Dr. Greg Chance, PhD, CEng, MInstP

Dr. Greg Chance
Trustworthy Systems Lab
University of Bristol
1 Cathedral Square
Bristol, BS1 5DD, UK
Email: greg.chance@bristol.ac.uk
URL: `www.bristol.ac.uk`

September 23, 2021

Prof. Eskandarian
Department Head
Nicholas and Rebecca Des Champs Chair
Mechanical Engineering Department, Virginia Tech
635 Prices Fork Road, 449 Goodwin Hall (MC 0238) Blacksburg, VA 24061

Review of paper T-ITS-21-05-1111.

Dear Prof Azim Eskandarian,

Thank you and the reviewers for taking the time to give feedback our manuscript. We have given each comment careful consideration between all the authors and we give a full account of the changes and rebuttals below. We also include a diff.pdf to show the changes made between the current and originally submitted manuscripts.

Reviewer 1
Comment 1: "the whole manuscript seems written from the perspective of Game Engines testing, which lacks practical values in the transportation area".

Thank you for this comment. We believe this comment requires a considered response to ensure that the main message of the paper is being properly conveyed to the reader. Carla is based on a game engine and Carla is currently a popular choice for simulation-based vehicle testing, so this is strong evidence that the underlying game engine of Carla is of practical value to the community of intelligent transportation. The paper is not focused on game engines per se, but rather the implications of using those game engines for vehicle testing.

We think that there must just be a misunderstanding here, we clearly state the importance of needing determinism for verification (from page 3):

> If the simulation is non-deterministic, e.g. it has a non-zero variance in, for example, actor positions, then this may, in the best case, lead to intermittent assertion failures, making it difficult to reproduce, understand and remove bugs and rendering verification results unstable. In the worst case, however, bugs that could have been identified in simulation remain undetected, leading to false confidence in the safety of the AV's control software.

We state that game engines are the underlying physics framework that support some driving simulators, such as Carla. Therefore, if you are using a driving simulator that is based on a game engine, this may lead to non-determinism and compromise your verification analysis. We agree with the reviewer that determinism for game engines may not be an important research field by itself, but we go on to state that this is not the case for AV verification (from page 3):

> When used for gaming, game engines do not need to be deterministic nor do they have any requirements on the limits of permissible variance; there are no safety implications from non-determinism in this domain, nor is finding and

> fixing all the bugs a high priority for games developers. It could even be argued that simulation variance is a feature that enhances gaming and improves the user experience. However, the situation is very different for AV development and testing. Thus, our main research questions are: How can one assess whether a simulation environment is deterministic? and How can one determine and control the simulation variance?

We also state on pg.4 the differences between the requirements of game engines for gaming, and for that of AV testing. We state that there is no need for determinism in gaming (it may even benefit gameplay) but that this is not the case for verification:

> Considering the objectives for gaming and comparing them to these for AV development and testing, there are fundamental differences. Providing game players with a responsive real-time experience is often achieved at the cost of simulation accuracy and precision. The gamer neither needs a faithful representation of reality (i.e. gamer accepts low accuracy) nor require repeated actions to result in the same outcome (i.e. gamer accepts low precision). In contrast, high accuracy and precision are necessary for AV development, testing and verification.

Carla, and other game engine based simulators, will be an entry point for many SME's and start-up companies looking to develop products and services in this area and we believe that this paper brings pertinent information to this community, many of whom may look to ITS for guidance. We believe this point is sufficiently clear to the reader and hope these examples bring this point to bear.

Comment 2: "One major concern is index selection. As we all know, the scale of deviation relies on the mean values of the investigated variables. Thus, I wonder why the authors pick maximum deviation to measure the performance of the simulation results. In my opinion, the average deviation seems better in measuring the overall performance for the whole simulation process".

The issue we found with taking the average deviation over many (100's or 1000's) of repeated runs is that a single 'failure' can be hidden in an average. For example 1 error in 1000 would be an insignificant difference to an average, but a single simulation run that exceeds permissible variance could result in a false negative result or even fail to detect the presence of a serious fault with the system under test.

The verification process, therefore, needs to be aware of any non-deterministic event beyond the permissible tolerance, as even a single event may coincide with a bug or error in the system that needs to be found and corrected. Hence, it is imperative that we discover any simulation that exceeds the simulation variance.

We could reframe this as thinking less about simulation performance, and more about error detection. It is important to know that any failure truly existed and if necessary can be repeated so that the bug can be tracked down and resolved. If the simulator is not deterministic then re-running the same test and hoping to observe the same error may not happen.

We have added (in italics) to the sentence on page 9 to clarify this point:

> Initial testing [26] indicated an actor path deviation of 1×10^{-13}cm for 997 out of 1000 tests, with three tests reporting a deviation of over 10cm. *Due to the low probability of a simulation trace exceeding permissible variance, using an average would 'hide' these errors and hence this is the reason that we use maximum variance and not an average.*

Comment 3: "Moreover, I also concern about the results from several tests. As stated in Section V-B, 1 cm considered sufficient for urban scenario assertion checking. However, the average vehicle speeds are setting as 20–35 km/h (i.e. 5.56-9.72 m/s) [1]. Thus, I *doubt the practical value of this permissible variance (too accurate for general AV testing) in real-world autonomous vehicle verification. The authors should provide more support materials for the setting used in the whole experiment".*

Thank you we take the point you are making. This seems to be less about the results quoted but more about what value of variance is appropriate for the given case study or application. The focus of this paper is not really the absolute values that we chose for permissible variance, indeed we state this value should be set by the user in the methodology (see section VII.A.3), or the speed of the vehicles (which was based on UK urban speed limit of 20mph or 30kmph), but more about the method of detecting and controlling simulation variance in an end-user's system. The value of permissible variance chosen in the paper (1cm) was chosen as a good point between, say 1m (too coarse) and 1mm (too fine), to illustrate the violation of this limit given certain simulator and system conditions.

The point is that repeated runs should give the exact same output, regardless of permissible variance or simulated vehicle speed. The reviewer suggests a paper that describes the speed of vehicles in an urban environment similar to those in our case study. We have included this reference including a line to point interested readers whom wish to set the speed of their simulated vehicles to an appropriate value.

Please see the additional line (in italics) add to page 7:

> To put this another way, we can accept a precision with a tolerance of ≤±1cm. *This tolerance may need to be chosen for the specific verification scenario and the speed of the vehicles within the environment [1].*

Comment 4: "The whole paper seems much more relevant to the UE4 engine than the CARLA platform. I understand the CARLA application is just a case study, but the first half part of the manuscript is less relevant to the AV testing. In particular, the authors said "the AV simulation domain introduces its unique challenges that were not considered in that paper." in section IV".

Carla uses Unreal Engine 4 (UE4) for vehicle testing, which is not the original purpose of UE4. The issues discussed in the paper do not have any negative implications for games, but (as the paper argues) they do have negative implications for vehicle testing. Thus, the issues are issues with Carla rather than UE4, and the same issues are likely to arise in any simulator based on a game engine.

There are examples of other simulators that run on game engines. AirSim developed by Microsoft for autonomous drone and vehicle development (see footnote[1]) is based on the same UE4 game engine as Carla. Unreal is also the simulator backbone for the research efforts at the University of Warwick where simulation based safety testing is done using a simulation cave and a real vehicle (see footnote[2]). The Unity game engine has also been used for development of the apollo driving simulator (see footnote[3]) which has partnerships with many vehicle manufacturers (see footnote[4]).

Comment 5: "The numbers of references are odd".

Please let us know any format changes that are required specifically. If you are referring to the document hyperlinks that link to each reference, these can be switched off if required.

[1] https://microsoft.github.io/AirSim/

[2] https://www.unrealengine.com/en-US/spotlights/meet-the-hybrid-real-time-simulator-for-testing-autonomous-vehicles

[3] https://apollo.auto/gamesim.html

[4] https://apollo.auto/about/ecological_partner.html#taboem

Comment 6: "In section I, the sentence "without the need for millions of miles of costly on-road testing" is too arbitrary, the on-road testing is still necessary for AVs verification".

We agree with the reviewer that the need for on-road testing is an essential component of the verification process. However, the previous sentence highlights the shortcomings of on-road testing where human fatalities have occurred due to a lack of verification which may result from insufficient exploration of the parameter space.

The point of this sentence was in the context of parameter space exploration and the fact that this space could never be explored sufficiently with on-road testing alone and hence the necessity of simulation for verification. We have modified the sentence to sound less arbitrary, and highlighted the complementary nature of both techniques:

> While on-road testing is an essential part of AV verification, it can be complementary to simulation, offering a means to explore the vast parameter space safely and efficiently [32] whilst reducing the amount of costly road trials [29].

Comment 7: "No need to provide such a detailed description of floating-point arithmetic since it does not cause non-zero simulation variance".

Floating point arithmetic does not contribute to simulation variance, but it is a common misconception that people often jump to when considering or discussing deterministic computation. We therefore believe it is important to fully explain the issue to ensure that there is no doubt going forward from this point.

To make this clearer we have modified this section to indicate this misconception. We have also made this section more concise. Please see page 4-5 starting:

> *A common misconception that is often jumped to when concerning non-determinism is the use of floating point number representation. This is an erroneous conflation with non-deterministic computational execution and we explain the reasons for this here.*

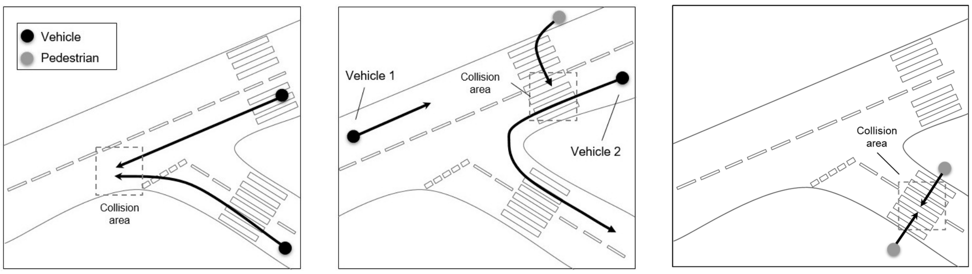
Comment 8: "There is no equal sign in Equation (1)".

Thank you we have created a symbol for this to turn this into an equation. See the amended equation on page 8:

$$\Psi = \max_{a,t} \sigma_a^2(t) \tag{1}$$

Comment 9: "TABLE I and Fig 3 are not consistent, it would be better if there are three subfigures in Fig 3".

Thank you we agree this would be clearer if there were three figures, this was done to save space in the document. Please see page 7 for the new figure.



Schematic of test scenarios for (a) Tests 1-2, (b) Tests 3-4, (c) Tests 5-6. Descriptions are given in Table 1.

Comment 10: "The size of simulation scenes should be pointed out, especially the boundary of (x, y, z)".

Thank you, we agree this would help describe the scene better to the reader. Please see section V.C on page 7 for the additional text:

> *The map size for the simulation test environment was 354m × 170m. The range of movement of the actors within the environment was up to 70m for vehicles and up to 25m for pedestrians.*

Comment 11: "Which vehicle collided with the pedestrian in test 4, please label them in Fig 3".

We agree it would be clearer for the reader if this information was presented. We have added the collision area to Fig.3 (on page 7) indicating where the collision happens and the ID of the vehicles.

Reviewer 2

Comment 1: "The current paper is a little bit too long and could be shortened. For instance, several definitions in Section II.A could be introduced in a more concise way. Sections II and III could be combined. Sections V and VI could be combined".

We agree the paper is a bit long, however if readers are familiar with the definitions it is likely that they will skip over these but it would be prudent to keep them in. We think that section V and VI are already long independently and combining them might become unwieldy.

We have attempted to reduce the paper size in different areas, please see the diff document. However, as the first reviewer wanted additional material added, including figures, this may have resulted in a net increase in overall manuscript length.

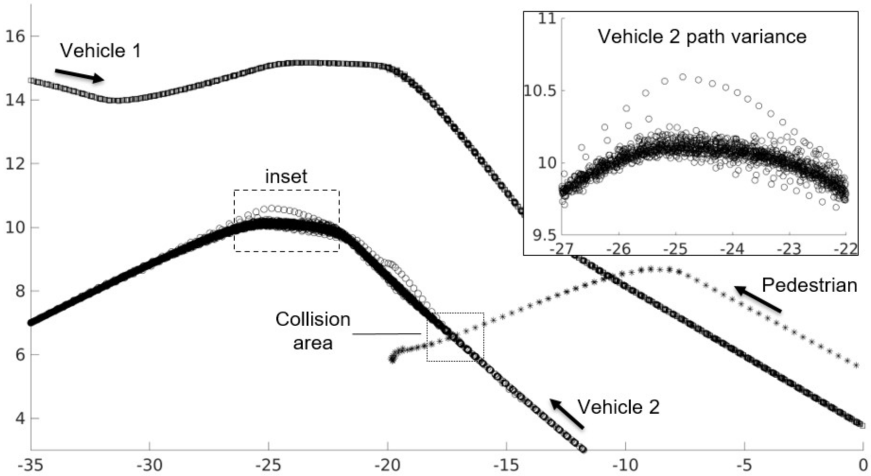
Comment 2: "In Section IV, several potential sources of non-determinism are reviewed. Through the case study, the authors should have established a general understanding of the relative significance of each potential source to the overall simulation variance (as least for CARLA). It would be good to comment on the relative significance alongside the review of these sources".

Giving a relative comparison is not easily achieved as each source would be difficult to quantify. Hopefully it is clear from Table.1 that pedestrian-vehicle collisions result in high simulation variance in our case study, and also the role that resource utilisation plays should be clear to the reader from viewing Fig.4. However, writing this as a general description would be difficult.

If the reader follows the methodology in section VII they should be able to discover the principle sources, quantify them and prioritise those which may lead to simulation variance above the permissible level for the user.

Comment 3: "It would be good if the authors could show the vehicle/pedestrian trajectories of a test case where a large maximum variance occur (such as the trajectories pre and post a collision). This could help the reader better understand the cause of the large trajectory variance".

Thank you, this is a great idea and we agree that this would give the reader a better insight into the issue. We have added an X-Y plane data plot showing the high simulation variance that occurs during test 4. In this plot you can see the area that vehicle 2 and the pedestrian collide within and the subsequent path variance (see Fig.5 on page 9).



Actor path plot in XY plane for Test 4 with 95% resource utilisation.

This new figure is referenced in the following text (additional in italics):

> However, all other scenarios involving vehicles or a mixture of actor types do not meet the required tolerance, with some deviation in actor path as large as 59cm. *A plot of actor position in the X-Y plane (plan view, units m) is shown in Fig.5, where the inset shows the divergence of the path of vehicle 2 post-collision with the pedestrian.* Clearly, such a large deviation cannot be acceptable for simulation to be considered a credible verification tool.

Thank you and the reviewers for your time and consideration to this paper. We hope given these changes you will consider this manuscript suitable for publishing.

Sincerely,

Dr. Greg Chance, PhD, CEng, MInstP