How Does Simulation-Based Testing for Self-Driving Cars Match Human Perception?

School of Engineering
InIT Institute of Computer Science

Christian Birchler, Tanzil Kombarabettu Mohammed, Pooja Rani, Teodora Nechita, Timo Kehrer, Sebastiano Panichella

Universität Zürich[UZH]
UNIVERSITÄT BERN

TO WHAT EXTENT DOES THE OOB QUALITY METRIC ALIGN WITH THE HUMAN PERCEPTION OF SAFETY AND REALISM?

(a) Failing Test: SDC driving off-lane (unsafe). (b) Passing Test: SDC driving in-lane (safe).

Fig. 1. Examples of simulation-based tests of an SDC.

(a) SDC in BeamNG.tech driving with 50 km/h close to obstacles

(b) SDC in CARLA crossing a red signal without stopping

Fig. 2. Examples of unsafe tests with valid OOB criteria

Problem statement: Software metrics such as coverage and mutation scores have been extensively explored for the automated quality assessment of test suites. While traditional tools rely on such quantifiable software metrics, the field of self-driving cars (SDCs) has primarily focused on simulation-based test case generation using quality metrics such as the out-of-bound (OOB) parameter to determine if a test case fails or passes. However, it remains unclear to what extent this quality metric aligns with the human perception of the safety and realism of SDCs, which are critical aspects in assessing SDC behavior.

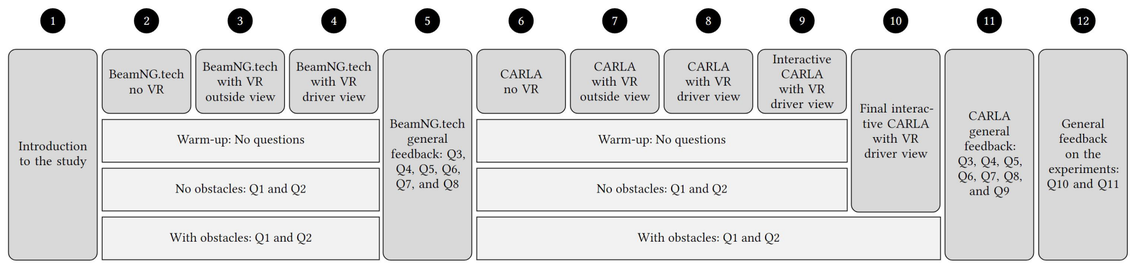
RESEARCH QUESTIONS

Empirical Study: It remains unclear to what extent this quality metric aligns with the human perception of the safety and realism of SDCs. To address this (reality) gap, we conducted an empirical study involving 50 participants to investigate the factors that determine how humans perceive SDC test cases as safe, unsafe, realistic, or unrealistic.

RQ1: To what extent does the OOB safety metric for simulation-based test cases of SDCs aligh with human safety assessment?

RQ2: To what extent does the safety assessment of simulation-based SDC test cases vary when humans can interact with the SDC?

RQ3: What are the main reality-gap characteristics perceived by humans in SDC test cases?

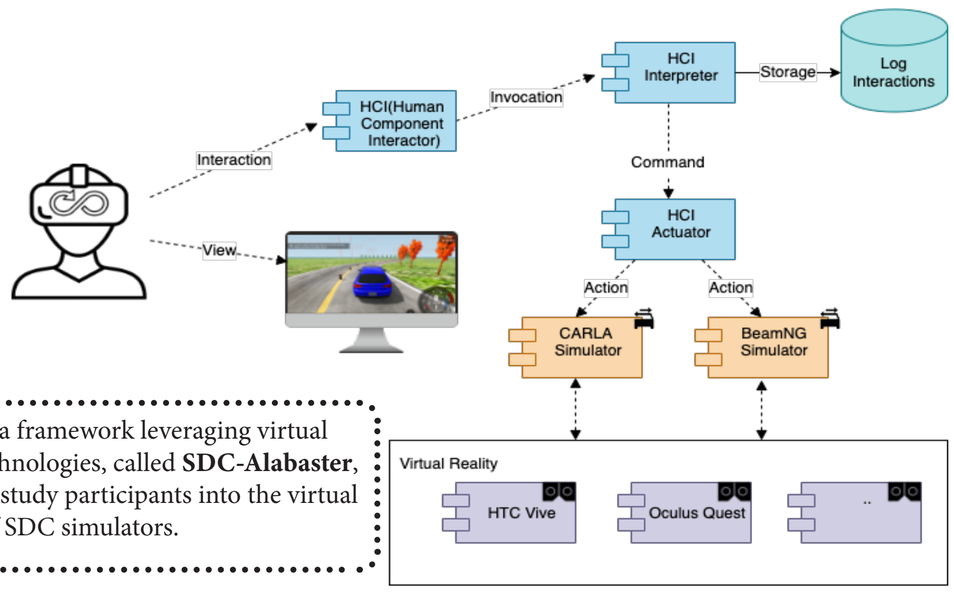


ID	Question	Type
Q1	What is the perceived safety of the Scenario?	LS
Q2	Justify the perceived safety of the Scenario.	OA
Q3	How would you scale the realism of scenarios generated by test cases in the simulator?	LS
Q4	Justify the level of realism of scenarios generated by test cases.	OA
Q5	How would you scale the driving of AI of the simulator?	LS
Q6	Justify the driving of AI from the simulator.	OA
Q7	How would you scale overall experience with the simulator?	LS
Q8	Justify overall experience with the simulator.	OA
Q9	How do you compare safety with and without interaction?	OA
Q10	Did this experiment change the way you thought about the safety of self-driving cars?	SC
Q11	Please write in a few words on your experience and suggestions.	OA

FRAMEWORK

(a) Participant with a VR headset (b) Outside view (c) Driver view

Fig. 3. One of our participants immersed in virtual SDCs with SDC-Alabaster

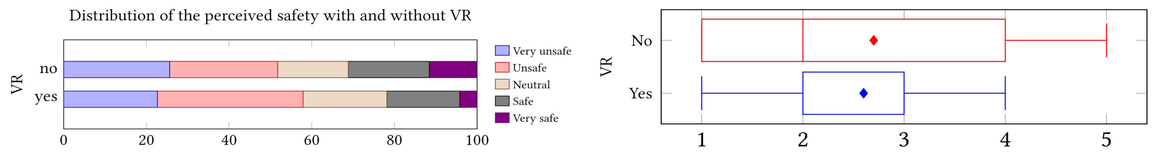


We developed a framework leveraging virtual reality (VR) technologies, called **SDC-Alabaster**, to immerse the study participants into the virtual environment of SDC simulators.

EMPIRICAL RESULTS

RQ1: To what extent does the OOB safety metric for simulation-based test cases of SDCs aligh with human safety assessment?

Finding 1: The passing test cases (i.e., the cases where the OOB metric is not violated) have a higher perception of safety from the participants than those failing (OOB metric is violated).



Finding 2: There is no statistical difference in safety perception between scenarios with and without obstacles when the OOB metric is not violated. However, when the car goes out of bounds, the scenario is perceived as significantly less safe with obstacles.

Finding 3: The utilization of VR had a minor impact on safety perception. However, participants using VR tended to perceive scenarios as somewhat less-safe, though this difference was not statistically significant (Wilcoxon rank-sum test, p = 0.16)

Finding 4: Overall, participants found the test cases less safe with obstacles.

RQ2: To what extent does the safety assessment of simulation-based SDC test cases vary when humans can interact with the SDC?

Finding 5: Safety perception of test cases is not static: When users can interact with the SDC, participants feel significantly safer (p = 0.013) compared to when they cannot.

Finding 6: Incorporating obstacles into the simulation, where participants interact with the SDC, leads to significantly lower perceived safety in test cases (p = 0.026) compared to obstacle-free interactive scenarios.

Finding 7: In the simulation, obstacles in non-interactive SDC test cases reduce the safety perception (p = 0.013). Yet, the ability to interact with the car raises more discomfort (making participants feel less safe) when obstacles are present.aA

RQ3: What are the main reality-gap characteristics perceived by humans in SDC test cases?

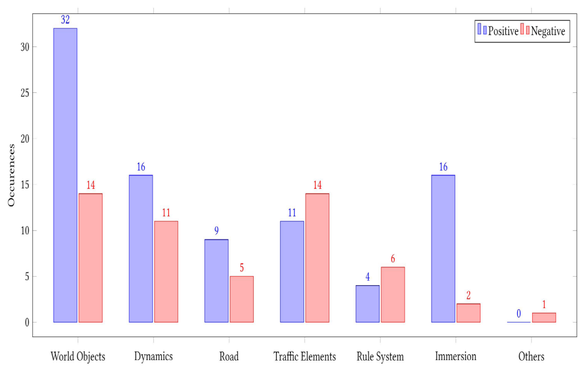


Fig. 9. Taxonomy of positive and negative factors impacting the perceived test cases' realism

Table 2. Taxonomy description including # of positive and negative comments on the perception of realism

Category	Description	Occurrences		
		Positive	Negative	Total
World Objects	This category relates to comments of participants on the accuracy of visual looks and design of all elements in the virtual environment, such as the weather, landscape, car design, traffic objects, etc., and how the graphical resolution is perceived.	32	14	46
Dynamics	This category relates to participants' comments on the physical dynamics of the elements in the virtual environment. For example, if the movement of the cars is physically realistic and reasonable or if crashes are realistically simulated from a physical perspective.	16	11	27
Road	This category relates to participants' comments on the road itself; to what extent the shape, surface, and structure are reasonably expected in the real world.	9	5	14
Traffic Elements	This category relates to participants' comments on the placement of the elements in the virtual environment. Furthermore, this category considers comments on the location and scale of the placed elements but also the quantity of the elements.	11	14	25
Rule System	This category relates to participants' comments on the traffic laws and the common sense of humans for resolving certain issues in specific traffic situations. A car should, for example, stop at a red signal and stop signs. Furthermore, the car should not drive recklessly and avoid dangerous situations (e.g., driving too close to other vehicles).	4	6	10
Immersion	This category relates to participants' comments on the immersive experiences. It applies to comments where participants express their feelings on how they experience the virtual environment and how they acoustically, visually, physically, and haptically sense it.	16	2	18
Others	This category relates to participants' comments that do not fit into the above categories.	0	1	1

MAIN FINDINGS & CONTRIBUTIONS

Our findings indicate that the human assessment of safety and realism of failing/passing test cases can vary based on different factors, such as the test's complexity and the possibility of interacting with the SDC. Especially for the assessment of realism, the participants' age leads to a different perception. This study highlights the need for more research on simulation testing quality metrics and the importance of human perception in evaluating SDCs.

Contributions:

- Proposing a methodology within the SDC-Alabaster framework, a VR-based approach, to study how quality metrics align with human perception of safety and realism in simulation-based testing, addressing the Reality Gap problem.
- Conducting the first empirical study on the perception of realism and safety in SDC test cases with 50 participants using VR technology, and publicly sharing a replication package with the code (Section 9).
- Sharing an initial taxonomy of factors influencing perceived realism of SDC simulators and discussing the confounding factors and implications of our work.

Our results demonstrate the impact of using VR in assessing SDCs, showing that "Safety perception of SDC test cases is not static." This underscores the importance of human interaction with the vehicle when evaluating SDCs using VR.