Toward the Definition of Competent Driving for the Assessment of Automated Driving Systems

Johannes Plaum*, Arturo Tejada†, Janine Günther* and Eric Sax‡

Abstract: One of the challenges for the assessment of automated driving systems (ADS) is the definition of reasonable release thresholds. Scenario-based reference models, like the "competent and careful human driver" introduced in UN regulation No. 157, can be integrated into a larger scenario-based testing process within a safety assessment program for ADS. This article extends the goal structuring notation (GSN) developed in the VVM project by a practically applicable methodology to derive scenario-based "competent driver" models from human reference driver data, which can serve as scenario-based assessment criteria. Based on an established role and procedures for safe on-road testing, the in-vehicle fallback test driver (IFTD), including presence of an in-cab safety conductor (SC) and adhering to a variety of safety management controls, is used as a human reference driver representing a competent and careful driver. The model development methodology is piloted using three collected on-road driving datasets.

Keywords: Driving Automation, Scenario-based testing, Safety argumentation

1 Introduction

Scenario-based testing has become a valuable component in state-of-the-art approaches for testing an automated driving system (ADS) [1]. As ADS deployment is in its infancy with real-world data lacking, particularly for commercial motor vehicles equipped with ADS according to SAE Level 4 [2], a prospective safety assessment has to be performed to predict the impact on traffic safety [3]. Using so-called criticality metrics, risks are identified, eliminated or reduced during development to a reasonable level [4]. Distinguishing what is reasonable from unreasonable is one of the more challenging elements of safety assessments, and so a variety of approaches are used, including comparing ADS performance relative to humans.

To establish language and qualitative reference points, Blumenthal et al. [5] interviewed different stakeholders and discussed multiple aspects of safety in their report "Safe enough," including safety as a threshold based on human performance. Using an average human driver for benchmarking is discussed as an "unsatisfying option, [whereas a] better-than-average, or safe, human driver is a preferable alternative" [5]. This principle was included in UN regulation No. 157 for active lane keeping systems (ALKS) [6], which

*Johannes Plaum and Janine Günther are with Torc Europe GmbH (e-mail: {johannes.plaum, janine.guenther}@torc.ai).

†Arturo Tejada is with the Dynamics and Control Group at the Mechanical Engineering Department at Eindhoven University of Technology (e-mail: arturo.tejadaruiz@tno.nl).

‡Eric Sax is with Institut für Technik der Informationsverarbeitung at Karlsruher Institut für Technologie (e-mail: eric.sax@kit.edu).

defines a ”competent and careful human driver” as the reference for collision avoidance performance and is the first UNECE regulation allowing type approval of SAE Level 3 systems [2]. Considering social acceptance of ADS, the German ethics commission on automated and connected driving proposed that an ADS needs to ”promise at least a reduction in damage in the sense of a positive risk balance” [7]. While alignment on the definition of a reasonable human reference for ADS and the implementation of such a reference are subject to ongoing discussion among researchers, industry, society broadly, regulators and policymakers, this paper contributes to this discussion in the following areas:

Section 3 proposes an extension of the goal structuring notation (GSN) forming the VVM safety argumentation [8] to include the derivation of a competent driver model based on collected human driving data. As the basis for a competent driver model, the in-vehicle fallback test driver (IFTD) [9], supported and monitored by an in-cab safety conductor (SC), is proposed as a baseline alternative to an average driver. Section 4 applies the methodology to collected real-world data from IFTDs' manual on-road driving and compares it to observational drone video recordings of a single highway site. This site collection, and the term 'average' is used as a stand-in to represent what, in the final application, would likely need to be a carefully designed, multi-site, multi-condition driver data set to cover the intended operational design domain (ODD). The commonly used intelligent driver model (IDM) [10] for the lead vehicle (LV) following scenario is calibrated based on the IFTD data. In the last step, the IFTD model is applied to an on-road dataset from an ADS system under test (SUT) to evaluate the similarity of ADS behavior to the competent driver model in contrast to the average driver model. This paper introduces a proposed method that might be used for comparing ADS performance to human performance, recognizing a need to further develop the method and establish what scale of data is appropriate for use as model inputs. Section 5 provides a conclusion, discussion of limitations and outlook of future work.

2 Scenario-Based Safety Assessment

Scenario-based testing describes and tests within an ODD in a structured and scalable approach that complements a suite of test methods including mileage accumulation approaches [11]. In the VVM project a safety argumentation structure based on a goal structuring notation was developed and includes an argumentation for the absence of unreasonable risk [8]. In addition to a traceable derivation and execution of test cases based on the system specification, release criteria and corresponding thresholds are required. Whereas the VVM argumentation provides an overall structure, it does not provide examples on how to apply the method or define release thresholds. Salem et al. [12] emphasize the definition of acceptance criteria as an indispensable part of societal discussion, but do not consider it in the scope of their work on the risk management core.

Favarò et al. [13] suggest organizing the evaluation of unreasonable risk at an aggregate-level and event-level. An example of a criterion to quantify the risk at an aggregate level is the collision rate estimation including the estimation error as shown by de Gelder and Op den Camp in [14]. Having restricted included data to represent the ODD and having selected events such that the selection process maintains the intent of the analysis, crash statistics can be developed and used to evaluate the ADS performance relative to the human driver performance [15]. Whereas the aggregated view provides a high-level positive or negative risk indication, scenarios with an ADS performance better or worse than the human driver could unintentionally be lost within the aggregate measure. For example, deploying an ADS exclusively using aggregate-level measures could lead to unreasonable risk in scenarios that might be controllable for a human driver. Consequently, additional event-level measures are recommended by Favarò et al. [13] to ensure that human performance is achieved or exceeded in each scenario. A practical way to apply such event-level measures is through so-called performance reference models. In addition to the models in UN regulation No. 157 [6], several models have been developed through industry efforts, including the non-impaired, with eyes on the conflict (NIEON) model [16], the stochastic cognitive model (SCM) [3], or the responsibility sensitive safety (RSS) model [17].

Whereas most of the previously mentioned performance reference models concentrate on collision avoidance, Tejada et al. [18] propose a practical methodology for the definition of competent driving with the focus on social and responsible driving or "roadmanship" [5]. This approach extracts driving rules and recommendations from driving manuals and formalizes them using assertions. Based on human data collected in a naturalistic driving study (NDS), the application of the assertions is determined. To define thresholds for competent driving, annotations of driving instructors are used to define the boundary between acceptable and unacceptable driving and extract corresponding assertion thresholds. The assertions are then used to describe the acceptable "envelope of operation" representing good roadmanship [18].

With the goal of defining a competent human driving model, a variation of the methodology proposed by Tejada et al. [18] is employed here, focusing on nominal driving and roadmanship. In this approach continuously recorded manual driving data from highly trained drivers (IFTDs), instead of driving instructor annotations at specific points-in-time, is used to represent competent driving.

3 Methodology

Given the interest in including the human reference in a safety assurance process, this proposed methodology aims to answer the question of how competent driving can be defined. The four steps of the proposed methodology for the data collection, model derivation and ADS assessment in comparison to the selected human reference driver are introduced in the following sections.

Fig. 1 shows the three strategies resulting from the methodology, extending the GSN developed in the VVM safety argumentation [8]. The linked sub-goals and solutions are discussed in the application example (see section 4). Note that these steps are intended as one method within a variety of approaches to define complementary risk acceptance criteria, indicated by the undeveloped element decorator below "Strategy_12".

3.1 Selection of the human reference driver

The concept of using a human reference for competent driving will be called "human reference driver" in the following and can be used to derive both aggregate- and event-level acceptance criteria (see Strategy_12 in Fig. 1). As mentioned, the UN regulation No. 157 for ALKS [6] is the first standard incorporating a limited-scope "competent and careful"

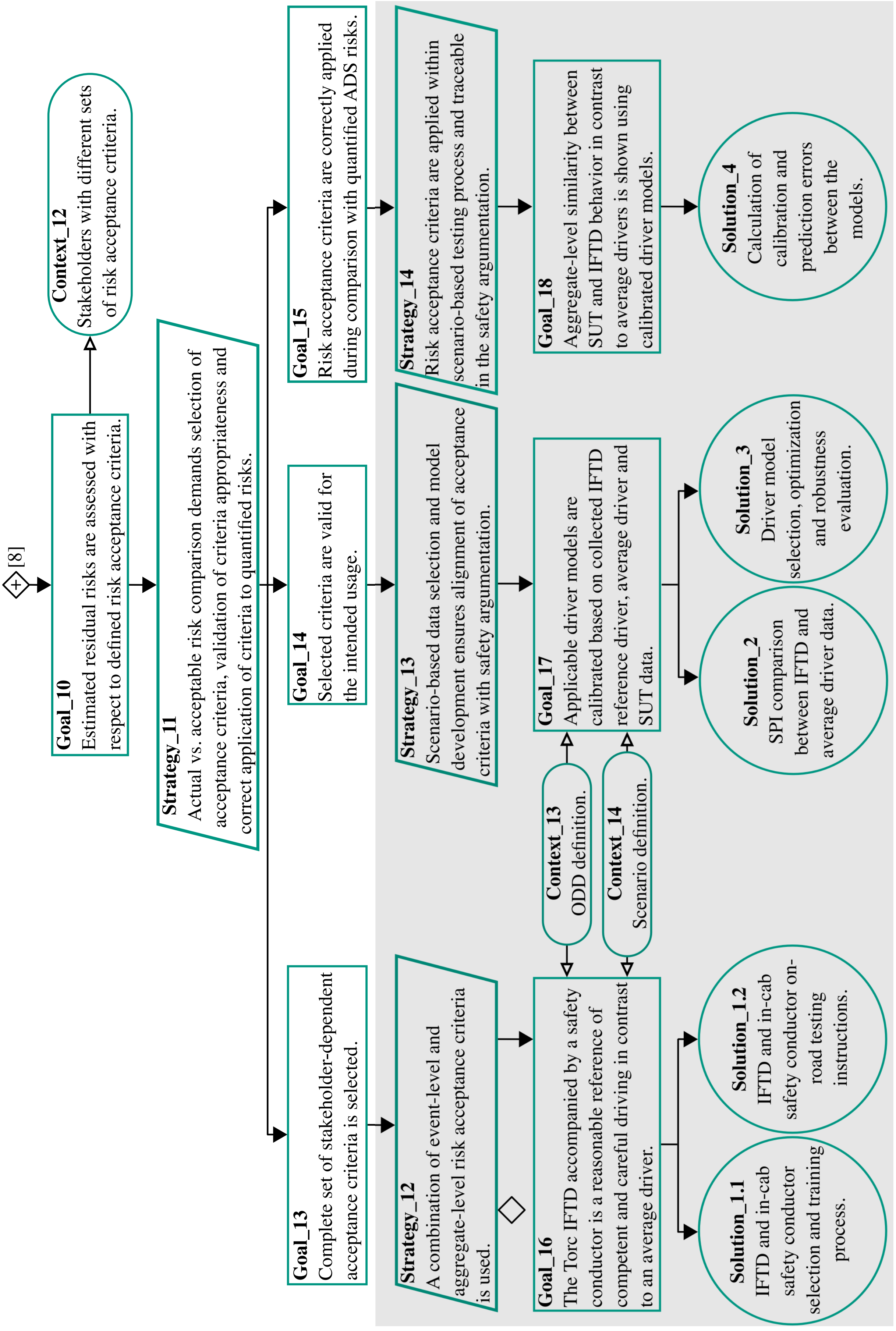

[8]
Goal_10
Estimated residual risks are assessed with respect to defined risk acceptance criteria.
Context_12
Stakeholders with different sets of risk acceptance crtiteria.
Strategy_11
Actual vs. acceptable risk comparison demands selection of acceptance criteria, validation of criteria appropriateness and correct application of criteria to quantified risks.
Goal_13
Complete set of stakeholder-dependent acceptance criteria is selected.
Goal_14
Selected criteria are valid for the intended usage.
Goal_15
Risk acceptance criteria are correctly applied during comparison with quantified ADS risks.
Strategy_12
A combination of event-level and aggregate-level risk acceptance criteria is used.
Strategy_13
Scenario-based data selection and model development ensures alignment of acceptance criteria with safety argumentation.
Strategy_14
Risk acceptance criteria are applied within scenario-based testing process and traceable in the safety argumentation.
Goal_16
The Torc IFTD accompanied by a safety conductor is a reasonable reference of competent and careful driving in contrast to an average driver.
Context_13
ODD definition.
Context_14
Scenario definition.
Goal_17
Applicable driver models are calibrated based on collected IFTD reference driver, average driver and SUT data.
Goal_18
Aggregate-level similarity between SUT and IFTD behavior in contrast to average drivers is shown using calibrated driver models.
Solution_1.1
IFTD and in-cab safety conductor selection and training process.
Solution_1.2
IFTD and in-cab safety conductor on-road testing instructions.
Solution_2
SPI comparison between IFTD and average driver data.
Solution_3
Driver model selection, optimization and robustness evaluation.
Solution_4
Calculation of calibration and prediction errors between the models.

Figure 1: GSN from the VVM safety argumentation [8] extended (grey) to apply a competent driver model based on a selected human reference driver

human reference model. From an industry perspective, human drivers play an important role as test drivers to ensure safe public road testing. SAE J3018 [9] defines a fallback test driver as a "person specially trained and skilled in supervising the performance of prototype ADS-operated vehicles in on-road traffic for testing purposes."

As mentioned by the Automated Vehicle Safety Consortium (AVSC) best practice [19], "an IFTD becomes a conventional (test) driver" when performing the complete dynamic driving task (DDT). Nevertheless, the term IFTD will be used for data collected with the driver performing the complete DDT, to emphasize that the selection and training of an IFTD differentiates the IFTD from a distribution of average driver performance (Solution_1.1). In addition, SCs are currently used to support and monitor the IFTD by leading and/or coordinating the remaining testing tasks [20].

Based on the established role of the IFTD, we propose using the IFTD accompanied by a SC as possible reference for the definition of competent driving (Solution_1.2). Whereas the main reason for using IFTDs is a safe response to unexpected hazardous operating situations during public road testing, fulfilling socially accepted driving standards is also taken into account.

3.2 Data collection

The definition of the human reference driver has a strong influence on the necessary data collection and follows the scenario-based testing process (Strategy_13). If the average driver is considered, an NDS like SHRP2 [21] or drone data [22] can be used to characterize the behavior and performance of drivers broadly or in specifically relevant ODD conditions or scenarios. For safety-critical scenarios, accident databases like GIDAS [23] or CRSS [24] serve as a potential source. Whereas this allows an aggregated comparison, accident databases typically do not provide sufficient detail for the derivation of a driver model. If a selected human driver like the IFTD serves as a reference, a custom driving data collection can be performed. The results for the selected reference driver can then be compared to larger datasets from an NDS [25].

3.3 Model derivation

For a competent human driver model, a generic model applicable to all kinds of traffic scenarios is preferable. As this is just as challenging as the development of an ADS itself, a scenario-based evaluation and calibration to the scenarios found within the intended ODD is a more targeted, and therefore feasible, approach. The main benefit is that a detailed behavior model depending on the corresponding scenario can be used and directly applied in the scenario-based testing process (Strategy_13). This approach requires a scenario identification method to select the data as a basis for the model derivation [26, 27].

3.4 ADS assessment

Based on the derived driving model, under our proposed approach, competent driving behavior is measured by the similarity of ADS behavior to the competent driver model. This is based on the assumption, that an unreasonable deviation of ADS behavior from the competent driver model could lead to traffic disturbances and possibly hinder social acceptance of ADS [18]. The scenario-based model definition allows a seamless integration into the scenario-based testing process and is traced in the GSN (Strategy_14). The similarity of ADS driving behavior to the model can be linked to the overall release argumentation as a safety performance indicator (SPI), focusing on one or both areas of nominal driving and safety-critical scenarios depending on the scope of the competent driver model.

4 Application

As a practical example, the proposed methodology is applied and evaluated using three collected real-world datasets described in this section.

4.1 Selection of the human reference driver

The IFTD fulfilling the complete DDT while being accompanied by a SC (see section 3) is proposed as the human reference driver to model competent driving in this application example. Specifically, we consider Torc IFTDs driving class 8 (gross vehicle weight rating>33.001 lbs) trucks.

4.2 Data collection

The first step for the data collection is the specification of the ODD and the SUT. For this application example, the ODD was restricted to highway driving in the southwest region of the USA. Highly automated class 8 trucks (SAE Level 4 [2]) are considered as the SUT.

4.2.1 IFTD dataset

The IFTD dataset was collected in on-road test drives with class 8 trucks, while the IFTD was performing the complete DDT manually. As the goal is to create a competent driver model and not to evaluate individual driver performance, the 23 individual Torc IFTDs involved in the data collection are treated as a group in the analysis. Various sensors, including camera, radar and lidar in combination with a proprietary object detection algorithm were used to extract the properties of the ego vehicle and all traffic participants.

4.2.2 Drone dataset

For the purpose of providing data for average driver modeling, a drone site data collection was used [28]. A total of six hours of drone recordings was collected for one location on the Interstate 40 near Albuquerque, New Mexico. The field of view captured approximately 300 meters in both driving directions with a speed limit of 65 mph (see Fig. 2). A 3D detection and tracking of all traffic objects was performed by DeepScenario[1] using computer vision algorithms. The object dimensions and movements are reconstructed in metric space and given in a high-definition map of the recorded area.

The main benefit of drone data is the unobstructed bird's-eye view on all traffic participants. From a scenario perspective, each recorded class 8 truck can be considered as

[1] https://www.deepscenario.com/

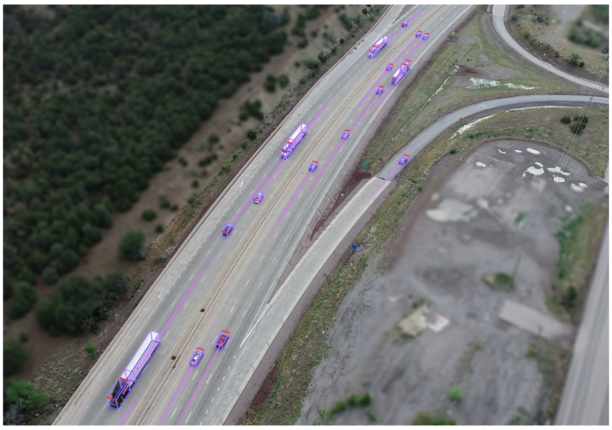


Figure 2: Drone recording annotated with 3D bounding boxes from DeepScenario

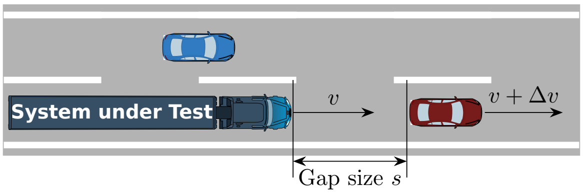


Figure 3: LV following scenario

the ego vehicle, leading to an increased number of captured scenarios per time for the creation of an average driver distribution from the recording site. The total number of detected vehicles in the six-hour dataset is shown in Table 1.

Car	Class 8 truck	Other truck	Bus	Motorcycle	Total
7657	2658	4539	43	59	14956

Table 1: Numbers of vehicles included in the drone dataset

4.2.3 Scenario identification

In the next step, scenario identification on both datasets is performed based on the StreetWise method [26]. First, all activities (e.g., acceleration or deceleration) of the vehicles are extracted. Second, the positions of all vehicles are matched to the map and lane assignments are created. Third, scenarios are identified based on a specified sequence of activities. In the last step, scenario parameters like the time gap to the LV are calculated. In this example, car-following behavior is investigated with the following scenario definition:

1. Ego vehicle is a class 8 truck
2. Steady-state car-following: LV acceleration between $-2.0\,\mathrm{m\,s^{-2}}$ and $2.0\,\mathrm{m\,s^{-2}}$
3. Highway with speed limit of 65 mph
4. Drone data captured only flowing traffic: lower limit of mean speed in IFTD scenarios set to the minimal value of 36.1 mph from the drone data
5. Minimum scenario duration of 2 s

Fig. 3 shows the main parameters used for the LV following scenario.

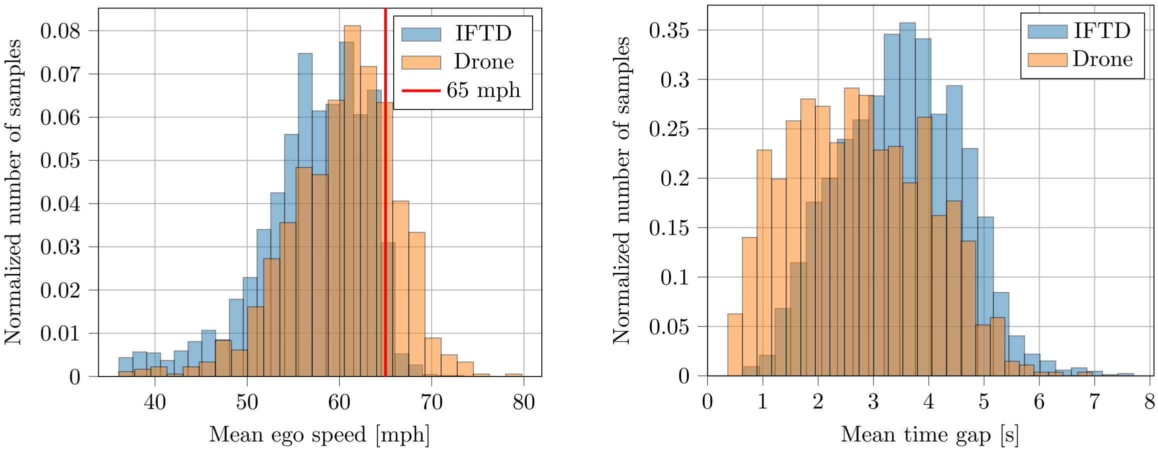
As a result from the scenario identification, a total of 3068 scenarios were identified in the IFTD dataset and 1029 scenarios in the drone dataset.

4.2.4 Comparison of datasets

To substantiate the selection of the manual driving IFTD data as a reasonable reference for competent driving, the following hypothesis was investigated by a comparison of the

IFTD and drone dataset (Goal_16): "The Torc IFTD accompanied by a SC is a reasonable reference of competent and careful driving in contrast to an average driver."

In a NHTSA report on pre-crash scenarios, Toma et al. [29] identified traveling too fast and unsafe following distance as main contributing factors for rear-end collisions of heavy trucks (see Fig. 3). Both factors were investigated for the drone site average driver and IFTD. The distribution of the mean ego speed for the drivers from the IFTD and drone dataset with a speed limit of 65 mph is shown in Fig. 4a. The mean of the distribution of speeds for the IFTD samples was 57.1 mph (standard deviation (SD) 6.01 mph) and 3 % of the scenarios exceeding the speed limit of 65 mph. The mean of the distribution of speeds for the drone samples was 60.4 mph (SD 5.81 mph) and 20 % of the scenarios exceeding the speed limit.



(a) Mean ego vehicle speed distribution

(b) Mean time gap distribution

Figure 4: Distributions in the LV following scenario

Fig. 4b shows the distribution of the mean time gap during the scenarios. The mean value of the mean time gap was 3.49 s for the IFTD samples (SD 1.09 s) and 2.74 s for the drone samples (SD 1.23 s). A time gap below 2 s occurred in 10 % of the IFTD scenarios, compared to 32 % for the drone data. We note that small time gaps are dependent on other traffic participants' behavior and cannot be avoided completely, e.g., if a cut-in of another vehicle in front of the ego precedes the LV following scenario, the initial time gap can be low until a larger following distance is created.

Extensive selection and training processes qualify Torc IFTDs. This includes the review of driving records and experience as well as classroom, proving ground and on-road training. Based on the collected data, the IFTD keeping lower speeds and reducing the occurrence of small time gaps supports the hypothesis that the IFTD is a reasonable representation of competent and careful driving compared to the drone site average driver (Solution_2). Overall, speeding was observed six times more frequently and with higher maximum speeds by the drone site average driver compared to the IFTD. The mean time gap for the IFTD was 27 % larger compared to the drone site average driver, with three times fewer occurrences of time gaps below 2 s.

4.3 Model derivation

Based on the identified scenarios, a suitable driver behavior model must be selected and calibrated to the IFTD data in order to derive a competent driver model. For the considered LV following scenario with limited deceleration, longitudinal car-following models are available. We chose the widely-used IDM developed by Treiber et al. [10] to derive an IFTD model for the LV following scenario. The IDM can represent multiple aspects of single-lane car-following, including accelerating on a free road, approaching a slower or stopped LV and actively modulating speed while following a LV [10]. The IDM's inputs include the ego vehicle longitudinal velocity v, the relative longitudinal velocity Δv and bumper-to-bumper distance s to the LV (see Fig. 3). The model parameters are the maximum acceleration a, the desired deceleration b, the acceleration exponent Δ, the minimum distance s_0 and desired time gap T_0 to the LV and the desired ego velocity v_0.

The desired distance s^* and acceleration $\dot{v}$ of the IDM are calculated as:

$$s^*(v, \Delta v) \;=\; s_0 + max\left(0, vT_0 + \frac{v\Delta v}{s\sqrt{ab}}\right) \tag{1}$$

$$\dot{v} \;=\; a\left[1 - \left(\frac{v}{v_0}\right)^{\delta} - \left(\frac{s^*(v, \Delta v)}{s}\right)^2\right] \tag{2}$$

In the next step, we use the global optimization algorithm called DIRECT-SQP, which was proposed by Li et al. [30] to calibrate the IDM to recorded driving data. As objective function the sum of squared velocity errors between the recorded velocity v and the simulated velocity $\hat{v}$ is selected. For the calculation of the discrete-time model, a fixed update time interval of 50 ms (20 Hz) was used for both datasets. N is the number of time steps and M is the number of scenarios used for the calibration. The calibrated parameters $\theta \;=\; [a, b, \Delta, s_0, T_0, v_0]$ are the solution to the optimization problem:

$$\min_{\theta} g(\theta) \;=\; \sum_{i=1}^{M}\sum_{j=1}^{N}\left(v_i - \hat{v}_i(\theta)\right)^2 \tag{3}$$

The same approach could also be applied to objective functions including the time gap [10] or the safety objective function proposed in [31]. Table 2 shows the calibrated parameters for the IDM for the IFTD dataset using the defined parameter boundaries.

	$a[\mathrm{m\,s^{-2}}]$	$b[\mathrm{m\,s^{-2}}]$	δ[-]	s_0[m]	T_0[s]	$v_0[\mathrm{m\,s^{-1}}]$
[min, max]	[0.1, 6.0]	[0.1, 6.0]	[2.0, 4.0]	[2.0, 5.0]	[0.5, 6.0]	[20, 40]
IFTD	0.26	6.00	2.00	5.00	1.92	30.67

Table 2: IDM parameter bounds used for optimization and results for IFTD dataset

The acceleration exponent $\delta = 2$ is at the lower optimization boundary, indicating smoother reduction of acceleration for trucks when reaching the desired velocity, compared to the value $\delta = 4$ commonly used for cars [10]. The minimum distance $s_0 = 5\,\mathrm{m}$ is at the upper boundary of the interval specified based on literature values and higher than the commonly used value of $s_0 = 2\,\mathrm{m}$ for cars [31]. The increased value represents the practical meaning of the parameter, as truck drivers must keep a larger distance to the vehicle in front compared to cars, in case they must maneuver or cut out of standing traffic. Therefore, we choose the optimization results $\delta = 2$ and $s_0 = 5\,\mathrm{m}$ as fixed values, which additionally reduces computation efforts.

Next, we apply the bootstrap method to estimate the distribution of car-following model parameters, as performed in [32]. The optimal solution is calculated for each resampled dataset using the DIRECT-SQP optimization approach, until the desired number of samples is achieved. The global optimal solution, the sample means and 95 % confidence intervals (CI) of the IDM parameters from 1000 bootstrap samples are shown in Table 3 for the IFTD, drone and SUT data (Solution_3).

		$a[\mathrm{m\,s^{-2}}]$	$b[\mathrm{m\,s^{-2}}]$	$T_0[\mathrm{s}]$	$v_0[\mathrm{m\,s^{-1}}]$
IFTD	Optimal solution	0.26	6.00	1.92	30.67
	Bootstrap mean	0.26	6.00	1.93	30.74
	95 % CI	(0.22, 0.30)	(6.00, 6.00)	(1.78, 2.10)	(30.01, 31.70)
Drone	Optimal solution	0.34	6.00	1.04	39.32
	Bootstrap mean	0.35	6.00	1.02	38.68
	95 % CI	(0.30, 0.44)	(6.00, 6.00)	(0.86, 1.18)	(35.42, 40.0)
SUT	Optimal solution	0.28	4.36	2.68	33.37
	Bootstrap mean	0.28	4.34	2.69	33.49
	95 % CI	(0.22, 0.33)	(3.74, 4.91)	(2.47, 2.95)	(32.02, 35.42)

Table 3: IDM optimal solutions, parameter means and 95 % CI

The results for the drone data show a 46 % lower value of T_0 and a 28 % higher value of v_0 compared to the IFTD model, representing the lower following distances and higher maximum speeds of the drone site average driver compared to the IFTD as discussed in subsection 4.2.

4.4 Assessment of ADS behavior

In the last step, the derived competent driver model is used for the assessment of ADS behavior. Given a recorded scenario of the SUT, the competent driver model is simulated using the initial speed and gap size of the ego vehicle as well as the recorded speed of the LV as input (using 50 ms/20 Hz for the calibration). The simulated IDM trajectory is then compared to the recorded trajectory of the SUT.

For this application example, 1147 scenarios from real-world testing of an ADS as SUT were collected. The same ODD, scenario definition and scenario identification method are used as described in subsection 4.2, effectively creating a unified basis from different data sources. The calibration and bootstrap approach is applied to the real-world testing dataset of the SUT. The results shown in Table 3 include the highest value of the desired time gap T_0, leading to the largest following distances of the three models.

In the next step, similarity of SUT behavior to the derived competent driver model is evaluated. We compare the results by calculating the root mean squared error (RMSE) between the simulated and recorded speed for each combination of the global optimal solutions of the IFTD, drone and SUT and the three underlying datasets used for calibration (see Table 4).

First, the RMSE of speed for each dataset shows that the lowest error per dataset is achieved for the optimal IDM solution fitted to the corresponding dataset, verifying the result of the IDM calibration. Second, the prediction error of $0.502\,\mathrm{m\,s^{-1}}$ calculated with

	Optimal IDM solution		
RMSE [m/s]	IFTD	Drone	SUT
IFTD dataset	0.548	0.981	0.602
Drone dataset	0.563	0.429	0.732
SUT dataset	0.502	1.154	0.490

Table 4: RMSE of speed calculated for each combination of IDM parameters and datasets

the optimal IFTD parameters for the SUT dataset is only 2 % higher compared to the fitting error of $0.490\,\mathrm{m\,s^{-1}}$ for the optimal SUT parameters, indicating a similarity in the behavior of IFTDs and SUT. Third, the prediction error of $1.154\,\mathrm{m\,s^{-1}}$ calculated with the optimal drone parameters for the SUT dataset is 136 % higher compared to the fitting error of $0.490\,\mathrm{m\,s^{-1}}$ for the optimal SUT parameters, showing a significant difference in drone site average driver and SUT behavior.

Overall, the difference between the prediction errors and the fitting error for the SUT dataset shows a higher similarity between SUT and IFTD behavior compared to the drone site average driver representation. The result confirms that, in contrast to the drone site average driver, the SUT aligns more closely with the behavior defined by the competent driver model (Solution_4).

5 Conclusion and future work

This article proposes a future methodology for the selection and derivation of a competent driver model based on the human driver. The methodology is integrated into the safety argumentation GSN from the VVM project, providing a practical applicable example to derive risk acceptance criteria based on a competent driver model. Building on the established role of the IFTD for safe public road testing, we explored the use of IFTD driving as a data source for a competent driver model. We show that through the comparison of this reference driver to a drone site average driver, the IFTD-informed competent driver model is useful as a reference. Finally, evaluating the fitting and prediction errors of the IDM on ADS SUT data collected during real-world testing shows the higher similarity of the SUT behavior to the competent driver model in contrast to the drone site average driver model. The method provides a promising approach to define competent driving and average driving within a scenario, and subsequent steps that could be used to assess ADS performance against these benchmarks.

Limitations of the application example include that effects not considered in the competent driver model can lead to false positives for the detection of deviations. Whereas the performed evaluation of the fitting and prediction errors shows the overall similarity between the datasets, a reasonable case-by-case comparison depends on the driver model. As the IDM only considers the behavior of the LV, a reaction to vehicles in adjacent lanes or in various weather conditions was not taken into account. To accurately represent the influence of other factors on human behavior, additional effort to extend the competent driver models would be required. Moreover, the collected limited datasets only serve as a pilot, and more data, including consideration of uncertainty, would be required for transition of these methods to an commercial application. Future work includes the application of the methodology to further scenarios and driver models. The approach could also be extended to onboard monitoring to detect deviations from the competent driving model.

References

[1] C. Neurohr, L. Westhofen, T. Henning, T. De Graaff, E. Mohlmann, and E. Bode, "Fundamental Considerations around Scenario-Based Testing for Automated Driving," *IEEE Intelligent Vehicles Symposium, Proceedings*, no. Iv, pp. 121–127, 2020.

[2] SAE International, "J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 2021.

[3] A. Fries, F. Fahrenkrog, K. Donauer, M. Mai, and F. Raisch, "Driver Behavior Model for the Safety Assessment of Automated Driving," *IEEE Intelligent Vehicles Symposium, Proceedings*, vol. 2022-June, no. Iv, pp. 1669–1674, 2022.

[4] L. Westhofen, C. Neurohr, T. Koopmann, M. Butz, B. Schütt, F. Utesch, B. Neurohr, C. Gutenkunst, and E. Böde, "Criticality metrics for automated driving: A review and suitability analysis of the state of the art," *Archives of Computational Methods in Engineering*, pp. 1–35, 2022.

[5] M. Blumenthal, L. Fraade-Blanar, R. Best, and J. Irwin, *Safe Enough: Approaches to Assessing Acceptable Safety for Automated Vehicles.* Santa Monica, CA: RAND Corporation, 2020. [Online]. Available: https://www.rand.org/pubs/research_reports/RRA569-1.html.

[6] U. N. E. C. for Europe (UNECE), "Proposal for the 01 series of amendments to UN Regulation No. 157 (Automated Lane Keeping Systems)," *Addendum 156 to the 1958 Agreement - UN Regulation No. 157*, 2022.

[7] U. Di Fabio, M. Broy, R. J. Brüngger, U. Eichhorn, A. Grunwald, D. Heckmann, and E. Hilgendorf, "Ethic commission: automated and connected driving, Report of ethics commission appointed by the federal minister of transport and digital infrastructure," 2017. [Online]. Available: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

[8] J. Reich, M. Nolte, N. F. Salem, T. Brade, M. Fistler, D. Hillen, and C. Lalitsch-Schneider, "VVM Safety Argumentation," 2023. [Online]. Available: https://www.vvm-projekt.de/en/argumentation

[9] SAE International, "J3018 - Safety-Relevant Guidance for On-Road Testing of SAE Level 3, 4, and 5 Prototype Automated Driving System (ADS)-Operated Vehicles," 2020.

[10] M. Treiber and A. Kesting, "Microscopic Calibration and Validation of Car-Following Models – A Systematic Approach," *Procedia - Social and Behavioral Sciences*, vol. 80, pp. 922–939, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.sbspro.2013.05.050

[11] F. Reisgys, J. Plaum, A. Schwarzhaupt, and E. Sax, "Scenario-based X-in-the-Loop Test for Development of Driving Automation," in *14. Workshop Fahrerassistenzsysteme und automatisiertes Fahren.* Uni-DAS e.V., 2022, pp. 207–216.

[12] N. F. Salem, T. Kirschbaum, M. Nolte, C. Lalitsch-Schneider, R. Graubohm, J. Reich, and M. Maurer, “Risk management core -toward an explicit representation of risk in automated driving,” *IEEE Access*, vol. 12, pp. 33 200–33 217, 2024.

[13] F. Favarò, L. Fraade-Blanar, S. Schnelle, T. Victor, M. Peña, J. Engstrom, J. Scanlon, K. Kusano, and D. Smith, “Building a Credible Case for Safety: Waymo’s Approach for the Determination of Absence of Unreasonable Risk,” 2023. [Online]. Available: https://arxiv.org/abs/2306.01917

[14] E. de Gelder and O. Op den Camp, “How certain are we that our automated driving system is safe?” *Traffic injury prevention*, vol. 24, no. S1, pp. S131–S140, 2023. [Online]. Available: https://doi.org/10.1080/15389588.2023.2186733

[15] J. M. Scanlon, K. D. Kusano, L. A. Fraade-Blanar, T. L. McMurry, Y.-H. Chen, and T. Victor, “Benchmarks for Retrospective Automated Driving System Crash Rate Analysis Using Police-Reported Crash Data,” 2023. [Online]. Available: http://arxiv.org/abs/2312.13228

[16] J. M. Scanlon, K. D. Kusano, J. Engström, and T. Victor, *Collision Avoidance Effectiveness of an Automated Driving System Using a Human Driver Behavior Reference Model in Reconstructed Fatal Collisions.* Waymo, LLC, 2022.

[17] S. Shalev-Shwartz, S. Shammah, and A. Shashua, “On a Formal Model of Safe and Scalable Self-driving Cars,” 2017. [Online]. Available: http://arxiv.org/abs/1708.06374

[18] A. Tejada, M. J. E. Legius, A. R. Kalose, P. F. V. Oliveira, E. V. Dam, and J. H. Hogema, “Towards a Practical Methodology for Defining Competent Driving,” in *2023 26th IEEE Intelligent Transportation Systems Conference (ITSC)*, 2023.

[19] Automated Vehicle Safety Consortium, “AVSC Best Practice for In-Vehicle Fallback Test Driver Selection, Training, and Oversight Procedures for Automated Vehicles Under Test,” 2019.

[20] Torc Robotics, Inc., “Innovating Safety and Efficiency - Torc Safety Report,” 2021.

[21] J. M. Hankey, M. A. Perez, and J. A. McClafferty, *Description of the SHRP 2 Naturalistic Database and the Crash, Near-Crash, and Baseline Data Sets.* Technical Report. Virginia Tech Transportation Institute, 2016.

[22] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, “The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems,” *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2118–2125, 2018. [Online]. Available: http://www.highd-dataset.com

[23] D. Otte, C. Krettek, H. Brunner, and H. Zwipp, “Scientific Approach and Methodology of a New In-Depth- Investigation Study in Germany so called GIDAS,” in *Proceedings of the 18th international conference on Enhanced Safety of Vehicles (ESV)*, 2003. [Online]. Available: http://www-esv.nhtsa.dot.gov/Proceedings/18/18ESV-000161.pdf

[24] National Highway Traffic Safety Administration, "Crash Report Sampling System." [Online]. Available: https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system

[25] Automated Vehicle Safety Consortium, "AVSC Best Practice for Developing ADS Safety Performance Thresholds Based on Human Driving Behavior," 2023.

[26] H. Elrofai, J.-P. Paardekooper, E. de Gelder, S. Kalisvaart, and O. op Den Camp, "Streetwise: Scenario-based safety validation of connected and automated driving," *Netherlands Organization for Applied Scientific Research, TNO*, p. 28, 2018.

[27] P. Elspas., Y. Klose., S. Isele., J. Bach., and E. Sax., "Time series segmentation for driving scenario detection with fully convolutional networks," in *Proceedings of the 7th International Conference on Vehicle Technology and Intelligent Transport Systems - VEHITS*, INSTICC. SciTePress, 2021, pp. 56–64.

[28] "DeepScenario Dataset: Unparalleled Sedillo (v2)," 2022. [Online]. Available: https://app.deepscenario.com

[29] S. Toma, E. Swanson, J. D. Smith, and W. G. Najm, *Heavy Truck Pre-Crash Scenarios for Safety Applications Based on Vehicle-to-Vehicle Communications.* No. DOT-VNTSC-NHTSA-11-14. United States. Department of Transportation. National Highway Traffic Safety Administration, 2014.

[30] L. Li, X. M. Chen, and L. Zhang, "A global optimization algorithm for trajectory data based car-following model calibration," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 311–332, 2016.

[31] K. Adjenughwure, A. Tejada, P. F. V. Oliveira, J. Hogema, and G. Klunder, "A New Safety Objective for the Calibration of the Intelligent Driver Model," 2023. [Online]. Available: https://arxiv.org/abs/2310.04259

[32] F. Wu, J. Lu, and J. Jiang, "Estimation of car-following model parameters distribution using bootstrap method," *Proceedings of the 3rd International Conference on Road Safety and Simulation*, 2011.