Scenario-based assessment of automated driving systems: How (not) to parameterize scenarios?

Erwin de Gelder[1*], Olaf Op den Camp[1]

***Abstract*— The development of Automated Driving Systems (ADSs) has advanced significantly. To enable their large-scale deployment, the United Nations Regulation 157 (UN R157) concerning the approval of Automated Lane Keeping Systems (ALKSs) has been approved in 2021. UN R157 requires an activated ALKS to avoid any collisions that are reasonably preventable and proposes a method to distinguish reasonably preventable collisions from unpreventable ones using "the simulated performance of a skilled and attentive human driver". With different driver models, benchmarks are set for ALKSs in three types of scenarios. The three types of scenarios considered in the proposed method in UN R157 assume a certain parameterization without any further consideration.**

This work investigates the parameterization of these scenarios, showing that the choice of parameterization significantly affects the simulation outcomes. By comparing real-world and parameterized scenarios, we show that the influence of parameterization depends on the scenario type, driver model, and evaluation criterion. Alternative parameterizations are proposed, leading to results the are closer to the non-parameterized scenarios in terms of recall, precision, and F1 score. The study highlights the importance of careful scenario parameterization and suggests improvements to the current UN R157 approach.

I. INTRODUCTION

Automated Driving Systems (ADSs) are expected to enhance traffic safety by eliminating human errors, providing more comfortable rides, and reduce traffic congestion [1]. Lower levels of automation, such as adaptive cruise control [2] and lane-keeping assist systems [3], are already common in modern vehicles. Initially, the deployment of higher automation levels (SAE level 3 and above [4]) was hindered by regulations requiring a human driver to be in charge, as per the Vienna Convention on Road Traffic of 1968 [5].

The World Forum for Harmonization of Vehicle Regulations approved the United Nations Regulation 157 (UN R157) in 2021 for the approval of ADSs, titled "Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping Systems (ALKSs)" [6]. This is the first regulation considering an automated system that fully takes over the driving task of a human driver for part of the ride. This regulation states that "the activated system shall not cause any collisions that are reasonably foreseeable and preventable" [6, Clause 5.1.1]. A method is proposed in [6, Annex 3, Appendix 4] to distinguish scenarios with preventable collisions from scenarios with unpreventable collisions "based on the simulated performance of a skilled and attentive human driver". Based on human driver models, benchmarks are set for ALKSs in three types of scenarios. The three types of scenarios considered in the proposed method in UN R157 assume specific parameterizations without further consideration.

In this work, we investigate the parameterization of the three types of scenarios in UN R157. While parameterizing scenarios enables statistical analyses of the performance of an ADS, the identification of potential failures of the ADS, and the upscaling of scenario-based testing, the particular choice of parameterization appears to have a significant influence on the outcome of the assessment. We show this by comparing the simulated performance of the driver models with scenarios as observed in real-world data as well as the same scenarios in a parameterized form. The results show that the choice of the parameterization has a significant influence on the outcome of the simulated performance that should not be neglected. We also show that the influence of the parameterization not only depends on the type of scenario, but also on the driver model and the test criteria that are used to evaluate the simulated performance. For each type of scenario, we propose several alternative parameterizations, and we demonstrate that other choices of scenario parameterization are leading to results that are closer, in terms of recall, precision, and F1 score, to the results with the real-world, non-parameterized scenarios.

This work is structured as follows. We first explain why and how scenarios are parameterized in Section II. Our proposed method for analyzing different parameterizations is presented in Section III. Section IV explains how we illustrate our method in a case study based on the UN R157 scenarios, with results shown in Section V. Sections VI and VII provide a discussion and conclusions, respectively.

II. BACKGROUND

As mentioned in the introduction, the choice of scenario parameterization significantly impacts assessment outcomes. This raises the question: should scenarios be parameterized? To address this, we first explain the benefits of parameterization in Section II-A, followed by a review of works on scenario parameterization for assessing ADSs in Section II-B.

The research presented in this work has been made possible by the SUNRISE project. This project is funded by the European Union's Horizon Europe Research & Innovation Actions under grant agreement No. 101069573. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

[1]TNO, Integrated Vehicle Safety, Helmond, The Netherlands

[*]Corresponding author: erwin.degelder@tno.nl

A. Why parameterizing scenarios?

Parameterizing scenarios enables comprehensive testing beyond observed road scenarios. Since the number of scenarios an ADS must be able to handle is virtually infinite, extensive testing with varied scenarios is essential. Relying solely on observed scenarios for testing would be impractical and costly as well as making the execution of all test scenarios infeasible.

Another reason to parameterize scenarios is to facilitate statistical analysis. By parameterization, we can estimate probability density functions for the parameters, allowing us to assess scenario exposure and quantify the risk associated with an ADS [7].

For assessing an ADS, it is essential to identify the scenarios it might encounter during its lifetime. A detailed representation of scenarios, with all state variables defined over time, would result in an impractically large number of scenarios. Instead of listing each scenario individually, we can use ranges of valid parameter values to define the scenarios an ADS must handle. The term "logical scenario," introduced in [8], refers to these scenario descriptions with parameter ranges. As shown in [9], [10], using parameter ranges also helps determining the bounds of reasonably foreseeable scenarios.

Due to the vast number of scenarios, assessing ADSs heavily relies on simulations. However, it is impossible to simulate every possible scenario. Consequently, research has focused on minimizing the number of simulations by targeting scenarios where the ADS exhibits critical behavior. A well-known method for this is importance sampling, which automatically selects scenarios for simulation to reduce the number of simulations needed while still providing sufficient confidence in the assessment results, e.g., see [11]–[13]. This would not be possible if scenarios would not be parameterized.

B. How to parameterize scenarios?

In UN R157, three different types of scenarios are considered: cut-in, cut-out, and Leading Vehicle Deceleration (LVD). The parameters of those scenarios are selected without further explanation. This approach is also common in many other studies, where specific parameterizations are chosen without additional justification or detailed consideration, e.g., see [9]–[13].

A common challenge with parameterizing scenarios is the so-called "curse of dimensionality", where the complexity of estimating statistics and/or performing numerical computations grows exponentially with the number of parameters. Therefore, reducing the number of parameters is desirable. Several studies focus on techniques for parameter reduction, e.g., see [14], [15]. In [14], a metric is proposed to find the optimal balance between minimizing the number of parameters for reliable statistics and maximizing parameters to reduce information loss. However, this metric does not account for the impact of parameter reduction on the simulation outcomes of these scenarios, which depends on the sensitivity of the results on the scenario parameters.

Typically, scenario parameters have a physical interpretation. However, Generative Adversarial Networks (GANs) offer a technique that does not rely on such parameters [16]. GANs have also been applied in the automotive domain [17], [18]. It is important to note that GANs still use parameters, known as "latent variables", which lack physical meaning. This essentially makes GANs another parameter reduction technique. Therefore, with GANs, it is still necessary to evaluate whether the automatically determined parameters are suitable.

III. METHOD

Due to the diverse range of scenarios, parameters relevant to one type of scenario may not be applicable to another. To differentiate between various types of scenarios, we assume that all scenarios — which are quantitative descriptions — can be categorized into one or more scenario categories, with scenarios within the same scenario category being parameterized similarly. Here, a scenario category can be seen as the qualitative counterpart of a scenario. This assumption does not limit the applicability of the proposed methodology, although it may require many scenario categories to cover all scenarios. We will denote a scenario category by $\mathcal{C}$.

The main idea of the presented method is to compare the simulation outcome of a non-parameterized scenario with the simulation outcome of the corresponding parameterized scenario. It is important to note that assumptions are also made for the non-parameterized scenario since it is being modeled. For instance, a specific sample frequency is chosen. Additionally, each simulation inherently includes limitations and simplifications of reality. When parameterizing a scenario, *additional assumptions* are made to simplify its representation. The objective of the proposed method is to examine whether the *additional assumptions* used for parameterizing the scenario can be justified.

To formalize the method, let $R(S) \in \{0, 1\}$ denote the outcome of the simulation of scenario S, where $R(S) = 1$ indicates a failure according to a particular criterion and $R(S) = 0$ indicates a pass according to the same criterion. For example, $R(S) = 1$ indicates that a collision happened, while $R(S) = 0$ means that the simulation finished without a collision. In the case study in Section IV, two additional examples are considered.

To examine the parameterization, a set of non-parameterized scenarios belonging to scenario category $\mathcal{C}$ is needed. For this work, we assume that such a set is available. One approach to obtain these scenarios is to extract them from real-world data. A method for identifying scenarios belonging to $\mathcal{C}$ is detailed in [19]. This method involves two steps: First, tags are used to describe activities, such as lane changing and braking, and statuses, such as "leading vehicle" and "driving slower". Each tag is typically associated with an object and includes a start and end time. Second, by searching for a specific combination of these tags that define the scenario category $\mathcal{C}$, the start and end time of the scenarios can be identified.

To measure the influence of the parameterization on the simulation outcome, we compare $R(S)$ with $R(S')$, where S' denotes the parameterized version of S. A True Positive (TP) is noted when $R(S) = R(S') = 1$. If $R(S) = 1$ and $R(S') = 0$, a False Negative (FN) is reported while $R(S) = 0$ and $R(S') = 1$ indicates a False Positive (FP). The recall is the ratio of the number of TPs and the total number of fails when considering the non-parameterized scenarios (TP+FN) and the precision is the ratio of the number of TPs and the total number of fails when considering the parameterized scenarios (TP+FP). To quantify the influence, we will look at the F1 score, which is the harmonic mean of the recall and the precision:

$$\mathrm{F1score} = 2 \cdot \frac{\mathrm{Recall} \cdot \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}.$$

IV. SETUP CASE STUDY

The scenarios will be discussed first. Next, we will mention the four different models that are used in this case study in Section IV-B. Section IV-C lists the three test criteria that are used. Finally, Section IV-D presents the different parameterizations.

A. Scenarios

Following UN R157, three different scenario categories are considered: cut-in, cut-out, and LVD. To obtain the non-parameterized scenarios belonging to these scenario categories, the HighD data set [20] is chosen. The data consists of trajectories of cars and trucks at six different locations on German motorways obtained using video footage from drones.

To obtain the scenario data, each of the more than $100\,000$ vehicles is treated as an ego vehicle once. I.e., from the total data set, more than $100\,000$ smaller data sets are created, where each of the smaller data sets contains a single ego vehicle and trajectory data relative to the ego vehicle as if the other vehicles are perceived from the ego vehicle. It is assumed that the ego vehicle can see all of its surrounding vehicles within a distance of $100\,\mathrm{m}$. Each of the smaller data sets stops whenever the ego vehicle is $100\,\mathrm{m}$ from its final position; this is done to avoid the sudden disappearance of vehicles in front of the ego vehicle, as these vehicles would be out of view of the drone camera. In total, this resulted in $109\,986$ data sets with a single ego vehicle.

From the data, 2992 cut-ins have been found. For this study, we consider only those cut-ins where the speed of the vehicle cutting in is less than $95\,\%$ of the ego vehicle's speed and the Time Headway (THW) is less than $2\,\mathrm{s}$. This has resulted in 362 cut-ins being used for the experiment. In total, 3069 cut-outs have been identified. When considering only those cut-outs with another vehicle in front of the vehicle cutting-out that is slower than the ego vehicle, only 819 cut-outs were left. We have found $20\,351$ LVD scenarios. To limit the number of LVD scenarios for the experiment, only the 482 LVD scenarios in which the deceleration exceeded $2\,\mathrm{m/s^2}$ have been considered.

B. Models

To demonstrate that the influence of the parameterization depends on the system-under-test, four different driver reference models are used. For the sake of brevity, these models will be summarized shortly; for a more elaborate description, see [21].

The first model is called Reg157 and is based on paragraph 5.2.5.2 of UN R157 that dictates that a collision should be avoided if the Time To Collision (TTC) is above a certain threshold. Here, the TTC is the time remaining until two vehicle collide if they would continue on the same course and speed [22]. With the Reg157 model, when the TTC becomes less than a certain threshold (depending on the speed), the ego vehicle decelerates with $6\,\mathrm{m/s^2}$ after a reaction time of $0.35\,\mathrm{s}$.

The second model is the Careful and Competent Human Driver Model (CCHDM), which is defined in Appendix 3 of UN R157. The ego vehicle with the CCHDM brakes with a delay of $0.6\,\mathrm{s}$ after the TTC is below $2\,\mathrm{s}$. The CCHDM differs from the Reg157 model in that it assumes different stages of braking (releasing foot from acceleration pedal and actual braking, resulting in a deceleration of $0.4\,\mathrm{m/s^2}$ and $7.59\,\mathrm{m/s^2}$, respectively) and a linear increase of the deceleration with a jerk of $12.65\,\mathrm{m/s^3}$.

The third model is based on [23], which introduces the Responsibility Sensitive Safety (RSS) model. The RSS model outlines some constraints that, under specific assumptions, ensure safety if a vehicle adheres to them. In this case, the ego vehicle will decelerate as soon as both the longitudinal and lateral safety distance margins are violated.

The fourth model is the Fuzzy Safety Model (FSM) [24]. Here, the ego vehicle brakes as soon as one of the two safety metrics are nonzero. The difference with the other three models is that the FSM allows for only gentle braking.

C. Pass/fail criteria

The influence of the parameterization depends on the pass/fail criteria that one is interested in. To demonstrate this, three different pass/fail criteria are used. The first criterion is that the ego vehicle should not collide. The second criterion is that the TTC, already mentioned in Section IV-B, should remain above $1\,\mathrm{s}$. Note that the TTC is only evaluated if the relative lateral position of the vehicle in front does not exceed the width of the ego vehicle.

The third criterion is that the Brake Threat Number (BTN) should remain below 0.8. The BTN describes how difficult a collision avoidance maneuver by braking is given a maximum deceleration and jerk. If the BTN is above 1, it is not possible to avoid a collision, while a BTN of 0 indicates that there is no threat. For the maximum deceleration and jerk, we use the same values as for the CCHDM. Since the original definition of the BTN [25] requires numerical solving for which convergence cannot be guaranteed up front, we use the modified version presented in [26]. Similar as for the TTC, the BTN is only evaluated if the relative position of the vehicle in front does not exceed the width of the ego vehicle.

D. Parameterizations

As shown in [21], the initial distance between the ego vehicle and the other vehicle(s) participating in the scenarios has a large influence on the outcome. If we would use the initial distance as observed in the data, there would be no collision in the simulations as the original data did not contain any (near-)collision. Therefore, following the approach in [27], for all scenarios, the initial THW is varied from $2\,\mathrm{s}$ down to $0.2\,\mathrm{s}$ in steps of $0.2\,\mathrm{s}$. Consequently, each of the observed scenarios is simulated ten times with varying THW.

For the *cut-in* scenarios, the following parameterizations are considered:

1) Similar to the parameterization in UN R157, i.e., the initial longitudinal and lateral velocity of the cutting-in vehicle ($v^{\mathrm{c}}_{x,0}$ and $v^{\mathrm{c}}_{y,0}$, respectively) and the initial longitudinal velocity of the ego vehicle ($v^{\mathrm{e}}_{x,0}$) are used. It is assumed that the longitudinal velocity of the cutting-in vehicle remains constant while the lateral velocity remains constant until the lane change, with a width of $3.5\,\mathrm{m}$, has been completed.
2) Similar to parameterization 1, but now a constant longitudinal acceleration/deceleration of the cutting-in vehicle ($a^{\mathrm{c}}_{x,0}$) is assumed, equal to the mean acceleration/deceleration in the non-parameterized scenario. In case the cutting-in vehicle comes to a standstill, it will remain stationary.
3) This parameterization is based on [7] and is similar to parameterization 1, but now the lane change is assumed to happen instantaneously. As a result, $v^{\mathrm{c}}_{x,0}$ and $v^{\mathrm{e}}_{x,0}$ are the only two parameters.
4) Similar to parameterization 2, but now the lane change is assumed to happen instantaneously.
5) Following the method from [14], the number of parameters are reduced using Singular Value Decomposition (SVD). The original time series are the lateral position and longitudinal velocity of the cutting-in vehicle and the additional parameters are $v^{\mathrm{e}}_{x,0}$ and the duration of the lane change of the cutting-in vehicle. With the parameter reduction, only $d = 3$ parameters are used. Note that SVD is employed by Principal Component Analysis (PCA) [28], so both SVD and PCA reduce the dimensionality of data while preserving as much variability as possible.
6) Similar to parameterization 5, but with $d = 4$.
7) Similar to parameterization 5, but with $d = 5$.

For the *cut-out* scenarios, the following parameterizations are considered:

a) Mostly similar to the parameterization in UN R157, i.e., the initial longitudinal and lateral velocity of the cutting-out vehicle ($v^{\mathrm{c}}_{x,0}$ and $v^{\mathrm{c}}_{y,0}$, respectively) and the initial longitudinal velocity of the ego vehicle ($v^{\mathrm{e}}_{x,0}$) are used. It is assumed that the longitudinal velocity of the cutting-out vehicle remains constant while the lateral velocity remains constant until the lane change, with a width of $3.5\,\mathrm{m}$, has been completed. Whereas UN R157 considers a stationary vehicle in front of the cutting-out vehicle, we consider a leading vehicle moving with a constant speed ($v^{\mathrm{l}}_{x,0}$) and an initial distance of D from the cutting-out vehicle.
b) Similar to parameterization a, but now a constant longitudinal acceleration/deceleration is assumed for the cutting-out vehicle ($a^{\mathrm{c}}_{x,0}$) and the leading vehicle ($a^{\mathrm{l}}_{x,0}$) equal to the respective mean acceleration/deceleration in the non-parameterized scenario. In case any of these vehicles comes to a standstill, it will remain stationary.
c) This parameterization is based on [7] and is similar to parameterization a, but now the lane change is assumed to happen instantaneously.
d) Similar to parameterization b, but now the lane change is assumed to happen instantaneously.
e) Following the method from [14], the number of parameters are reduced using SVD. The original time series are the lateral position of the cutting-in vehicle and the longitudinal velocity of the cutting-in and leading vehicles. The additional parameters are $v^{\mathrm{e}}_{x,0}$, D, and the duration of the lane change. With the parameter reduction, only $d = 3$ parameters are used.
f) Similar to parameterization e, but with $d = 5$.
g) Similar to parameterization e, but with $d = 7$.

For the *LVD* scenarios, the following parameterizations are considered:

i) Similar to the parameterization in UN R157, i.e., with the initial longitudinal velocity ($v^{\mathrm{l}}_{x,0}$), final longitudinal velocity ($v^{\mathrm{l}}_{x,1}$), and the mean deceleration ($\bar{a}^{\mathrm{l}}_{x}$) of the leading vehicle are used. It is assumed that the deceleration is linear. After the deceleration, the leading vehicle maintains its speed. The initial speed of the ego vehicle equals $v^{\mathrm{l}}_{x,0}$.
ii) Similar to parameterization i, but now the deceleration follows a sinusoidal shape [7].
iii) Using the longitudinal velocity over time of the lead vehicle and the total duration of the deceleration, the number of parameters are reduced following the approach in [14]. In total, $d = 3$ parameters are used.
iv) Similar to parameterization iii, but with $d = 4$.
v) Similar to parameterization iii, but with $d = 5$.

V. RESULTS

Table I lists the total number of failures per scenario category, model, and pass/fail criteria for the baseline simulations, i.e., the simulations of the non-parameterized scenarios. All models experience a substantial number of collisions in *cut-in scenarios*. These collisions include instances where the cutting-in vehicle hits the ego vehicle from the side. In those cases, the TTC and BTN are never calculated since the cutting-in vehicle was never in front of the ego vehicle. This explains why there may be more collisions than simulations where the TTC is below $1\,\mathrm{s}$.

Table I shows that the Reg157 model and CCHDM frequently fail the TTC and BTN criteria, with the Reg157 model failing much more often. This occurs because these models only decelerate when the TTC drops below a certain

TABLE I

NUMBER OF FAILS PER SCENARIO CATEGORY, MODEL, AND PASS/FAIL CRITERION FOR THE BASELINE SIMULATIONS. IN TOTAL, 3620 CUT-IN, 8190 CUT-OUT, AND 4820 LVD SIMULATIONS ARE PERFORMED.

Scenario	Model	Collisions	TTC	BTN
Cut-in	Reg157	349	2249	2481
	CCHDM	538	457	677
	RSS	266	157	329
	FSM	387	205	476
Cut-out	Reg157	195	4623	4607
	CCHDM	65	1219	366
	RSS	21	94	65
	FSM	23	91	65
LVD	Reg157	2202	4819	4819
	CCHDM	1303	3221	2482
	RSS	6	26	20
	FSM	15	32	22

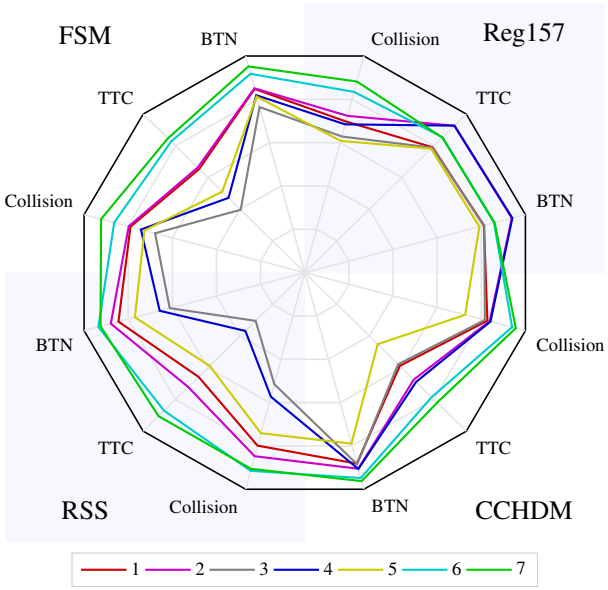

Fig. 1. F1 scores for the *cut-in* parameterizations. The distance from the center represents the F1 score, starting at 0 in the center until a maximum of 1 at the outer black line. The different lines denote the different parameterizations as listed in Section IV-D.

threshold. Since both low TTC and high BTN signals a threat, failing the TTC criterion often results in failing the BTN criterion. In contrast, the RSS model and FSM anticipate earlier to threats, resulting in significantly fewer failures. Another observation is the high number of collisions for the Reg157 model and CCHDM in *LVD scenarios*. This happens because these models do not account for the leading vehicle's deceleration, causing them to begin braking too late.

Fig. 1 shows a radar plot with F1 scores for the *cut-in* parameterizations. Adding acceleration as a parameter ($a^{\mathrm{c}}_{x,0}$) improves the F1 scores (parameterization 2 vs. 1), though it introduces an extra parameter. However, assuming an instantaneous lane change by removing parameter $v^{\mathrm{c}}_{y,0}$ results in significantly worse results, except for the TTC and

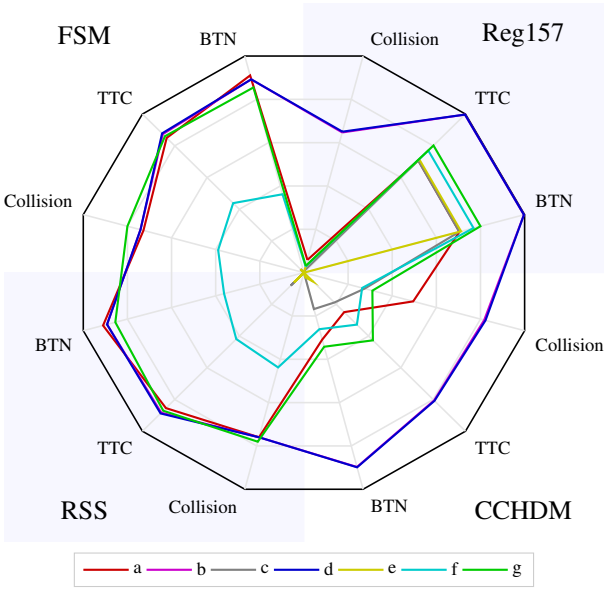

Fig. 2. F1 scores for the *cut-out* parameterizations. Parameterization b is hardly visible because its results are almost similar to the results of parameterization d.

BTN criteria with the Reg157 model. Generally, assuming an instant lane change leads to many FPs (meaning that a critical TTC or BTN is reached in the parameterized scenario but not in the non-parameterized scenario) because the ego vehicle has less time to react in the parameterized scenario. Consequently, this lowers the precision and F1 score. Exceptions are the TTC and BTN criteria with the Reg157 model, which is highly sensitive to the front vehicle's acceleration, making parameterizations 2 and 4 perform best in these cases. Generally, using the parameter reduction method with $d = 4$ (parameterizations 6) or $d = 5$ (parameterizations 7) provides the best result in terms of F1 scores.

For *cut-out* parameterizations, results vary significantly by model and pass/fail criteria, as shown in Fig. 2. With fewer collisions, FPs or FNs have a higher impact on the F1 score. This generally leads to lower F1 scores. The Reg157 model and CCHDM show low F1 scores for all parameterizations except b and d because even slight changes in acceleration/deceleration can alter simulation outcomes. Parameterizations a and c assume constant speed while parameterizations e to g contain accelerations, but the parameter reduction may cause slight deviations in the actual acceleration/deceleration values.

For the RSS model and FSM, parameterization c reports an F1 score of almost zero due to the many FNs, resulting in a low recall. Similarly, parameterizations e and f have low F1 scores are obtained, suggesting that reducing the parameterization to $d = 3$ or $d = 5$ parameters seems insufficient.

Fig. 3 displays the F1 scores for the *LVD scenarios*. All parameterizations achieve perfect F1 scores for the Reg157 model with the TTC and BTN criteria. Other than that, however, parameterizations i and ii perform poorly, mainly

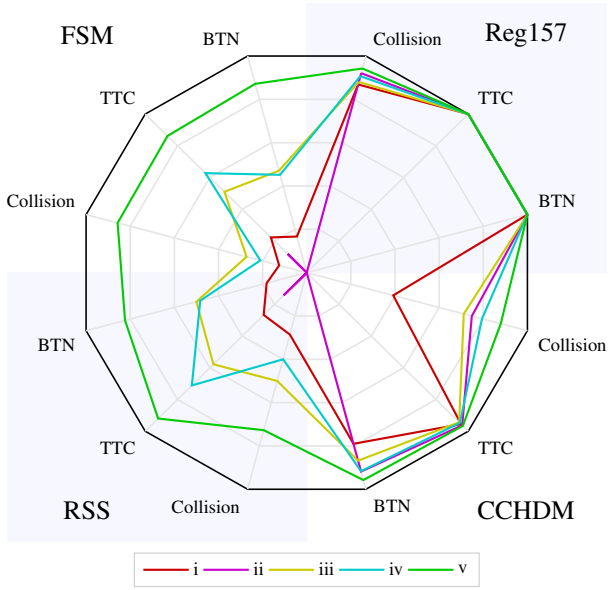

Fig. 3. F1 scores for the *LVD* parameterizations.

due to many FNs. When parameter reduction is applied (parameterizations iii to v), the F1 scores improve significantly, with parameterization v yielding the best results in all cases.

VI. DISCUSSION

In this work, we have demonstrated that scenario parameterization affects simulation outcomes. Also, different systems under test and test criteria may require different parameterizations. Specifically, we have found that the parameterization used in the UN R157 might not be optimal. Although not considered in the current study, the level of detail in a simulator also impacts the optimal parameterization; more detailed simulators likely require more parameters to describe a scenario. For example, environmental conditions like precipitation could influence systems, and if modeled, would need to be considered as scenario parameter(s).

To measure the impact of parameterizations, we have used the F1 score, which requires both high recall and precision. Achieving both is generally challenging, so a trade-off is often made to maximize the F1 score. For safety purposes, focusing solely on recall might be sufficient, which means that FPs are not penalized. This approach could result in a conservative design where a potentially safe system produces failures due to a low precision.

In addition to considering metrics like the F1 score, recall, and precision, there are other factors that may influence the choice of parameterization. For instance, when estimating a probability density function for the parameters, the number of parameters plays a crucial role in the accuracy of the estimation. Generally, the more parameters there are, the lower the accuracy of the estimation tends to be. Moreover, having more parameters can make it harder to achieve full coverage of the parameter space when sampling scenario parameter values. Thus, even if a parameterization yields a lower F1 score, a simpler model with fewer parameters might sometimes be preferable.

We have shown that scenario parameterization can significantly influence the simulations outcome, raising the question of whether to parameterize scenarios at all. In Section II-A, we have explained the benefits of parameterization. However, combining simulations of both parameterized and non-parameterized scenarios can still be useful. Replaying real-world, non-parameterized scenarios provides different insights. In addition, also simulating non-parameterized scenarios helps monitoring if the parameterization is still appropriate. Provided that parameterization of scenarios remain useful, future work involves researching the potential of additional parameterization techniques, such as feature selection methods and autoencoders [29].

Four different driver models have been considered in this work, all aiming to represent a "skilled and attentive human driver" [6, Annex 4, Appendix 3]. As presented in Section V, the four different models displayed significant differences in performance. This study did not focus on evaluating the actual performance of these models; further research is necessary to establish a true baseline for an ADS.

VII. CONCLUSIONS

Scenario-based assessment is crucial for evaluating Automated Driving Systems (ADSs). The United Nations Regulation 157 (UN R157) concerning the approval of Automated Lane Keeping Systems (ALKSs) recommends a scenario-based approach to benchmark for ALKSs against "the simulated performance of a skilled and attentive human driver". In this study, we have examined the scenario parameterizations proposed in UN R157 and demonstrated that they significantly impact simulation outcomes. This paper shows the need for careful consideration when adopting specific scenario parameterizations and that the optimal choice of a parameterization depends on factors like the system under test and the test criteria. We have proposed a method to assess the scenario parameterization and, by applying it to the UN R157 scenarios, identified and tested alternatives parameterizations that show potential improvements over the scenario parameterization currently used in UN R157.

Based on the research presented in this work, we recommend that future amendments to UN R157 include an additional requirement. Currently, the system's performance in parameterized scenarios is compared with the performance of a human reference model to show that the activated system does not cause any collisions that are reasonably preventable. In addition, it should be necessary to justify that the chosen parameterization of scenarios is appropriate. This justification should consider the system's performance, the type of scenario, and the specific metrics of interest.

REFERENCES

[1] C.-Y. Chan, "Advancements, prospects, and impacts of automated driving systems," *International Journal of Transportation Science and Technology*, vol. 6, no. 3, pp. 208–216, 2017.

[2] I. Mahdinia, R. Arvin, A. J. Khattak, and A. Ghiasi, "Safety, energy, and emissions impacts of adaptive cruise control and cooperative adaptive cruise control," *Transportation Research Record*, vol. 2674, no. 6, pp. 253–267, 2020.

[3] A. Mammeri, G. Lu, and A. Boukerche, "Design of lane keeping assist system for autonomous vehicles," in *7th International Conference on New Technologies, Mobility and Security (NTMS)*, 2015, pp. 1–5.

[4] SAE J3016, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," SAE International, Tech. Rep., 2021.

[5] N. E. Vellinga, "Automated driving and the future of traffic law," in *Regulating New Technologies in Uncertain Times*, Springer, 2019, pp. 67–82.

[6] E/ECE/TRANS/505/Rev.3/Add.156, "Uniform provisions concerning the approval of vehicles with regard to automated lane keeping systems," World Forum for Harmonization of Vehicle Regulations, Standard, 2021. [Online]. Available: https://unece.org/sites/default/files/2021-03/R157e.pdf.

[7] E. de Gelder, H. Elrofai, A. Khabbaz Saberi, O. Op den Camp, J.-P. Paardekooper, and B. De Schutter, "Risk quantification for automated driving systems in real-world driving scenarios," *IEEE Access*, vol. 9, pp. 168 953–168 970, 2021.

[8] T. Menzel, G. Bagschik, and M. Maurer, "Scenarios for development, test and validation of automated vehicles," in *IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1821–1827.

[9] H. Nakamura, H. Muslim, R. Kato, S. Préfontaine-Watanabe, H. Nakamura, H. Kaneko, H. Imanaga, J. Antona-Makoshi, S. Kitajima, N. Uchida, E. Kitahara, K. Ozawa, and S. Taniguchi, "Defining reasonably foreseeable parameter ranges using real-world traffic data for scenario-based safety assessment of automated vehicles," *IEEE Access*, vol. 10, pp. 37 743–37 760, 2022.

[10] E. de Gelder and O. Op den Camp, "A quantitative method to determine what collisions are reasonably foreseeable and preventable," *Safety Science*, vol. 167, p. 106 233, 2023.

[11] S. Feng, Y. Feng, C. Yu, Y. Zhang, and H. X. Liu, "Testing scenario library generation for connected and automated vehicles, part I: Methodology," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1573–1582, 2020.

[12] L. Li, N. Zheng, and F.-Y. Wang, "A theoretical foundation of intelligence testing and its application for intelligent vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, pp. 6297–6306, 2020.

[13] S. Thal, H. Znamiec, R. Henze, H. Nakamura, H. Imanaga, J. Antona-Makoshi, N. Uchida, and S. Taniguchi, "Incorporating safety relevance and realistic parameter combinations in test-case generation for automated driving safety assessment," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2020, pp. 666–671.

[14] E. de Gelder, J. Hof, E. Cator, J.-P. Paardekooper, O. Op den Camp, J. Ploeg, and B. De Schutter, "Scenario parameter generation method and scenario representativeness metric for scenario-based assessment of automated vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 18 794–18 807, 2022.

[15] J. Cai, W. Deng, H. Guang, Y. Wang, J. Li, and J. Ding, "A survey on data-driven scenario generation for automated vehicle testing," *Machines*, vol. 10, no. 11, p. 1101, 2022.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *27th International Conference on Neural Information Processing Systems*, vol. 2, 2014, pp. 2672–2680. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[17] A. Demetriou, H. Allsvåg, S. Rahrovani, and M. H. Chehreghani, "Generation of driving scenario trajectories with generative adversarial networks," in *IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–6.

[18] J. Spooner, V. Palade, M. Cheah, S. Kanarachos, and A. Daneshkhah, "Generation of pedestrian crossing scenarios using ped-cross generative adversarial network," *Applied Sciences*, vol. 11, no. 2, p. 471, 2021.

[19] E. de Gelder, J. Manders, C. Grappiolo, J.-P. Paardekooper, O. Op den Camp, and B. De Schutter, "Real-world scenario mining for the assessment of automated vehicles," in *IEEE International Transportation Systems Conference (ITSC)*, 2020, pp. 1073–1080.

[20] 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," in *IEEE 21st International Conference on Intelligent Transportations Systems (ITSC)*, 2018, pp. 2118–2125.

[21] K. Mattas, G. Albano, R. Donà, M. C. Galassi, R. Suarez-Bertoa, S. Vass, and B. Ciuffo, "Driver models for the definition of safety requirements of automated vehicles in international regulations. application to motorway driving conditions," *Accident Analysis & Prevention*, vol. 174, p. 106 743, 2022.

[22] J. C. Hayward, "Near miss determination through use of a scale of danger," Pennsylvania State University, Tech. Rep. TTSC-7115, 1972. [Online]. Available: https://onlinepubs.trb.org/Onlinepubs/hrr/1972/384/384-004.pdf.

[23] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374*, 2017. [Online]. Available: https://arxiv.org/abs/1708.06374.

[24] K. Mattas, M. Makridis, G. Botzoris, A. Kriston, F. Minarini, B. Papadopoulos, F. Re, G. Rognelund, and B. Ciuffo, "Fuzzy surrogate safety metrics for real-time assessment of rear-end collision risk. a study based on empirical observations," *Accident Analysis & Prevention*, vol. 148, p. 105 794, 2020.

[25] M. Brannstrom, J. Sjoberg, and E. Coelingh, "A situation and threat assessment algorithm for a rear-end collision avoidance system," in *IEEE Intelligent Vehicles Symposium*, 2008, pp. 102–107.

[26] A. Andersson, "Multi-target threat assessment for autonomous emergency braking," M.S. thesis, Chalmers University of Technology, 2016. [Online]. Available: https://publications.lib.chalmers.se/records/fulltext/245725/245725.pdf.

[27] M. R. Zofka, F. Kuhnt, R. Kohlhaas, C. Rist, T. Schamm, and J. M. Zöllner, "Data-driven simulation and parametrization of traffic scenarios for the development of advanced driver assistance systems," in *18th International Conference on Information Fusion*, 2015, pp. 1422–1428. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7266724.

[28] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[29] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.