulation and, in case of human performance, driving simulator studies offer the advantage of controlling all aspects of a driving scenario without putting anyone at risk. In addition, different participants/HAVs can repeatedly go through exactly the same scenario, which is optimal for comparison of HAV and human performance. It should yet be noted that, in contrast to simulation, only a few scenarios can be tested within one driving simulator study. This requires careful selection of safety-critical factors and scenarios. Severity and exposure, following e.g. ISO 26262, and the functional scope of the system to be tested might guide the selection process. Finally, the efficiency of this method will heavily depend on whether results for a specific scenario are generalizable to a broad range of scenarios. As this study focused on one very specific scenario, this has to be studied in further experiments. Another important aspect concerns the technical configuration: if simulation or driving simulators are used for comparison of human and HAV performance, it is most important to have realistic input signals and driving dynamics to obtain a realistic driving performance (e.g., collision probability). If the observed driving behaviour does not generalize to real driving, the comparison of human and HAV might be invalid.

To determine the limit of (human) performance, adequate methods are needed. This study borrowed from psychophysics by adapting the method of constant stimuli, which is originally used for sensory threshold detection [34]. This method (in combination with logistic regression) seems appealing because it allows quantifying performance by one value, which can then be directly used for comparison. However, one could question whether a collision probability

of 50% represents the limit of human performance. To avoid specifying a definite limit of performance (in terms of a specific collision probability), it would be possible to compare the full regression curve and not just one value. The method of constant stimuli is limited so far as it needs (continuous) stimulus “intensity” and a binary response. Additionally, it should be noted that the presented method requires stimulus repetition and therefore allows participants to prepare, at least partially, for the upcoming stimulus. It thus identifies the upper limit of human performance, i.e., what humans are capable of under optimal circumstances. This is seen as an advantage since it sets higher expectations for HAVs.

Even though the focus is on developing a method, the aim of this study is also to quantify the upper limit of human performance. For this it was important that participants were not distracted but focused on the task. Their level of alertness was rated based on response times. Following Green [25], response times for expected brake signals take about 0.7 to 0.75 s, whereas brake reactions to unexpected events follow 1.25 to 1.5 s after the signal. In critical situations response times might fall below 1 s for unexpected events [39]. The average response time in this study (approx. 0.27 s) is thus shorter than would be expected. As described in the method section, the lane change might have been visible before the selected stimulus onset (the moment when the target vehicle touched the line between the two lanes). The time between the beginning of the lane change and the defined onset was approximately 0.5 s. The actual average response time should, accordingly, lie between 0.27 and 0.77 s (see Supplement A), which is still far less than one second. This indicates that participants expected the cut-in manoeuvre and therefore concentrated on the task.

Human performance, in this study, corresponds to collision avoidance. It describes the final output of an encounter, but does not provide any information about the strategy used to avoid the collision or how “well” a collision was avoided. The critical driving task in the cut-in scenario was stabilisation [9]. Because of the instruction (keep lane and speed) and the suddenness of the lane changes, participants had to rely on highly automated skills, i.e., stimulus–response–automatisms, to avoid a collision [11]. Further studies should address whether the presented method can measure other facets of (human) performance.

Although collision probability did not differ significantly with respect to age and experience, it decreased over time and varied between participants. The difference between the inflection points, i.e., the limit of performance, of the top and last 10% was 0.37 s, which might make a difference for HAVs’ performance rating. This raises the question of which group of participants should serve as the reference for HAV performance. In the present study, the sample was aligned with the total population of German car drivers and then divided into performance groups. In the future, it could be

considered to select a specific performance group from the outset, e.g. novice or professional drivers. In addition, one might only use certain trials to decrease variance in performance caused by fatigue or learning effects. The selection of trials and reference group will be a question of the desired safety performance level of HAVs.

The results show that collision probability was only close to zero for a TTC of 1.3 s and 1.5 s, respectively, which is in line with the suggestion to rate encounters with a TTC of 1.5 s or less as safety critical [40]. However, since the participants of this study were prepared to encounter critical situations, the use of higher TTC values as safety limit seems also reasonable [37, 38]. When reviewing the observed TTC values, one should keep in mind that TTC calculation was based on the moment when the target vehicle was in the middle of the two lanes and that the lane change was visible before that. Accordingly, the reported TTC values should only be used for comparison if calculation is identical.

6 Conclusions

Taking a human factors point of view on the issue of verification methods, human performance was used as a benchmark to assess HAVs’ safety performance. Because methods to quantify and directly compare HAV and human performance are rare, an appropriate method was developed. This method was based on a classical psychophysical approach and allows to quantify (the limit of) human and HAV performance in a specific scenario by means of logistic regression. In addition, it was shown that implementation as part of a driving simulator study enables testing of critical scenarios that are otherwise not available in sufficient numbers or too dangerous. The presented method therefore closes the gap between the recording and description of human performance in uncritical and critical scenarios to the point of accidents and adds to the ongoing development of adequate testing and verification methods for HAVs. Whether the identified limit of human performance holds for different scenarios and whether the method can be used for other facets of human performance has to be addressed in further research.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42154-021-00152-2>.

Acknowledgements The authors thank Eric Nicolay and Dirk Assmann for their technical support.

Authors’ Contributions KP and CS designed the work, KP acquired the data, LQ and MZ analysed the data, and LQ, CS and MZ interpreted the data. The work was drafted by LQ and MZ and revised by KP and CS. All authors have approved the submitted version.

Funding Open Access funding was enabled and organized by Projekt DEAL. The work of this paper was part of the project PEGASUS funded by the German Ministry for Economic Affairs and Energy (Bundesministerium für Wirtschaft und Energie).

Availability of Data and Materials The dataset analysed during the current study is available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alonso Raposo, M., Grosso, M., Després, J., Fernández Macías, E., Galassi, C., Krasenbrink, A. et al.: An analysis of possible socio-economic effects of a Cooperative, Connected and Automated Mobility (CCAM) in Europe - Effects of automated driving on the economy, employment and skills. EUR 29266 EN. Publications Office of the European Union, Luxembourg (2018). <https://doi.org/10.2760/777>
- Fagnant, D.J., Kockelman, K.: Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transp. Res. Pol. Pract.* **77**, 167–181 (2015). <https://doi.org/10.1016/j.tra.2015.04.003>
- Kolarova, V., Steck, F., Bahamonde-Birke, F.J.: Assessing the effect of autonomous driving on value of travel time savings: a comparison between current and future preferences. *Transp. Res. Pol. Pract.* **129**, 155–169 (2019). <https://doi.org/10.1016/j.tra.2019.08.011>
- SAE International: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (J3016_201806). https://www.sae.org/standards/content/j3016_201806/ (2018). Accessed 18 May 2021
- Lee, D., Hess, D.J.: Regulations for on-road testing of connected and automated vehicles: assessing the potential for global safety harmonization. *Transp. Res. Pol. Pract.* **136**, 85–98 (2020). <https://doi.org/10.1016/j.tra.2020.03.026>
- PEGASUS: PEGASUS method. An overview. <https://www.pegasusprojekt.de/files/tmpl/Pegasus-Abschlussveranstaltung/PEGASUS-Gesamtmethode.pdf> (2018). Accessed 18 May 2021
- Federal Ministry of Transport and Digital Infrastructure: Ethics commission automated and connected driving. <https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.html?nn=187598> (2017). Accessed 18 May 2021
- Fuller, R.: Towards a general theory of driver behaviour. *Accid. Anal. Prev.* **37**, 461–472 (2005). <https://doi.org/10.1016/j.aap.2004.11.003>
- Donges, E.: A conceptual framework for active safety in road traffic. *Veh. Syst. Dyn.* **32**, 113–128 (1999). <https://doi.org/10.1076/vesd.32.2.113.2089>
- Donges, E.: Driver behavior models. In: Winner, H., Hakuli, S., Lotz, F., Singer, C. (eds.) *Handbook of Driver Assistance Systems. Basic Information, Components and Systems for Active Safety and Comfort*, pp. 19–33. Springer International Publishing, Switzerland (2016)
- Rasmussen, J.: Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Trans. Syst. Man. Cybern.* **3**, 257–266 (1983)
- Caird, J.K., Simmons, S.M., Wiley, K., Johnston, K.A., Horrey, W.J.: Does talking on a cell phone, with a passenger, or dialing affect driving performance? An updated systematic review and meta-analysis of experimental studies. *Hum. Factors* **60**, 101–133 (2018). <https://doi.org/10.1177/0018720817748145>
- Dotzauer, M., Preuk, K., Patz, D., Schießl, C.: Das autonome Fahrzeug oder der Mensch: Wer ist besser und leistungsfähiger? In: *VDI-Berichte 2335, 34. VDI/VW-Gemeinschaftstagung Fahrerassistenzsysteme und automatisiertes Fahren*, pp. 299–314. VDI Wissensforum GmbH, Wolfsburg, Germany (2018)
- Kalra, N., Paddock, S.M.: Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transp. Res. Pol. Pract.* **94**, 182–193 (2016). <https://doi.org/10.1016/j.tra.2016.09.010>
- Powelleit, M., Vollrath, M.: Situational influences on response time and maneuver choice: development of time-critical scenarios. *Accid. Anal. Prev.* **122**, 48–62 (2019). <https://doi.org/10.1016/j.aap.2018.09.021>
- Precht, L., Keinath, A., Krems, J.F.: Identifying the main factors contributing to driving errors and traffic violations—results from naturalistic driving data. *Transp. Res. F Traffic. Psychol. Behav.* **49**, 49–92 (2017). <https://doi.org/10.1016/j.trf.2017.06.002>
- Winkle, T.: Safety benefits of automated vehicles: extended findings from accident research for development, validation and testing. In: Maurer, M., Gerdes, J.C., Lenz, B., Winner, H. (eds.) *Autonomous Driving*, pp. 335–364. Springer, Berlin (2016)
- GIDAS (German in-depth accident study) project. <https://www.gidas.org/en/willkommen/> (2021). Accessed 18 May 2021
- Barnard, Y., Utesch, F., van Nes, N., Eenink, R., Baumann, M.: The study design of UDRIVE: the naturalistic driving study across Europe for cars, trucks and scooters. *Eur. Transp. Res. Rev.* (2016). <https://doi.org/10.1007/s12544-016-0202-z>
- Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J. et al.: The 100-car naturalistic driving study, Phase II - Results of the 100-car field experiment (DOT HS 810 593). United States Department of Transportation. National Highway Traffic Safety Administration. <https://rosap.nhtl.gov/view/dot/37370> (2006). Accessed 18 May 2021
- Lyu, N., Deng, C., Xie, L., Wu, C., Duan, Z.: A field operational test in China: Exploring the effect of an advanced driver assistance system on driving performance and braking behavior. *Transp. Res. F Traffic. Psychol. Behav.* **65**, 730–747 (2019). <https://doi.org/10.1016/j.trf.2018.01.003>
- Weinberger, M., Winner, H., Bubb, H.: Adaptive cruise control field operational test - the learning phase. *JSAE Rev.* **22**, 487–494 (2001). [https://doi.org/10.1016/S0389-4304\(01\)00142-4](https://doi.org/10.1016/S0389-4304(01)00142-4)
- Preuk, K., Schießl, C.: Menschliche Leistungsfähigkeit als Gütekriterium für die Zulassung automatisierter Fahrzeuge: Methode zur Ermittlung der Grenzen menschlicher Leistungsfähigkeit. Paper presented at the 9th VDI-Fachtagung Der Fahrer im 21. Jahrhundert, DLR Braunschweig, 21–22 November 2017

24. Enke, K.: Possibilities for improving safety within the driver-vehicle-environment control loop. In: 7th International Technical Conference on Experimental Safety Vehicle Proceedings, Washington, USA, 5–8 June 1979
25. Green, M.: “How long does it take to stop?” Methodological analysis of driver perception-brake times. *Transp. Hum. Factors* **2**, 195–216 (2000)
26. Hacker, W.: *Allgemeine Arbeitspsychologie: Psychische Regulation von Wissens-, Denk- und körperlicher Arbeit. Schriften zur Arbeitspsychologie: Vol. 58.* Huber, Bern (2005)
27. Rasmussen, J.: Human errors. A taxonomy for describing human malfunction in industrial installations. *J. Occup. Accid.* **4**, 311–333 (1982)
28. Zimmer, A.: Wie intelligent darf/muss ein Auto sein? Anmerkungen aus ingenieurspsychologischer Sicht. In: Jürgensohn, T., Timpe, K.P. (eds.) *Kraftfahrzeugführung*, pp. 39–55. Springer, Berlin (2008)
29. Graab, B., Donner, E., Chiellino, U., Hoppe, M.: Analyse von Verkehrsunfällen hinsichtlich unterschiedlicher Fahrerpopulationen und daraus ableitbarer Ergebnisse für die Entwicklung adaptiver Fahrerassistenzsysteme. Paper presented at 3. Tagung Aktive Sicherheit durch Fahrerassistenz, TU München, Garching bei München, 7–8 April 2008
30. Chiellino, U., Winkle, T., Graab, B., Ernstberger, A., Donner, E., Nerlich, M.: Was können Fahrerassistenzsysteme im Unfallgeschehen leisten? *Zeitschrift für Verkehrssicherheit* **3**, 131–137 (2010)
31. National Highway Traffic Safety Administration: National motor vehicle crash causation survey: Report to congress (DOT HS 811 059). <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059> (2008). Accessed 18 May 2021
32. Hayward, J.C.: Near-miss determination through use of a scale of danger (Report TTSC-7115). The Pennsylvania State University. Pennsylvania Transportation and Traffic Safety Center (1972)
33. PEGASUS: Bericht zu Meilenstein 1: Festlegung grundlegende Anforderungen an das Testen (2016)
34. Fechner, G.T.: *Elemente der Psychophysik.* Breitkopf und Härtel, Leipzig (1860)
35. Lewald, J., Ehrenstein, W.H.: Influence of head-to-trunk position on sound lateralization. *Exp. Brain Res.* **121**, 230–238 (1998)
36. Nolden, S., Haering, C., Kiesel, A.: Assessing intentional binding with the method of constant stimuli. *Conscious Cognit.* **21**, 1176–1185 (2012)
37. Mahmud, S.S., Ferreira, L., Hoque, M.S., Tavassoli, A.: Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs. *IATSS Res.* **41**, 153–163 (2017). <https://doi.org/10.1016/j.iatssr.2017.02.001>
38. Vogel, K.: A comparison of headway and time to collision as safety indicators. *Accid. Anal. Prev.* **35**, 427–433 (2003). [https://doi.org/10.1016/S0001-4575\(02\)00022-2](https://doi.org/10.1016/S0001-4575(02)00022-2)
39. Summala, H.: Brake reaction times and driver behavior analysis. *Transp. Hum. Factors* **2**, 217–226 (2000)
40. Svensson, Å.: A method for analysing the traffic process in a safety perspective. Dissertation, University of Lund, Sweden (1998)