applied sciences

MDPI

Article

Objectively Scoring the Human-Likeness of Artificial Driver Models

Teresa Rock [1,2,*], Taras Hryhoruk [3], Thomas Bleher [1], Mohammad Bahram [1], Stefanie Marker [2], Arslan Ali Syed [3] and Maya Sekeran [3]

[1] BMW Group, Research and Technology, 80788 Munich, Germany
[2] Chair of Naturalistic Driving Observation for Energetic Optimisation and Accident Avoidance, Technical University of Berlin, 13355 Berlin, Germany; stefanie.marker@tu-berlin.de
[3] Chair of Traffic Engineering and Control, Technical University of Munich, 80333 Munich, Germany
* Correspondence: teresa.rock@bmw.de

Abstract: Several applications of artificially modeled drivers, such as autonomous vehicles (AVs) or surrounding traffic in driving simulations, aim to provide not only functional but also human-like behavior. The development of human-like AVs is expected to improve the interaction between AVs and humans in traffic, whereas, in a driving simulation, the objective is to create realistic replicas of real driving scenarios to investigate various research questions under safe and reproducible conditions. In urban traffic, driving behavior strongly depends on the situational context, which introduces new challenges not only for modeling but also for the evaluation of such models. However, current objective assessment strategies rarely consider situational context and human similarity, whereas subjective approaches are not suitable for iterative development processes. In this paper, we present a first attempt to make the plausibility and human-likeness of vehicles' trajectories objectively measurable. A multidimensional quality function is presented that incorporates various parameters characterizing human-like driving behavior and compares each of those parameters to human driving behavior under similar conditions. Among other things, our validation results show that the presented evaluation methodology is scalable to a wide range of situations has the ability to identify model weaknesses, and is able to reflect the way people distinguish between artificial and human behavior.

Keywords: traffic data; driver models; autonomous driving; driving simualtion; urban; human factors

check for updates

Citation: Rock, T.; Hryhoruk, T.; Bleher, T.; Bahram, M.; Marker, S.; Syed, A.A.; Sekeran, M. Objectively Scoring the Human-Likeness of Artificial Driver Models. *Appl. Sci.* **2023**, *13*, 10218. https://doi.org/10.3390/app131810218

Academic Editor: Christos Bouras

Received: 10 July 2023
Revised: 16 August 2023
Accepted: 28 August 2023
Published: 11 September 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

There are various applications for artificially modeled road users. In driving simulations, for example, the objective is to recreate a realistic driving scenario including artificially modeled drivers. In Driver-in-the-Loop (DiL) applications, the aim is to achieve a high degree of presence of participants during the experiment by providing realistic interactions between the driver and artificially modeled drivers [1]. Software-in-the-Loop (SiL) applications, on the other hand, require the human-like behavior of surrounding traffic to investigate individual research questions involving interactions between an AV function and other road users. The constantly growing field of autonomous driving also ultimately faces the challenge of replacing the human driver with a complex model that can safely and meaningfully handle the complex task of driving. The development of human-like driving capabilities in AVs is expected to enhance the ability of surrounding drivers to understand and anticipate the behavior of AVs, resulting in more natural interactions [2]. As a result, AVs are required to mimic human-like driving behavior [3]. Following, human-like behavior should be considered in the automated driving design, as mentioned by Hang et al. [4] and in driver models noted by Lindorfer et al. [5]. Due to the complexity of such tasks, most likely there will not be a single perfect solution for all applications, and there is still a long development journey ahead. Therefore, meaningful evaluation strategies are important to determine the capabilities and limitations of developed models.

In urban traffic, driving behavior is highly dependent on the situation, which introduces even more complexity to both modeling and evaluation. For the resulting model behavior, provided by a driver model, the planning or prediction module can be described by the spatiotemporal movement, the trajectory. However, current metrics for assessing the quality of trajectories rarely consider situational context and are often bound to specific ground truth (GT) data for comparison. Common evaluation strategies for trajectory prediction models, for example, usually rely on spatiotemporal distance measures to compare GT and artificially generated trajectories for quantifying model performance [6,7]. Previously published research identified cases in which behavior deviates from the real trajectory but is still plausible [8]. In some cases, for example, the error value was large because the model chose a longer time gap than the individual human in a right-of-way situation. Both trajectories did not lead to a collision and were not critical. Thus, when comparing artificial trajectories to any individual human-driven trajectory, the result may show large error values but be still plausible, and vice versa. This can be remedied by evaluating a trajectory detached from individual behavior using a general metric that provides insight into how the artificial trajectory fits within the range of human behavior in similar situations.

In summary, current research is intensively addressing the problem of developing driver models, planning, and prediction modules for complex urban situations, whereas the challenge of adequately evaluating the results is rarely considered.

Therefore, this paper aims at creating a plausibility metric that can be applied to driver models and sub-modules by evaluating the human-likeness of trajectories while considering the situational context. In the following paper, human-likeness is understood as a significant similarity to the behavior exhibited by humans in comparable traffic situations, measured by different objective parameters. Driving behavior is assumed to be conditional and therefore dependent on various external influences [9].

This paper is organized as follows. After a broad overview of state-of-the-art approaches to evaluating driver models, the methodology is presented in Section 3. For evaluation, a multidimensional quality function, including various objective parameters to characterize human-like driving behavior, presented in Section 3.1, is formulated. Trajectories are categorized into different driving situations by assigning contextual information. The characterization of each driving situation allows for a conditional comparison of model behavior to that of humans in similar situations by selecting a subset of human traffic data showing the respective driving situation. The concept of context assignment is presented in Section 3.2. For processing trajectory and environmental data, some data processing is required, discussed in Section 3.3. Based on the behavior and context parameters, the degree of human similarity of the artificial model can be evaluated by employing statistical analysis within a quality function explained in Section 3.4. Since this method attempts to objectify human-like behavior that is difficult to attribute to any objective truth, the developed method is validated with the help of a subjective survey explained in Section 3.5. The purpose of the validation is to find out whether the human-likeness score obtained by the metric corresponds to people's subjective assessment of model behavior.

Section 4 provides specific details on the implementation, the datasets, the parametrization of the quality function, and the setting of the survey.

Section 5 discusses the results of the survey and the objective results of the method for the presented datasets are presented. To show how the presented method can be used to investigate and improve a driver model in detail, an exemplary case study is provided.

2. Related Work

Common evaluation strategies can be categorized into objective and subjective approaches. In the area of AV development, most of the objective metrics rely on the direct comparison of modeled driving data to a unique single driving sample using any distance measure [10]. Common metrics for evaluating prediction or planning frameworks employ displacement errors, measured, for example, as the distance between the actual and predicted trajectories [7]. Such metrics indicate how accurately the predicted trajectory matches the individual, human-driven trajectory. However, in the case of larger displacement errors, no conclusions can be drawn as to whether the trajectory was still plausible and only the behavior deviated with regard to safety, for example, or whether the trajectory exhibited functional problems. Therefore, in some individual cases, more sophisticated evaluation strategies are applied, e.g., taking into account functional errors such as road violations [11] or unrealistic headways [12]. To quantify the similarity between driver models and human traffic behavior in driving simulation, macroscopic analyses are performed. With the help of endurance tests, synthetic data are generated and compared with real traffic data in typical highway scenarios, such as cut-in maneuvers [13]. Typical indicators to describe human behavior in related works are average and maximal velocity, frequency and exceeding of speeding [14], acceleration and headway [15,16], as well as Time-to-Collision (TTC) and longitudinal distance [17]. Based on such parameters, the relative validity of the macroscopic behavior of driver models can be determined [18,19]. Such methods compare observed macroscopic parameters of artificial vehicles with a distribution of respective parameters among real vehicles. For comparing distributions, statistical approaches such as Kolmogorov–Smirnov in the study conducted by Wang et al. [20], or Kullback–Leibler divergence as in research by Kuefler et al. [6] are applied.

However, most approaches focus on highway traffic and do not consider contextual influences, which in turn raises doubts about the applicability of such methods to more complex urban traffic since driving behavior is affected by various external influences [21]. Subjective approaches, on the other hand, measure human-likeness using questionnaires, interviews, or surveys that automatically consider behavior within an individual context. The underlying assumption of such methods is that either what is perceived as real or can not be distinguished from artificial behavior defines human-likeness. Y. Zhang et al., for example, adapted the Turing test and asked participants to classify the driving behavior of another vehicle into either artificial or human-driven [22]. Similarly, Dumbuya et al. asked subjects to rate how realistic they perceived a drive completed by different driver models and how likely it was that the drive was conducted by a real human driver [23]. Further research investigates the human-likeness of driver models and the extent to which the perceived realism of a VE is affected by the behavior [1,24]. Since subjective methods always require an experiment involving participants, such methods are not suitable for an iterative development process due to the effort associated with the evaluation.

In summary, current evaluation strategies either do not consider contextual information, focus on macroscopic assessments, or are associated with a high implementation effort and are therefore not scaleable or inappropriate for quantifying human-like behavior in urban traffic.

3. Method

The core idea in this paper is to develop a metric that can objectively quantify the plausibility, and thus the human-likeness, of an artificially generated trajectory. The method is required to be transferable and suitable for complex urban traffic. The metric obtains artificial driving data (trajectories) in combination with situational information describing the context, e.g., whether the vehicle is currently yielding the right-of-way to someone. The metric returns a summarized human-likeness score and a multidimensional quality function incorporating various parameters characterizing behavioral plausibility, as shown in Figure 1.

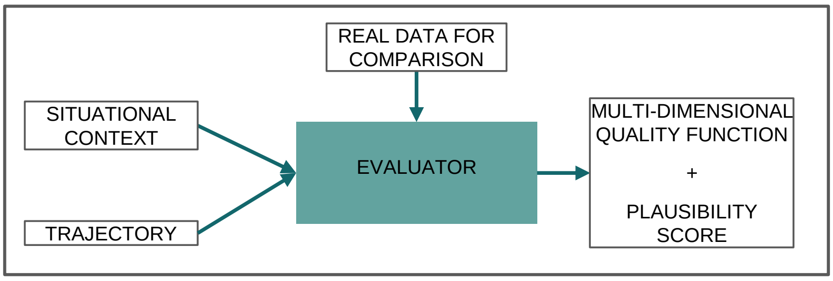

Figure 1. The general idea of evaluating model behavior within situational context.

The multidimensional quality function is composed of functional, dynamic, and interaction-related parameters. All parameters are checked against human behavior in similar situations and can be weighted to be adaptable to different applications and requirements. Finally, all weighted checks are combined into one human-likeness score. The value ranges to either pass or fail a check for each parameter are extracted from real traffic data, and thresholds can be assigned to narrow the desired range of human similarity based on application-specific requirements, such as safety. This modular design is intended to provide high scalability of the method, allowing developers to choose which parameters to consider and how to weight them, depending on the application. For example, the evaluation of a trajectory planner might require a higher level of safety in behavior compared to artificial road users in driving simulation. This work focuses on interactive situations at urban intersections, as this is where the largest gap in state-of-the-art evaluation strategies has been identified.

Context parameters are assigned to identify an individual situation in which the artificial driving data were generated. Based on knowledge of the given situation, a subset of similar situations can be extracted from real traffic data. In this way, artificial behavior can be compared with human behavior under similar conditions by considering the context. Instead of evaluating longitudinal acceleration at a macroscopic level, this method allows, for example, a comparison of acceleration measured when approaching an intersection with human acceleration values in such situations.

In order to investigate the extent to which the proposed metric is able to replace the subjective evaluation of a human assessing artificial behavior, the method is validated through a subjective survey. During the survey, participants are asked to rate the human-likeness of drivers in different situations, without knowing whether the vehicle was driven by a human or by a model. Based on the participants' ratings and the automatically calculated objective human-likeness score, the proposed metric can be validated. The present concept can be divided into the following steps: specification of parameters to characterize human driving behavior, identification of context-based similarity of situations, preparing databases, formulation of the quality function, and validation, as shown in Figure 2.

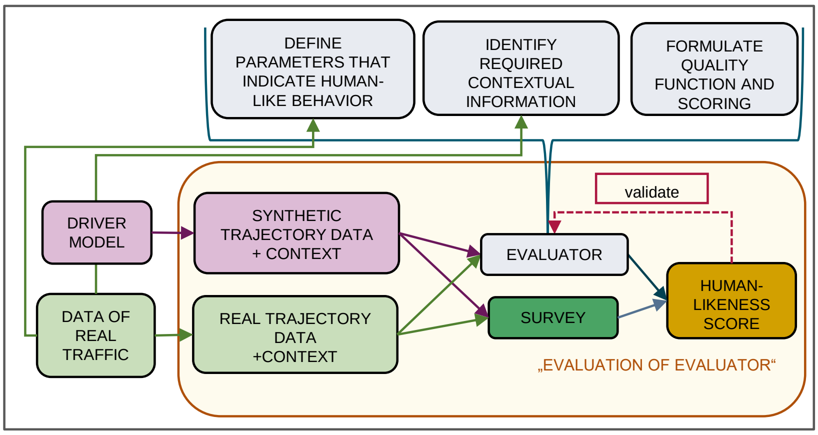
DEFINE PARAMETERS THAT INDICATE HUMAN-LIKE BEHAVIOR
IDENTIFY REQUIRED CONTEXTUAL INFORMATION
FORMULATE QUALITY FUNCTION AND SCORING
validate
DRIVER MODEL
SYNTHETIC TRAJECTORY DATA + CONTEXT
EVALUATOR
SURVEY
HUMAN-LIKENESS SCORE
DATA OF REAL TRAFFIC
REAL TRAJECTORY DATA +CONTEXT
„EVALUATION OF EVALUATOR“

Figure 2. Overview of the entire toolchain and concept.

3.1. Parameter Specification

In the first step, parameters for the evaluation of human-like driving behavior have to be defined. Inspired by literature, Table 1 provides an overview of potential parameters to characterize human-like driving behavior. The parameters aim to cover the following categories:

- Functional: is there a collision, or does the trajectory stay within the driveable area?
- Dynamic values: is motion measured by acceleration, jerk, and velocity in a human-like range?
- Interactive: are cautiousness and criticality in interactive situations, measured by time gaps and distances, in a plausible range?

Table 1. Overview parameters for describing human-likeness and plausibility of behavior.

Category	Parameter	References
Functional	Road violation	[7,11,25]
	Collision check	[26]
Dynamic	Lateral Velocity	[3]
	Longitudinal Velocity	[3,27–30]
	Lateral Acceleration	[3,27,29]
	Longitudinal Acceleration	[3,27,29,30]
	Longitudinal Jerk	[29]
Interaction	Relative Velocity	[3,27,29,30]
	Distance to the partner	[3,29]
	Time-to-Collision (TTC)	[28,29]
	Time Exposed Time-to-Collision (TET)	[28]
	Max. Value for critical time gap when interacting	[31]
	Post Encroachment Time (PET)	[28]

The parameters can be calculated for all samples of real and artificially generated behavioral data, provided that the spatiotemporal motion, road user classification, and information about the static environment, i.e., the map, are available. Algorithms to calculate interaction-related information are based on a fusion of map and time-series data according to our previous work [8].

3.2. Identification of Context-Based Similarity of Situations

In order to compare artificial behavior with the behavior of humans in similar situations, contextual information must be assigned to the data. Inspired by Scholtes et al. [32], a multi-layer approach to describe urban scenarios is used to create the basis for contextual parameters, shown in Table 2. The parameters were inspired by Schlote's approach in combination with the real traffic database, which shows right-of-way-controlled intersections. For an extension of the methodology to other traffic scenarios, additional context parameters might be required. Algorithms to extract context information are based on a fusion of map and time-series data according to our previous work [8]. Contextual parameters will be assigned to all samples in all databases.

Table 2. Overview context parameters for distinguishing scenarios sorted by priority order (P) to identify similarity.

P	Scenario	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
1	Infrastructure maneuver: relation to intersection	Before	Just Before	Inside	Just After	After
2	Infrastructure maneuver: lane turn direction	Left	Right	Straight	-	-
3	Vehicle state maneuver	Acc	Dcc	Steady	Stop	-
4	Object-related maneuver	Following	Waiting for gap	Approaching	Waiting queue	-
5	Number of legs	Three	Four	-	-	-
6	Number of interactive vehicles	Zero	One	Multiple	-	-
7	Number of interactive VRU	Zero	One	Multiple	-	-
8	Intersection density	Low	Moderate	High	-	-
9	Right-of-way relationship	Giving	Receiving	-	-	-
10	Closest interacting vehicle class	Car	Truck / Bus	-	-	-
11	Closest interacting VRU class	Bicycle	Pedestrian	-	-	-

3.3. Preparing the Database

To effectively compare real and artificial driving behaviors, it is crucial to analyze them under similar situational conditions. The situation is described by the context parameters shown in Table 2. Based on these parameters, a subset of real data can be selected. The subset of the real traffic database forming the basis for the comparison is required to contain sufficient samples to be comparable. The time-series data are aggregated into sequences of one-second time windows to determine the context. Context information is also used to select the parameters to be considered, since, for example, calculating *PET* in the absence of interaction partners would not be reasonable. A defined threshold is used to determine if the number of GT data is sufficient for comparison. If the number of remaining samples in the GT subset is insufficient, the level of abstraction is increased and context-describing parameters are gradually removed from the filtering. The priority order of context parameters for increasing abstraction is shown in Table 2 in column *P*. The priority order was determined empirically by expert knowledge and scanning the real traffic data. Knowledge of the number of matching parameters is later used to provide additional information about the certainty of the comparison. For this purpose, the Jaccard Index is used to quantify the similarity of subsets [33].

3.4. Quality Function Formulation

Once all behavioral parameters have been calculated and a subset of real data is extracted, the comparison of driving behavior can be conducted. In order to select an appropriate statistical test to measure the difference between real and artificial behavior, the underlying statistical distribution for each parameter under consideration must be examined first. For the proposed metric, the similarity of the distribution of driving behavior parameters of the artificial vehicle to the driving behavior of real vehicles in the same context is measured by using the Kolmogorov–Smirnov two-sample test method [34]. In addition, the extremes of some driving parameters are evaluated to verify that driving behavior exhibited by artificial vehicles lies within the limits defined by the minimum and maximum values obtained from real drivers in matching scenarios. The following parameters were selected for additional extrema evaluation: maximum longitudinal velocity, maximum lateral velocity, maximum longitudinal acceleration, minimum longitudinal acceleration, minimum distance to partners, and maximal critical time gap. By assigning thresholds, all individual parts of the metric either pass or fail in human similarity, are weighted by w_i, and finally calculated into a human similarity score according to Equation (1). Inspired by the parameters identified in the literature, presented in Table 1, the metric includes the parameter checks listed in Table 3.

$$score = \frac{\sum(w_i + pass\ check_i)}{\sum(w_i + total\ check_i)} * 100 \tag{1}$$

Please note that not all parameters are suitable for all driving situations; for example, if no vehicle interacts when turning, no time value can be calculated for the gap acceptance. The unavailability of such parameters was assumed as *pass* when calculating the final score.

Table 3. Thresholds for measuring human-likeness for different parameters extracted from inD data [35]. Distributions of velocities, accelerations, and jerk are compared using KS statistics. Percent ratios are assigned for maximum values and raw thresholds are assigned for context-free parameters. *Parameters marked with * are calculated context-free.*

Name of Parameter Check	Initial	Fine-Tuned
KS Longitudinal Vel.	>0.993648	>0.668406
KS Lateral Vel.	>0.952358	>0.624929
KS Longitudinal Acc.	>0.780503	>0.559198
KS Lateral Acc.	>0.926266	>0.647957
KS Jerk	>0.685024	>0.511960
Max. Longitudinal Vel.	<66.67%	<66.67%
Max. Lateral Vel.	<73.33%	<83.13%
Max. Longitudinal Acc.	<64.00%	<71.80%
Min. Longitudinal Acc.	<78.00%	<72.00%
Min. Distance to partners	<84.00%	<93.60%
PET *	<0.64 s	<0.50 s
TET *	>4.96 s	>3.90 s
Max. Critical Time Gap *	>6.98 s	>5.40 s

3.5. Strategy for Validating the Method

Since the proposed metric attempts to quantify the complex construct of human-like behavior in the absence of any official objective truth, validation is required. The validation of the method explores whether the human-like driving scores obtained by the metric correlate with people's subjective ratings. The strategy is based on the assumptions that, first, human-likeness is defined by the ability of subjects to distinguish between artificial and human behavior and second, that if a correlation is found between scores obtained by the metric and subjective ratings, validation demonstrates the ability of the metric to objectively quantify human-likeness. The subjective experiment was inspired by the work of Zhang et al. [22], who adjusted the Turing Test to quantify the human-likeness of their proposed methodology in a driving simulator. In order to validate the metric, participants are asked to assess the behavior of vehicles in short videos without knowing if one marked subject vehicle is driven by a real human or an artificial driver. The videos are designed to remove any indication of whether the behavior comes from artificial or real data. After each video, participants were asked to rate the behavior, of one marked vehicle, by the following scale: 1: Completely artificial driving; 2: Somewhat artificial driving; 3: Not sure; 4: Somewhat human-like driving; 5: Completely human-like driving. Based on this, it can be investigated whether the scores given by humans correlate with those of the proposed metric. In addition, it can be evaluated to what extent participants are able to distinguish between artificial and human behavior. For further validation of the method, the metric is applied to some real driving and artificial driving data, assuming that the human-likeness score of real data is significantly higher compared to synthetic behavioral data. The metric includes several aspects that can be tuned, such as how narrowly the range of human similarity is defined or how individual parameters are weighted for the final score. The insights from the survey provide a basis for tuning the metric toward how people would distinguish between artificial and real.

4. Implementation

4.1. Used Databases and the Exemplary Driver Model

For representing real human driving behavior, the open-source dataset inD (https://ind-dataset.com/, accessed on 1 February 2023) was selected, which provides recordings of four German unsignalized intersections from a birds-eye perspective shown in Figure 3 (right) [35]. The dataset is composed of *tracks data* describing the spatiotemporal motion of all road users, *meta data* describing dimensions and classifications of road users, and the respective openDRIVE map for the location. For creating the human behavior database, all recordings except recording 12 were selected, whereby this recording was retained for validation purposes as described in Section 3.5 (as referred to inD12). For testing the proposed metric, artificial driving data were created on two intersections, which are shown in Figure 3 (left). For testing on a large scale of interactive situations, data were created with the help of the simulation framework *Spider* at BMW [36]. Driving behavior is generated by the BMW proprietary driver model TRM. The model follows a hierarchical and heuristic structure, categorizing the driving task into maneuvers like car-following, changing lanes, overtaking, stopping due to an obstacle or red lights, slowing down for speed limits, or curved roads. Based on the situational and environmental information, multiple maneuver evaluators identify their need for action in each time step. The chosen maneuver is subsequently performed by applying the respective motion model such as the Wiedemann-following model for car-following [1]. Behavioral and contextual parameters are computed according to Section 3 for all datasets employing the fusion and interaction identification algorithms presented in our previous work [8]. The algorithms first fuse the time-series and map data by assigning all vehicles to the respective lane they are driving on. Based on this assignment and the knowledge from the map, interactions can be identified and described by semantic parameters.

4.2. Metric Formulation and Thresholds

By assigning thresholds and weights, all individual parts of the metric can be combined into one human-likeness score according to Equation (1). For *PET*, *TET*, and *Max. Critical Gap*, global thresholds for human similarity were derived from the real traffic database since not enough situational samples could have been extracted for contextual comparison due to the heuristics applied to calculate these parameters. All other parameters could be compared considering the situational context, i.e., in comparison to the behavior of real drivers in comparable situations. Accordingly, threshold values represent the limits for statistical similarity. The human-likeness scores for the two synthetic locations and the retained real recording (inD12) were calculated. Since the driver models used in the synthetic data showed quite high scores, thresholds for each parameter were further narrowed to fulfill the assumption that the overall results of the real test data would significantly exceed those of the synthetic data. The tuning resulted in a score of 89.62% for the real dataset (inD12), 77.31% for interSection 1 (synthetic), and 77.87% for interSection 2 (synthetic). Initially extracted and tuned thresholds for all parameters are shown in Table 3.

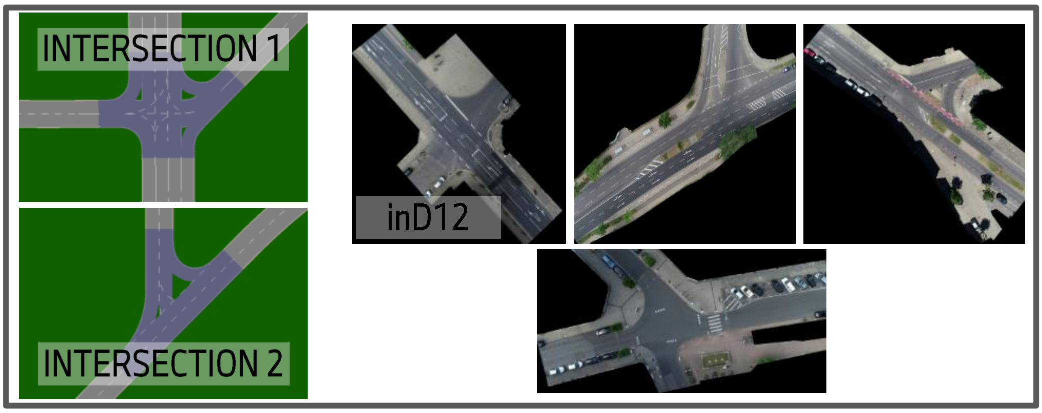

Figure 3. Locations for data gathering—**Left:** synthetic intersections for creating artificial driving behavior. **Right:** locations from inD Data [35].

4.3. Survey for Validation

The survey was conducted online and involved 23 participants rating the behavior of vehicles in 12 videos. Four videos showed real driving behavior and eight artificial behavior. Real scenarios were extracted from the inD dataset of recording 12, whereby artificial behavior was created on the two synthetic intersections as described in Section 4.1. The order of the videos was randomized to eliminate the possibility of order effect bias. In each video, the vehicle to assess was marked red whereas all other vehicles were blue, as shown exemplarily in Figure 4. The visualization was abstracted to eliminate any indicators that might help distinguish between real and artificial. After each video, participants were asked to rate on a five-point scale whether they perceived the red vehicle's behavior as real or artificial.

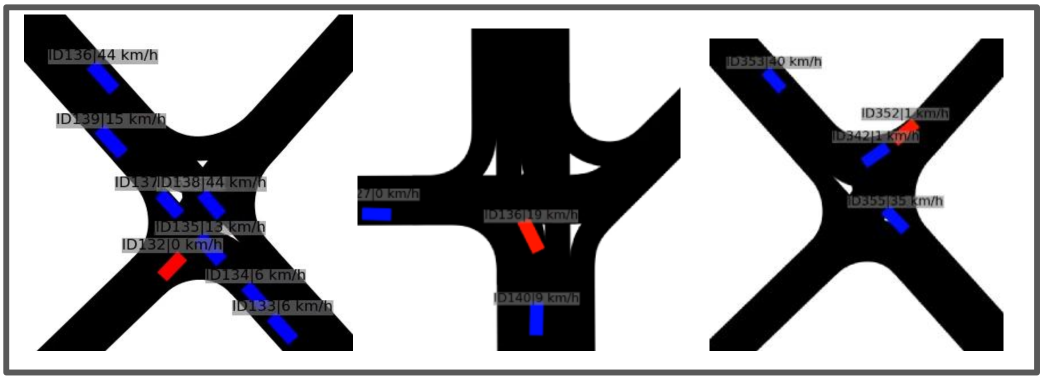

Figure 4. Exemplary screenshots of a video shown to participants for rating human-likeness of a subject vehicle (marked red).

5. Results

In the following section, the results of the survey, providing subjective assessments of human-likeness, are compared with the objective scores obtained by the proposed metric. In order to apply the proposed method to a broader range of samples, the human-likeness score is additionally computed for the datasets described in Section 4.1. Finally, a case study is presented as exemplary, showing how to apply the proposed method for model improvement.

5.1. Objective Metric Results versus Subjective Human Ratings

When investigating the results of the survey, two aspects are of interest. First, subjects were able to distinguish between real and artificial behavior, and second, participants' subjective ratings correlated with objective scores calculated by the proposed metric. The mean value of the Turing test (6.37), indicated that participants' ability to distinguish between artificial and real drivers is only slightly higher than random responses or the result (exactly 6.0) when selecting "Not sure" for all vehicles. Figure 5 visualizes the subjective ratings for all test vehicles shown during the survey associated with the objective scores calculated by the proposed metric. Real vehicles are green, whereas artificial vehicles are colored red. The y-axis shows the rating scale presented during the survey. The blue value above each vehicle rating describes the objective human-likeness score obtained by the proposed metric. During the survey, artificial vehicles were selected that exhibited high and low levels of human-likeness. Furthermore, for the real vehicles, samples with more and less objective human-likeness scores were selected to determine if the metric is able to detect both good and bad results.

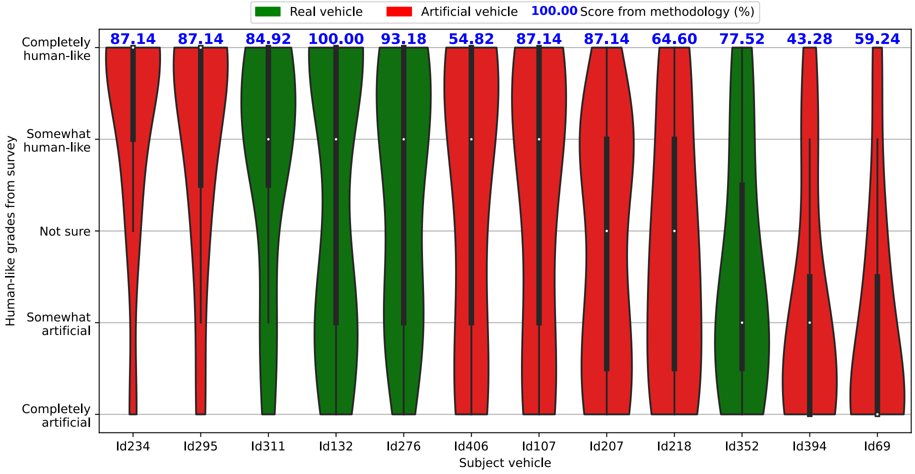

Figure 5. Subjective human-likeness rating obtained from participants (y-axis) associated with objective human-likeness scores obtained by the proposed metric (blue value above). Vehicles are sorted by the average rating assigned by participants, in descending order from left to right.

Comparing the objective scores from the metric to participants' subjective ratings from the survey, a positive correlation could be observed with a Spearman correlation coefficient of 0.62 and a p-value of 0.030, and a Pearson correlation coefficient of 0.65 and a p-value of 0.023 shown in Figure 6. The p-values indicate that there is a moderately monotonic positive relationship at the 97% confidence level and a moderately linear positive relationship at the 97% confidence level, which can be considered statistically significant.

Based on the insights into which vehicles were rated as more human-like by the participants and the multidimensional quality function, the individual parameters could be analyzed in terms of the extent to which they contribute to the decision for real or artificial behavior. Table 4 provides the results of this analysis using the Spearman correlation, which is further transformed into a weighting of the individual parameters with the aim of reflecting how people prioritize differences in the individual behavioral parameters.

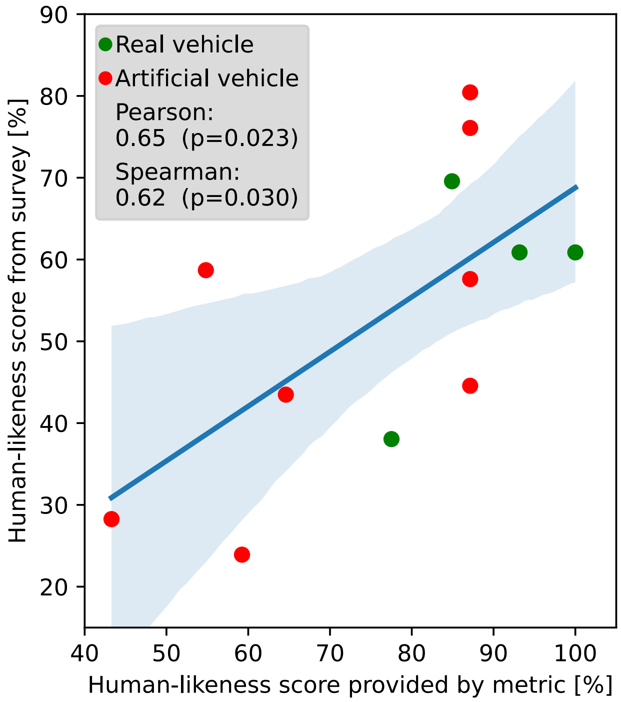

Figure 6. Relationship between the fine-tuned objective human-like driving behavior scores provided by the proposed methodology and subjective ratings of participants during the survey.

Table 4. Spearman correlation between the parameters and average human-like driving behavior score from the survey, with correlations converted into weights.

Parameter	**Spearman Correlation**	***p*-Value**	**Conversion to Weight**
KS Longitudinal Vel.	0.073555	0.820285	0.016269
KS Lateral Vel.	0.178634	0.578567	0.03951
KS Longitudinal Acc.	−0.30823	0.329698	0.068174
KS Lateral Acc.	0.021016	0.948312	0.004648
KS Jerk	−0.51839	0.084229	0.114656
Max. Longitudinal Acc. ratio	−0.42732	0.165877	0.094514
Min. Longitudinal Acc. ratio	−0.50794	0.0918	0.112346
Max. Longitudinal Vel. ratio	−0.12151	0.706773	0.026876
Max. Lateral Vel. ratio	0.309234	0.328046	0.068396
Min. Distance to partners ratio	−0.3632	0.245869	0.080333
Max. Critical Time Gap (s)	−0.94286	0.004805	0.208539
TET (s)	−0.52179	0.288343	0.115409
PET (s)	−0.22755	0.587845	0.050329

5.2. Human-Likeness of Investigated Datasets

As described in Section 3.5, the metric is applied to retained real driving (inD12) and artificial driving data, assuming that the human-likeness score of real vehicles should be significantly higher compared to synthetic behavioral. The used database is described in Section 4.1 and results of the objective scores obtained by the metric are shown in Figure 7. The scores were calculated once with the initial thresholds and without weighting according to Table 3 (right), and once with the tuned quality function, incorporating weights from the survey and adjusted thresholds from Table 3 (left). First of all, the difference in those two figures demonstrates the sensitivity of results to weights and thresholds within the quality function. Regarding the initial setting, only the comparison of real data and artificial behavior of interSection 2 showed significant differences in the human-likeness scores when using the Mann–Whitney U test (U = 82,727.0, p_{value} = 4.06×10^{-13}), whereas the comparison of behavior on interSection 1 compared to real humans showed no significant difference (U = 58,561.0, $p_{value} = 0.78$). This can be explained by the quite far-developed driver model, which was used to create the artificial data. Therefore, without weighting or tightening the thresholds of the quality function, only *harsh* outliers of behavior can be detected. When using the tuned quality function, (Figure 7 left) a clear difference between real and artificial behavior could be measured (human-like grades of inD12 vs. interSection 1: U = 103,568.5, p_{value} = 4.34×10^{-68}; human-like grades of inD12 vs. interSection 2: U = 113,910.0, p_{value} = 1.54×10^{-60}).

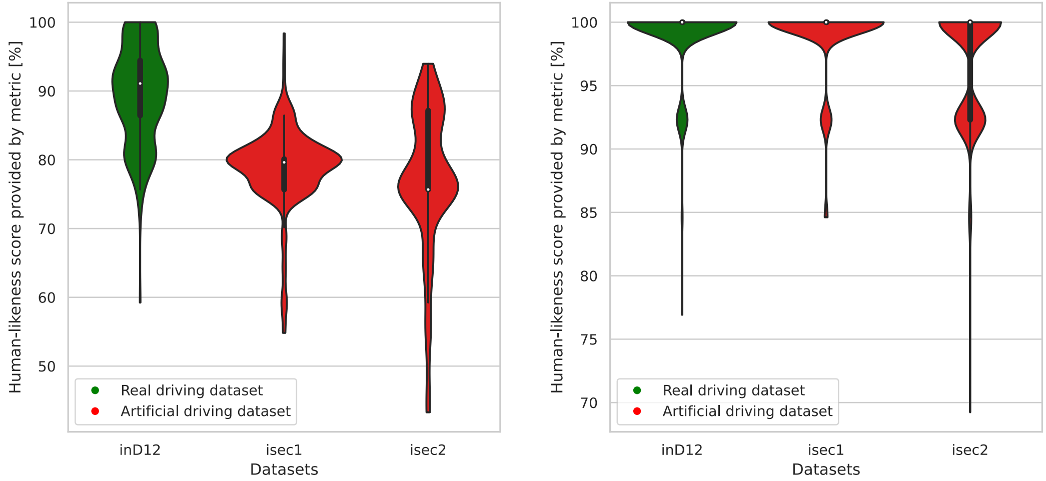

Figure 7. Results for human-like scores for real and synthetic datasets: with tuned thresholds and weights (**left**) and initial setting (**right**).

5.3. Case Study: Exemplary Application of the Method for Model Improvement

The proposed method is characterized by two main aspects. First, by using various parameters involving functional, dynamic, and interactive behavior, the multi-modality of driving behavior is objectively measurable at different levels. Secondly, behavior is assumed to be conditional and compared within a situational context instead of comparing parameters on a macroscopic level. Therefore, this approach provides a high level of transparency and enables targeted model improvement. How the proposed method can be used for model improvement is presented in the following case study.

The overall scores measured for the artificial driver model in Figure 7 show a mean human-likeness score of 78.89% showing a variance of 33.33 on interSection 1. The scoring method enabled for quantifying model behavior among multiple situations measured by various parameters. Based on the proposed method, we are able to address the following questions to enable model improvement:

- In which scenarios does the model show less human-like behavior?
- Which parameters mostly fail when comparing the model to human behavior?
- Why do those parameters fail; how does the distribution of parameters distinguish when comparing model behavior and human behavior?

Based on the correlation analysis presented in Table 4, the critical time gap was found to have a high negative correlation with the subjective ratings of human-like driving obtained through the survey. Therefore, the critical time gap value is further investigated. In the real traffic data, the value was determined to be 6.98 seconds, whereas the same parameter in the synthetic data was determined to be 10.05 seconds at InterSection 1 and 9.98 seconds at InterSection 2. This shows a significant difference in behavior and needs to be improved in the driver model. To further investigate in which situations model behavior mostly differs from that of humans, the distribution of failing scenario parameters is investigated. Considering all context parameters and the distribution of failed parameter checks, the following conclusions could be drawn. Vehicles are more likely to fail:

- In the maneuver states: approaching, accelerating, decelerating, or stopping;
- Giving right-of-way;
- In the object-related maneuver: approaching or waiting for a gap;
- Interacting with fewer partner vehicles;
- Driving in lower intersection density.

The calculated behavior parameters of the synthetic data are analyzed to identify which behavior parameters mostly fail. Illustrated in Figure 8, the five most often failing behavior parameter checks were identified to be: KS Longitudinal Vel., KS Lateral Vel., KS Longitudinal Acc., KS Lateral Acc., and KS Jerk.

In the next step, the value distributions of the extracted parameters can be analyzed. For this purpose, the distribution of the dynamic parameters extracted from real data is compared with the dynamic parameters of the artificial samples showing a human-likeness score of less than 70%. The distributions are extracted on a macroscopic level and situationally under the scenario conditions identified above as causing most of the failures. All distributions are summarized in Table 5.

Table 5. Analysis of the value distribution for dynamic parameters for model behavior and real humans on a macroscopic level (left) and scenario-specific under situational conditions (right).

Parameter		**Real Data: Macroscopic**	**Syn Data: Macroscopic**	**Real Data: Situational**	**Syn Data: Situational**
Longitudinal Vel. [m/s]	mean	7.18	3.38	5.71	3.48
	std	5.24	3.92	4.68	3.48
	min	−4.35	0.00	−0.21	0.00
	max	27.48	19.05	20.52	12.34
Lateral Vel. [m/s]	mean	0.04	0.03	0.04	0.14
	std	0.19	0.15	0.26	0.25
	min	−2.87	−0.77	−0.99	−0.07
	max	10.89	1.45	1.07	0.85
Longitudinal Acc. [m^2/s]	mean	0.06	0.06	−0.08	−0.47
	std	0.87	1.02	1.01	2.72
	min	−6.25	−9.58	−4.34	−9.58
	max	6.54	5.25	4.72	2.61
Lateral Acc. [m^2/s]	mean	−0.04	0.13	−0.08	0.65
	std	0.61	0.78	0.89	1.14
	min	−5.57	−3.97	−4.62	−0.20
	max	4.98	4.78	4.12	3.27
Jerk [m^3/s]	mean	−0.05	0.03	0.06	0.14
	std	0.84	2.91	1.05	12.20
	min	−20.35	−21.40	−6.96	−18.17
	max	16.16	220.04	16.16	220.04

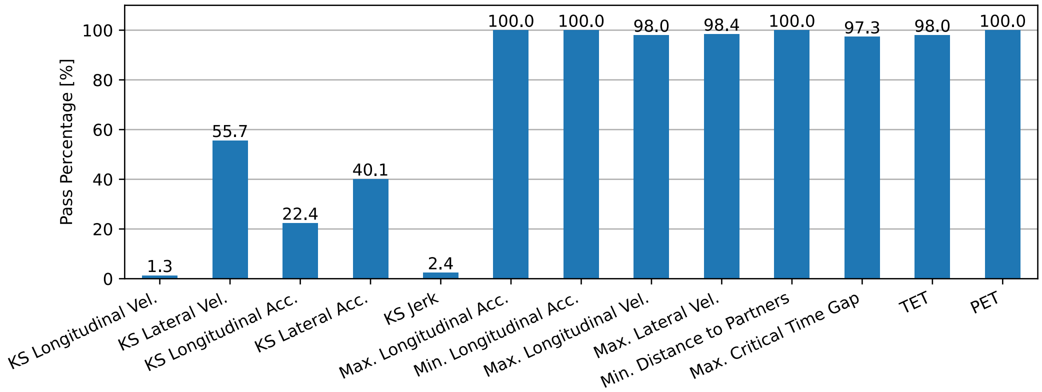

Figure 8. Analysis to identify which parameters mostly fail when comparing artificial and human behavior.

As shown in Table 5, the jerk parameter stands out in particular. The maximal values measured in artificial behavior are much larger compared to those observed by human drivers. It should be noted that the jerk values observed by real humans are determined by tracking algorithms that process video data from drones. Since the open-source real traffic dataset is preprocessed, it is not known to what extent the values in this data are smoothed. However, the jerk values of the driver model show significantly high maximum values, which should be further investigated. Therefore, some trajectories were extracted and analyzed. Figure 9 shows some example trajectories that illustrate the *jerking problem* that occurs when switching between driving maneuvers. Since the driver model used is based on heuristic decision-making, the temporal behavior and motion are more discrete compared to humans. However, from a subjective visual point of view, the *jerking problem* is not perceptible, as shown by the survey, and therefore not critical for driver models in the context of DiL traffic simulation.

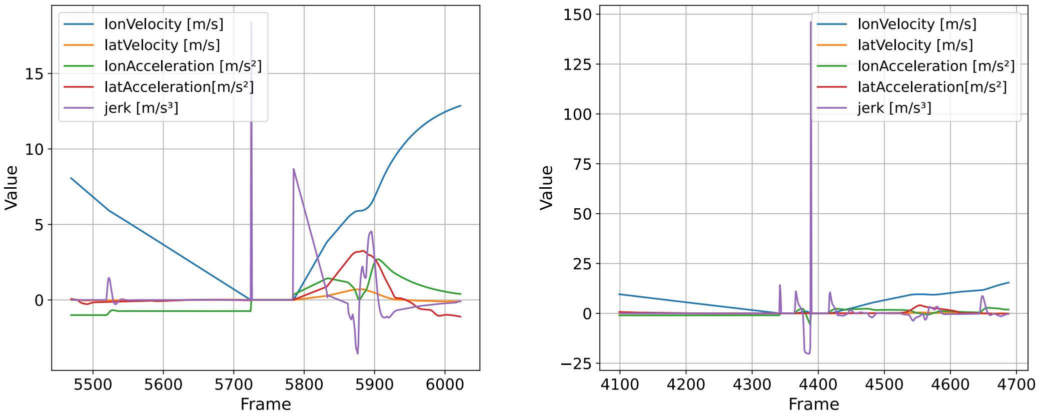

Figure 9. Exemplary trajectory of two vehicles showing significant high jerk values when approaching an intersection.

Further interesting insights are provided by the comparison between macroscopic behavior parameter distribution (left) and situational behavior (right) in Table 5. When considering longitudinal acceleration, for example, one can observe that distributions are in line with human value ranges in a macroscopic perspective but not when comparing the parameter situationally . This shows that macroscopic comparison can result in misleading conclusions regarding the extent of human-like behavior.

In summary, the case study demonstrates the potential of the method to reveal model weaknesses and enable better model parameterization.

6. Conclusions

In the present work, a new method was introduced to objectively measure the degree of human-likeness of artificial driver models. Driving behavior is characterized by various parameters, which are then compared to the behavior of humans in real traffic. Using statistical analysis in the context of a quality function, a final human-likeness score can be computed for each trajectory. Since behavior in urban scenarios is influenced by various factors and assumed to be conditional, the situational context for each trajectory in real and artificial data is characterized by automatically computed context parameters that aim to distinguish between different situations that may occur in urban traffic. Thus, a subset of GT data showing human behavior under similar conditions can be used to compare behavioral parameters.

Since there is no clear definition of human-likeness, we investigated the ability of the proposed method to reflect the subjective ratings of humans when assessing driving behavior in a survey inspired by the Turing test. Results of the survey showed a significant correlation between the scores calculated by the proposed metric and the scores assigned by participants. In addition, the survey provided interesting insights into which parameters contribute most to the distinction between artificial and human driving behavior. These findings were used to parameterize the quality function and provided valuable insights into specific weaknesses of the used driver models.

Evaluation of large datasets has shown that the proposed metric has the potential to evaluate models or sub-modules in a wide range of situations, which is crucial for developing reliable solutions for urban traffic. The case study exemplified how the proposed metric can be used for detailed model evaluation and targeted model improvement. The modular structure of the metric allows models to be evaluated according to application-specific requirements. In a driving simulation, for example, the priority is more on human-like behavior than on rule compliance and safety. In contrast, when evaluating a trajectory planner for AVs in real traffic, the focus might be more on non-critical behavior, accepting a lower level of aggressiveness during interactions. By weighting and narrowing the thresholds for individual parameters, the method can be used for a broad range of applications.

7. Limitations and Future Work

The proposed method presents a first attempt to objectify human-like driving behavior, taking into account the situational context. The multi-modality of human behavior is mapped into individual parameters, which are then statistically evaluated by assigning thresholds for “pass” or “fail” and potentially additional weights. Weighting and threshold assignment have a significant impact on the final metric score, resulting in high sensitivity to individual tuning. In the future, more parameters can be added to better account for the multimodality of human driving behavior, and extensive surveys could provide a basis for fine-tuning when focusing on replacing subjective evaluations of humans. Instead of the binary approach of either passing or failing a parameter, a more sophisticated concept could be used whereby the range of human-likeness is discretized into bins for each parameter.

In order to determine similar matching situations from real driving behavior, context parameters are assigned to the data. The heuristics used in this paper correspond to the scenarios encountered in real traffic data showing unsignalized intersections. To extend the proposed metric for other traffic scenarios, additional context parameters should be considered. The algorithms for computing context parameters, such as the number of interaction partners and related parameters, are based on heuristics. Such heuristics, of course, do not guarantee that meaningful results are provided across the entire diversity of interpersonal situations occurring in urban traffic. Some parameters, such as *PET*, were calculated only for situations in which reliable results could be guaranteed, resulting in a significant reduction in samples. Future work will attempt to increase the validity of heuristics to allow a contextual comparison for all parameters.

A simple abstraction strategy was used to measure the similarity between situations in artificial and real data. State-of-the-art techniques offer alternative approaches for measuring similarity between datasets, such as those presented by Heuer [37], which could be used in the future.

Author Contributions: Conceptualization T.R. and T.H.; Methodology, T.R. and T.H.; Writing—original draft, T.R.; Writing—review & editing, T.B., M.B., S.M., A.A.S. and M.S.; Supervision, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Written informed consent has been obtained from the patients to publish results of the survey.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to information protection at BMW.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rock, T.; Bahram, M.; Himmels, C.; Marker, S. Quantifying Realistic Behaviour of Traffic Agents in Urban Driving Simulation Based on Questionnaires. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 4–9 June 2022; pp. 1675–1682.
2. Wei, J.; Dolan, J.M.; Litkouhi, B. A learning-based autonomous driver: emulate human driver's intelligence in low-speed car following. In Proceedings of the Unattended Ground, Sea, and Air Sensor Technologies and Applications XII, SPIE, Orlando, FL, USA, 5–9 April 2010; Volume 7693, pp. 93–104.
3. Sharath, M.N.; Velaga, N.R.; Quddus, M.A. 2-dimensional human-like driver model for autonomous vehicles in mixed traffic. *IET Intell. Transp. Syst.* **2020**, *14*, 1913–1922. [CrossRef]
4. Hang, P.; Lv, C.; Xing, Y.; Huang, C.; Hu, Z. Human-like decision making for autonomous driving: A noncooperative game theoretic approach. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 2076–2087. [CrossRef]
5. Lindorfer, M.; Mecklenbraeuker, C.F.; Ostermayer, G. Modeling the imperfect driver: Incorporating human factors in a microscopic traffic model. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 2856–2870. [CrossRef]
6. Kuefler, A.; Morton, J.; Wheeler, T.; Kochenderfer, M. Imitating driver behavior with generative adversarial networks. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017.
7. Bahari, M.; Saadatnejad, S.; Rahimi, A.; Shaverdikondori, M.; Shahidzadeh, A.H.; Moosavi-Dezfooli, S.M.; Alahi, A. Vehicle trajectory prediction works, but not everywhere. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 17102–17112. [CrossRef]
8. Rock, T.; Marker, S.; Bleher, T.; Bahram, M. Data-Driven Prediction of Other Road Users' Intention for Better Scene Understanding in Traffic Agents. In Proceedings of the Driving Simulation Conference 2022 Europe VR, Strasbourg, France, 15–16 September 2022; Kemeny, A.; Chardonnet, J.R.; Colombet, F., Eds.; Driving Simulation Association: Strasbourg, France, 2022; pp. 9–16.
9. Wang, W.; Wang, L.; Zhang, C.; Liu, C.; Sun, L. Social Interactions for Autonomous Driving: A Review and Perspectives. *Found. Trends Robot* **2022**, *10*, 198–376. [CrossRef]
10. Zhu, M.; Wang, X.; Wang, Y. Human-like autonomous car-following model with deep reinforcement learning. *Transp. Res. Part C Emerg. Technol.* **2018**, *97*, 348–368. [CrossRef]
11. Schäfer, M.; Zhao, K.; Bühren, M.; Kummert, A. Context-Aware Scene Prediction Network (CASPNet). In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 3970–3977. [CrossRef]
12. Su, J.; Beling, P.A.; Guo, R.; Han, K. Graph convolution networks for probabilistic modeling of driving acceleration. In Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; pp. 1–8.
13. Fries, A.; Fahrenkrog, F.; Donauer, K.; Mai, M.; Raisch, F. Driver Behavior Model for the Safety Assessment of Automated Driving. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 4–9 June 2022; pp. 1669–1674. [CrossRef]
14. Kim, H.; Yoon, D.; Shin, H.; Park, C.H. Driving characteristics analysis of young and middle-aged drivers. In Proceedings of the 2016 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 19–21 October 2016; pp. 864–867. [CrossRef]
15. Moertl, P.; Festl, A.; Wimmer, P.; Kaiser, C.; Stocker, A. Modelling driver styles based on driving data. In Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference, Rome, Italy, 28–30 September 2017.
16. Bae, I.; Moon, J.; Jhung, J.; Suk, H.; Kim, T.; Park, H.; Cha, J.; Kim, J.; Kim, D.; Kim, S. Self-driving like a human driver instead of a robocar: Personalized comfortable driving experience for autonomous vehicles. *arXiv* **2020**, arXiv:2001.03908.
17. Wang, W.; Cheng, Q.; Li, C.; André, D.; Jiang, X. A cross-cultural analysis of driving behavior under critical situations: A driving simulator study. *Transp. Res. Part F Traffic Psychol. Behav.* **2019**, *62*, 483–493. [CrossRef]
18. Kostikj, A.; Kjosevski, M.; Kocarev, L. Validation of a microscopic single lane urban traffic simulator. In Proceedings of the 2014 International Conference on Connected Vehicles and Expo (ICCVE), Vienna, Austria, 3–7 November 2014; pp. 850–854. [CrossRef]

19. Rudloff, C.; Schoenauer, R.; Fellendorf, M. Comparing Calibrated Shared Space Simulation Model with Real-Life Data. *Transp. Res. Rec.* **2013**, *2390*, 44–52. [CrossRef]
20. Wang, J.; Dixon, K.K.; Li, H.; Ogle, J. Normal deceleration behavior of passenger vehicles at stop sign–controlled intersections evaluated with in-vehicle Global Positioning System data. *Transp. Res. Rec.* **2005**, *1937*, 120–127. [CrossRef]
21. Marina Martinez, C.; Heucke, M.; Wang, F.Y.; Gao, B.; Cao, D. Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 666–676. [CrossRef]
22. Zhang, Y.; Hang, P.; Huang, C.; Lv, C. Human-Like Interactive Behavior Generation for Autonomous Vehicles: A Bayesian Game-Theoretic Approach with Turing Test. *Adv. Intell. Syst.* **2022**, *4*, 2100211. [CrossRef]
23. Dumbuya, A.; Booth, A.; Reed, N.; Kirkham, A.; Philpott, T.; Zhao, J.; Wood, R. Complexity of traffic interactions: improving behavioural intelligence in driving simulation scenarios. In *Complex Systems and Self-organization Modelling*; Bertelle, C., Duchamp, G.H., Kadri-Dahmani, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 201–209. [CrossRef]
24. Al-Shihabi, T.; Mourant, R.R. Toward More Realistic Driving Behavior Models for Autonomous Vehicles in Driving Simulators. *Transp. Res. Rec.* **2003**, *1843*, 41–49. [CrossRef]
25. Liu, J.; Mao, X.; Fang, Y.; Zhu, D.; Meng, M.Q. A Survey on Deep-Learning Approaches for Vehicle Trajectory Prediction in Autonomous Driving. In Proceedings of the 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), Sanya, China, 27–31 December 2021; pp. 978–985.
26. Tao, C.; Jiang, Q.; Duan, L.; Luo, P. Dynamic and Static Context-aware LSTM for Multi-agent Motion Prediction. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
27. Kesting, A.; Treiber, M.; Helbing, D. Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2010**, *368*, 4585–4605. [CrossRef] [PubMed]
28. Sharath, M.N.; Mehran, B. A Literature Review of Performance Metrics of Automated Driving Systems for On-Road Vehicles. *Front. Future Transp.* **2021**, *2*, 759125. [CrossRef]
29. Zlocki, A.; König, A.; Bock, J.; Weber, H.; Muslim, H.; Nakamura, H.; Watanabe, S.; Antona-Makoshi, J.; Taniguchi, S. Logical Scenarios Parameterization for Automated Vehicle Safety Assessment: Comparison of Deceleration and Cut-In Scenarios From Japanese and German Highways. *IEEE Access* **2022**, *10*, 26817–26829. [CrossRef]
30. Sama, K.; Morales, Y.; Liu, H.; Akai, N.; Carballo, A.; Takeuchi, E.; Takeda, K. Extracting human-like driving behaviors from expert driver data using deep learning. *IEEE Trans. Veh. Technol.* **2020**, *69*, 9315–9329. [CrossRef]
31. Raff, M.S.; Hart, J.W. *A Volume Warrant for Urban Stop Signs*; Eno Foundation for Highway Traffic Control: Saugatuck, CT, USA, 1950.
32. Scholtes, M.; Westhofen, L.; Turner, L.R.; Lotto, K.; Schuldes, M.; Weber, H.; Wagener, N.; Neurohr, C.; Bollmann, M.H.; Körtke, F.; et al. 6-layer model for a structured description and categorization of urban traffic and environment. *IEEE Access* **2021**, *9*, 59131–59147. [CrossRef]
33. Jaccard, P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 241–272.
34. Facchinetti, S. A procedure to find exact critical values of Kolmogorov–Smirnov test. *Stat. Appl. Ital. J. Appl. Stat.* **2009**, *21*, 337–359.
35. Bock, J.; Krajewski, R.; Moers, T.; Runde, S.; Vater, L.; Eckstein, L. The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020.
36. Strobl, M. *SPIDER—Das innovative Software-Framework der BMW Fahrsimulation/ SPIDER—The Innovative Software Framework of the BMW Driving Simulation*; Nr. 1745; VDI-Berichte: Düsseldorf, Germany, 2003.
37. Heuer, F.M. Scenario Generation for Testing of Automated Driving Functions based on Real Data. Master’s Thesis, Institut für Softwaretechnik und Fahrzeuginformatik, Braunschweig, Germany, 2022. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.