Master Thesis

January 6, 2026

‘"Safety Perception in Virtual Reality Testing of Self-Driving Cars: The SDC-Alabaster Approach"

Tanzil Kombarabettu Mohammed

of Mangalore, India (20-744-561)

supervised by

Prof. Dr. Davide Scaramuzza

Dr. Sebastiano Panichella

University of Zurich^UZH

ROBOTICS & PERCEPTION GROUP

Master Thesis

‘"Safety Perception in Virtual Reality Testing of Self-Driving Cars: The SDC-Alabaster Approach"

Tanzil Kombarabettu Mohammed

University of Zurich^UZH

ROBOTICS & PERCEPTION GROUP

Master Thesis

Author: Tanzil Kombarabettu Mohammed, tanzil.kombarabettumohammed@uzh.ch

URL: http://rpg.ifi.uzh.ch/index.html

Project period: 25.05.2022 - 25.11.2022

Robotics and Perception Group
Department of Informatics, University of Zurich

Acknowledgements

I want to express my gratitude to Professor Dr. Davide Scaramuzza for supervising my thesis, providing invaluable suggestions, and agreeing to supervise such a fascinating topic involving self-driving cars. I would also like to extend my thanks to Dr. Sebastiano Panichella, who guided me throughout the thesis by providing constant feedback, innovative advice, and a lot of insightful ideas. It would never have been possible for me to conduct such a novel and user-intensive experiment without his belief and constant support. Additionally, I want to thank Christian Birchler whose expertise in testing Self-Driving cars and his research work in SDC-Scissor provided me with a lot of in-depth understanding about the Self-Driving Cars which was base for my work, and I would also like to thank Sajad Khatiri for his guidance in scientific research and data analysis, along with Nechita Teodora for sharing her expertise with virtual reality and being part of my pilot study. I also want to thank all the participants who invested their valuable time and provided me with some great feedback, which played a huge role in helping me complete my experiment. I am extremely grateful for the constant and unconditional support I received from my parents for the past six months, without which this wouldn’t have been possible. I would also like to thank Albert and Akshay for their emotional support throughout this journey. Last but certainly not least, I would like to thank my friends and colleagues for all the support and encouragement I received from their end while being a part of this wonderful journey.

Abstract

Simulation-based testing helps in the improvement of cyber-physical systems (CPS) such as self-driving cars (SDC) because it increases the efficiency, diversity, and relevance of tests from a human perspective. The importance of human feedback in validating test cases cannot be overstated. Despite this, testing SDCs in simulated environments does not take human factors into account. Previous research demonstrates how to optimize the test case through selection, improve classification and accuracy when test cases result in a fault, and improve testing cost-effectiveness. However, test validity, relevance, and safety perception from a human point of view were not addressed. In this thesis, we investigate the variety of possible scenarios (static and dynamic obstacles) and examine how humans perceive safety and the level of realism of the SDC test case with various factors such as interaction with the car and different views (i.e., the VR view, the outside view, and the driver's view). We propose an approach called SDC-Alabaster (SDC hum**A**n-in-the **L**oop simul**A**tion-**BAS**ed **T**esting s**E**lf-driving ca**R**s) that uses a virtual reality (VR) headset to illustrate SDC test scenarios, create the sensation of being in SDCs and to enable users to experiment with the experience. Our results show the perception of realism and safety without obstacles is higher than with obstacles, and CARLA was more realistic and safer than the BeamNG simulator with a p-value $>$ 0.01e-16, The distribution is 85%($\hat{A}_{12}$). Our results also show interactions with vehicles make humans safer compared to those without interactions with a p-value $>$ 0.001, and the distribution is 36%($\hat{A}_{12}$), and users' perceptions of safety and realism vary with and without VR headsets, and the failure cases that are most important to test are also regarded as less realistic by participants'. In addition, we discovered factors such as using an advanced AI agent for traffic cars, using voice feedback in VR, and integrating participants' driving will help test scenarios be more realistic, and the perception of participants' safety can be improved in simulation-based testing of SDCs.

Zusammenfassung

Simulationsbasierte Tests helfen bei der Verbesserung von Cyber-Physical Systems (CPS) wie selbstfahrende Autos (SDC), weil sie die Effizienz, Vielfalt und Relevanz von Tests aus menschlicher Sicht erhöhen. Die Bedeutung des menschlichen Feedbacks bei der Validierung von Testfällen kann gar nicht hoch genug eingeschätzt werden. Trotzdem werden beim Testen von SDCs in simulierten Umgebungen menschliche Faktoren nicht berücksichtigt. Frühere Forschungen haben gezeigt, wie man den Testfall durch die Auswahl optimieren kann, sowie die Klassifizierung und Genauigkeit verbessern kann, wenn Testfälle zu einem Fehler führen, und wie man die Kosteneffizienz von Tests verbessern kann. Die Validität, Relevanz und Sicherheitswahrnehmung von Tests aus menschlicher Sicht wurden jedoch nicht berücksichtigt. In dieser Arbeit untersuchen wir die Vielfalt möglicher Szenarien (statische und dynamische Hindernisse) und untersuchen, wie Menschen die Sicherheit und den Realitätsgrad des Testfalls unter Berücksichtigung verschiedener Faktoren wie Interaktion mit dem Fahrzeug und verschiedenen Blickwinkeln (d. h. VR-Ansicht, Aussenansicht und Fahrersicht) wahrnehmen. Wir schlagen einen Ansatz namens SDC-Alabaster (SDC hum**A**n-in-the **L**oop simul**A**tion-**BAS**ed **T**esting s**E**lf-driving ca**R**s) vor, der ein Virtual Reality (VR)-Headset verwendet, um SDC-Testszenarien zu veranschaulichen, das Gefühl zu erzeugen, sich in SDCs zu befinden, und den Benutzern zu ermöglichen, mit der Umgebung zu experimentieren. Unsere Ergebnisse zeigen, dass die Wahrnehmung von Realismus und Sicherheit ohne Hindernisse höher ist als mit Hindernissen, und CARLA war realistischer und sicherer als der BeamNG Simulator mit einem p-Wert > 0,01e-16 und einer Verteilung von 85%($\hat{A}_{12}$). Unsere Ergebnisse zeigen auch, dass Interaktionen mit Fahrzeugen das Gefühl von mehr Sicherheit geben als solche ohne Interaktionen. Mit einem p-Wert > 0,001, und einer Verteilung von 36%($\hat{A}_{12}$), variiert die Wahrnehmung von Sicherheit und Realismus durch die Benutzer, sowohl mit und ohne VR-Headsets, und die Fehlerfälle, die am wichtigsten zu testen sind, werden von den Teilnehmern auch als weniger realistisch angesehen. Darüber hinaus haben wir herausgefunden, dass Faktoren wie die Verwendung eines fortschrittlichen KI-Agenten für Verkehrsfahrzeuge, die Verwendung von Sprachfeedback in VR und die Integration des Fahrverhaltens der Teilnehmer dazu beitragen, dass die Testszenarien realistischer sind und die Wahrnehmung der Sicherheit der Teilnehmer bei simulationsbasierten Tests von SDCs verbessert werden kann.

Contents

1 Introduction **1**
1.1 Problem Statement and Research Questions 2
1.2 Summary of Results & Contributions 4

2 Background and Related Work **7**
2.1 Background on SDCs Simulator 7
2.1.1 BeamNG 8
2.1.2 CARLA 9
2.1.3 Compare Simulator 11
2.1.4 Other simulators for SDCs 11
2.2 Simulation-based testing 13
2.2.1 Simulation-based testing of SDCs 13
2.2.2 Simulation-based Testing of CPS in Virtual Environments 14
2.2.3 Simulation-based Testing of Lane Keeping Systems 14
2.2.4 VR in Cyber-Physical Systems 15
2.2.5 VR and User perception 15
2.3 Thesis Terminology 15

3 Approach **17**
3.1 SDC-Alabaster Architecture Overview 17
3.1.1 Internal Components 17
3.1.2 External Components 19
3.2 SDC-Alabaster Test Case Scenarios Generation and Selection 20
3.2.1 BeamNG 21
3.2.2 CARLA 22
3.3 SDC-Alabaster Workflow 23
3.4 SDC-Alabaster Tool 24
3.5 Technical Aspects 27
3.5.1 Technical Aspect of RQ_1 28
3.5.2 Technical Aspect of RQ_2 28

4 Methodology **31**
4.1 Research questions 31
4.2 Five steps of research methodology 32
4.3 Test case generation and dataset preparation 33
4.4 Study Procedure, Material, and Test Case Environmental setting 34

4.4.1 Pre-survey 36
4.4.2 Environmental setup 37
4.4.3 Log & Survey data collection and analysis 38
4.5 Research method 40
4.5.1 RQ$_3$: How closely does the SDC test case resemble a real-world driving event, and what is the human perception of SDCs test failures/safety? . . 41
4.5.2 RQ$_4$: What is the human perception of SDC's test failures/safety when humans can interact with the car? 43

5 **Results** **45**
5.1 RQ$_3$: How closely does the SDC test case resemble a real-world driving event, and what is the human perception of SDCs test failures/safety? 45
5.2 RQ$_4$: What is the human perception of SDC's test failures/safety when humans can interact with the car? 59

6 **Discussion** **65**
6.1 Safety perception and realism of test cases with various participants' factors (RQ$_3$) 65
6.1.1 Does the participants' age affect their perception? 65
6.1.2 Does the participants' gender play any role in safety perception and realism in SDC test cases? 67
6.1.3 Does the participants' field of expertise or the studies (IT v.s non-IT experts) they've done affect the level of perceived realism and safety of test scenarios? 69
6.1.4 Does the prior experience with VR participants impacting their perceived level of safety and realism? 72
6.1.5 Do the participants' who play pc games as any impact on the level of safety perception and realism in SDC test cases? 74
6.1.6 Does the participants' years of experience in testing as any impact on level of safety perception and realism in SDC test cases? 76
6.2 Lesson learned on comparing the safety perception of the two control groups of participants with one group WITHOUT interaction and another WITH interaction (RQ$_4$) 78
6.3 Does the safety perception change from a different view (outside view versus driver's view)? 80
6.4 Summarizing qualitative aspects from participants on the simulators 82
6.4.1 CARLA simulator 82
6.4.2 Participant' feedback on the experiment 83
6.4.3 Lessons Learned 85

7 **Threat to Validity** **87**

8 **Conclusion and Future Work** **89**
8.1 Future Work 90

9 **Appendices** **91**
9.1 Additional data on the participant 91

List of Figures

2.1 Lidar Sensor [5] . . . 9
2.2 Damage Sensor [5] . . . 10

3.1 SDC-Alabaster Architecture Overview . . . 18
3.2 Virtual Reality (VR) technologies . . . 20
3.3 Road points as SDC test specication . . . 21
3.4 BeamNG with static objects . . . 21
3.5 View angles in VR with vorpX . . . 22
3.6 CARLA with static and dynamic objects . . . 23
3.7 SDC-Alabaster's workflow overview . . . 24
3.8 SDC-Alabaster's Graphical User Interface (GUI) for different use cases . . . 29

4.1 Participants with BeamNG and CARLA Simulator . . . 32
4.2 Overview of research approach . . . 33
4.3 Participants' age distribution . . . 35
4.4 Participant in experiment with VR(P25) . . . 37
4.5 Overall study setup . . . 38
4.6 BeamNG study setup . . . 41
4.7 CARLA study setup . . . 41

5.1 From the graphs, the safety perception with and without VR has an almost similar distribution. . . . 45
5.2 Figure of BeamNG simulator from participants' experiment . . . 46
5.3 Graphs visualizing Safety and realism with test outcome from SDC-Alabaster . . 48
5.4 From the graphs, it seems clear that the CARLA simulator is perceived as more realistic compared to BeamNG . . . 48
5.5 Figure of CARLA simulator from participants' experiment . . . 49
5.6 From the graphs, It shows CARLA slightly more realistic than BeamNG . . . 53
5.7 From the graphs, it seems clear that the safety perception CARLA simulator is higher than simulator BeamNG . . . 55
5.8 Participants' P17 outside view car crash to the roundabout. . . . 56
5.9 From the graphs, it seems clear that when test scenarios are complex safety perception is low compare with simple scenarios(Without obstacles) . . . 57
5.10 From the graphs, the perception increases when participants can interact with the car. . . . 59
5.11 From the graphs, participants can interact with the car, their perception improves. . . . 60
5.12 Figure of CARLA simulator with interaction from participants' experiment . . . 61

6.1 From the graphs, it is evident that participant age has no effect on safety perception, as the distributions are almost identical. . . . 66
6.2 Graphs visualizing level of realism based on age of the participants' . . . 67
6.3 From the graphs, the Gender has an effect on the safety perception. . . . 67
6.4 The graphs make it very clear that one of the women felt test scenarios were more realistic than men . . . 68
6.5 From the graphs, it seems clear that gender has different impacts on the level of realism of the test scenarios. . . . 69

6.6 This graph clearly shows that male participants played more PC games compared to female participants. 70
6.7 From the graphs, it is evident that non-IT participant safety perception higher than IT experts. 70
6.8 The graphs make it very clear that non-IT experts felt test scenarios were more realistic than IT experts 71
6.9 From the graphs, it seems clear that IT versus non-IT experts has major impacts on the level of realism of the test scenarios. 72
6.10 From the graphs, it is evident that participant who already used VR has no effect on safety perception. 73
6.11 From the graphs, it is evident that participants who already used VR felt test scenarios were less realistic. 74
6.12 From the graphs, it is evident that participant who play pc Games has effect on safety perception. 74
6.13 The graphs make it very clear that participants who has experience on PC games felt test scenarios are more realistic than those who don't play PC games. 75
6.14 From the graphs, it seems clear that participants who has experience on PC Games impacts the level of realism of the test scenarios. 76
6.15 From the graphs, it is evident that participant years of experience in testing has no effect on safety perception. 77
6.16 From the graphs, it is evident that participants who had 5 years and more experience in testing felt the test scenarios were more realistic. 77
6.17 It is clear from the graphs that participants in test Group B (which did not involve any interaction) experienced a greater feeling of vulnerability on safety compared to participants in test Group A. (with interaction). 79
6.18 Visualization of Safety perception from the outside view versus the driver's view 80
6.19 The figure of CARLA simulator where the car didn't stop for stop sign 83

9.1 This graph demonstrates that the majority of participants were trying SDC for the first time. 91
9.2 This graph clearly shows most of the participants have some kind of experience driving. 92

List of Tables

2.1 Compare Simulator CARLA vs BeamNG 11

3.1 Overview of components' input and output 18
3.2 Feature overview of HTC Vive Pro 2 and Oculus Quest 2 20

4.1 Test Case 34
4.2 Summarizes participants *(*Higher Professional education includes researchers, professor)* 35
4.3 Pre-survey questions. (MC: Multiple Choice, OA: Open Answer) 36
4.4 Safety levels (Likert-scale intensity) 39
4.5 Simulator Based Survey questions. (MC: Multiple Choice,OA: Open answer, LS: Likert scale (1-5 were 1 is worst and 5 is best)) 42

5.1 Interpretation of the Vargha-Delanay effect size 46
5.2 Statistics for the test scenario WITH or WITHOUT Virtual reality 47
5.3 Qualitative comments on the safety perception 47
5.4 Statistics for the test scenario WITH or WITHOUT obstacle 50
5.6 What makes the safety perception low for CARLA and BeamNG 51
5.5 What makes the safety perception high for CARLA and BeamNG? 54
5.7 Statistics for the test scenario of CARLA and BeamNG simulator 55
5.8 Some qualitative comments on the CARLA simulator's Quality 57
5.9 Some qualitative negative comments on the simulator's realism 58
5.10 Statistics for the test scenario WITH or WITHOUT interaction 60
5.11 Qualitative comments on the CARLA simulator's safety with or without interaction 61
5.12 What makes the interaction test case for the CARLA simulator safer? 61

6.1 Statistics for the test scenario based of different age group. 66
6.2 Statistics for the test scenario based of different gender. 68
6.3 Statistics for the test scenario based on IT versus non-IT experts 71
6.4 Statistics for the test scenario based on participants who already used VR. 73
6.5 Statistics for the test scenario based on participants who play PC Games. 75
6.6 Statistics for the test scenario based on participants years of experience in testing. 77
6.7 Statistics for the test scenario based on test groups. 79
6.8 Statistics for the test scenario based on different views (outside view versus driver's view) 81
6.9 What makes the safety perception low for CARLA and BeamNG 84

9.1 Summarizes participants with which type of vehicle drive and how many years of driving experience they have 92

Chapter 1

Introduction

According to Waymo[1] there were 1.36 million deaths worldwide due to vehicle crashes in 2016 and an additional 836 billion in harm from loss of life and injury each year [64]. This number of tragedies has been growing over the years. Self-driving Cars hold the promise of improving road safety and offering new mobility options to millions of people. Whether they're saving lives or helping people run errands, commute to work, or drop kids off at school, fully autonomously driven vehicles to hold enormous potential to transform people's lives for the better [64].

Testing of self-driving cars (SDCs) has been getting more and more attention from developers and test engineers in recent years. There have been multiple fatal events with Tesla self-driving cars [48, 59, 63], prompting us to consider the necessity of SDC testing, as it may be fatal to human beings if they are not adequately tested to assure safety. Ideally, the tests for life-critical SDCs should ensure trust in the system, but they should also be well integrated into the development process without increasing the costs too much [38].

Testing SDC software in complex (physical) conditions (*i.e.*, with dense traffic and adverse weather conditions) is not only costly but also dangerous, with fatalities already occurring [24]. Virtual testing, in which SDC software is tested in computer simulations, is a more efficient, cheaper, and safer option [2]. Simulators for robotics are used to test robots in a controlled environment without requiring physical hardware [2]. Industrial robots, unmanned aerial vehicles, and autonomous (self-driving) cars have been simulated using popular simulators like Gazebo [36], V-REP [53], and Webots [44].

SDCs represent a specific use case of Cyber-Physical Systems (CPSs). Hence, as for general CPSs, SDCs also face significant challenges in terms of verification and validation for safety assessment to prevent road accidents and traffic congestion [50]. Because Artificial Intelligence (AI) is potentially unpredictable [66], its use in SDCs raises concerns that must be addressed through appropriate verification and validation processes that can address trustworthy AI and safe autonomy. *i.e.*, Deep Blue IBM Watson [2] and AlphaZero (Go) [3] did not know what specific decisions their AI would make for every turn. On the other hand, creating appropriate test scenarios is time-consuming and challenging to replicate the real-world environment [24].

Immersive Computing Technology (ICT), which is another name for Virtual Reality (VR), is a new way to interact with the digital world, which is always changing. VR is often described as a set of technologies that enable humans to have an immersive experience of a

[1] `https://waymo.com/`

[2] `https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/`

[3] https://www.deepmind.com/research/highlighted-research/alphago

world beyond reality [6]. The concept of combining virtual and real-world domains is not new, and it has been used in a variety of contexts, such as the exploration of the International Space Station (ISS) [56], which would allow astronauts to perform mission-critical activities during training sessions (*i.e.*, docking cargo capsules, conducting spacewalks, and performing mission-critical activities) [4], [5]. Another domain where VRs have been used concerns e-health training sessions: surgery planning and surgical simulators on a virtual patient are reality medical healthcare [55], which have a human life-critical nature like SDCs. Moreover, telesurgery, or remote surgery, which uses augmented and virtual reality, represents another critical VR application [18].

In the real world, there are blind spots in the AI of SDCs that can only be fixed by simulating user feedback before putting a person in the same situation as in the real world [33]. In previous work, they showed how incorporating humans into the artificial intelligence training loop can improve SDCs, thereby improving the training efficiency and performance of the deep reinforcement learning algorithm under human supervision [65]. Accidents are unpredictable due to different constraints. AI will not be trained for every possible accident scenario. In contrast, no previous work has attempted to fully immerse humans in the context of simulation-based testing environments; this will be the focus of the thesis. The following section will look at the problem statement and research questions.

1.1 Problem Statement and Research Questions

Previous studies on SDCs proposed tools for test selection, prioritization, and assessing the cost-effectiveness of simulation-based testing for SDCs [8, 9, 21, 35]. Test selection aims to choose only relevant test cases that are likely to fail, whereas test prioritization specifies the order in which the selected tests are executed, allowing faults to be found earlier in the testing process. During the test selection process, certain scenarios may be overlooked or disregarded. In the real world, this could lead to situations or collisions going undetected [27].

SDC-Scissor cuts down on the number of long and complex simulations executed and drastically increases the cost-effectiveness of simulation-based testing of SDCs software [35]. With 12000 tests in a dataset, it was shown that SDC-Scissor achieved a higher classification F1-score (between 47% and 90%) [35].

Due to the limited predictability of safety-critical domains within SDC, some verification and validation issues arise. By having humans evaluate the SDC's [50] software, we could reduce the number of undiscovered scenarios and produce more reliable software [50]. Human assessment and perception of SDCs can be a source of pertinent information for ensuring that SDCs behave safely in a broader range of scenarios [37].

In this thesis, we enhance the test case generation process by including more obstacles and dynamic behaviours, and we use a human-based experiment to verify the validity of the generated test scenarios. In order to give users the virtual experience that they are actually in SDC, we have incorporated virtual reality technology into the simulator. By including a human review in the testing, the procedure improves the test case for this issue. User feedback gives context-specific information about discovered problems or exceptions during automated testing.

To address the aforementioned challenges, we concentrate on technology and research-based questions. Whereas technological research questions are primarily concerned with the

[4] https://www.nasa.gov/specials/trackartemis/

[5] https://www.nasa.gov/marsxr-challenge

technological implementation of a solution to a problem statement, research questions seek to validate data through scientific experiments.

Technological question - $\mathbf{RQ_1}$: How far can we automate test cases to generate a variety of scenarios with various environmental conditions and static and dynamic object placement, and safe and unsafe SDCs tests?

Answering the technical question RQ_1 regarding the extent to which we can generate test scenarios with various environmental conditions and semaphores, such as static objects such as cylinders, road humps, and trees, and dynamic objects There are other vehicles and pedestrians on the road, simulating actual traffic. In addition, to evaluate the performance of SDCs in different weather conditions, These obstacles will simulate real-world scenarios and enhance the SDC's test cases, and we will be able to classify how safe or unsafe the SDCS is with various obstacles.

Technological question - $\mathbf{RQ_2}$: To what extent is it possible to integrate simulator scenarios into virtual reality (VR)?

$RQ_2$is also a technological question that asks how far the SDC's simulator test scenarios can be integrated with virtual reality (VR). VRs may aid in the evaluation of testing scenarios by allowing people to immerse themselves in an immersive experience to test as if in reality, generating significant insights that may reduce the number of RQ_2 test cases. Virtual reality lets people see roads from a different point of view. This brings the scenarios closer to reality and gives useful feedback that helps choose better tests and provides human-based criteria that could support future research on test minimization, such as identifying potentially dangerous road stretches from a human point of view.

$\mathbf{RQ_3}$: How closely does the SDC test case resemble a real-world driving event, and what is the human perception of SDCs test failures/safety?

In RQ_1 and RQ_2, we will generate test scenarios with obstacles to simulate real-world situations and integrate SDC's simulator with VR. In RQ_3, we investigate the degree to which the generated test scenarios resemble the real world and how humans perceive whether a test case is safe or unsafe. also investigates how human perception of safety differs depending on the mode of visualization, such as a desktop screen, VR with an outside view (view from the rear of the vehicle), or the driver's view. In order to answer this question, we conducted a survey-based controlled experiment.

$\mathbf{RQ_4}$: What is the human perception of SDC's test failures/safety when humans can interact with the car?

Similar to RQ_3, RQ_4 we investigate how humans perceive safety when they have some kind of interaction and compare the results without interaction.

$\mathbf{RQ_5}$:(Future work) What are the road segments considered unsafe from a human's perspective?

For the future work of this thesis, we are interested in determining how we can perform a test minimization by considering the road segments to be hazardous from a human standpoint. Test case minimization techniques are used to minimize the testing cost in terms of execution time, resources, etc. The purpose of test case minimization is to generate a representative set from a test suite that satisfies all the same requirements as the original test suite with a minimum number of tests. Collaboration with non-technical users and developers from other fields to collaborate alongside software developers in CPS development and empower users by providing new opportunities such as remote collaboration and training [45].

1.2 Summary of Results & Contributions

Results of our study show that SDC-Alabaster test case classification (pass/fail) closely resembles the human perception of SDCs' test failures/safety (RQ3), and perceptions of safety and realism vary with the simulating environment (*i.e.*, with or without VR). In addition, we discovered that the failure cases that are most important to tests are perceived as less realistic (RQ3). Our results show the perception of realism and safety among users is significantly dependent on the presence of obstacles in the given scenario. Our results also show CARLA was more realistic and safer than BeamNG because participants found CARLA's scenarios more realistic. We also found that interactions with cars make humans safer compared to when there is no interaction. We also found that interactions with vehicles increased humans' perception of safety compared to when there was no interaction. In addition, we discovered that the age, gender, field of expertise, previous use of virtual reality, computer gaming experience, and the number of years of testing experience of the participants all play a significant role in the level of safety and perception of realism.

The contributions of this thesis can be summarized as follows:

- **Extending of SDC-Scissor**: We enhanced SDC-Scissor by incorporating SBST22[6] test generators.
- **Integrate CARLA** *Integrating the CARLA simulator and adding obstacles to scenarios* (RQ1): We have incorporated the CARLA simulator into the testing pipeline. Added static (bumps, trees, cylinders) and dynamic (other vehicles/pedestrians/weather) obstacles to the test case scenarios that were generated.
- **VR simulation integration** (RQ2): Integration of virtual reality (HTC Vive Pro, Oculus Quest 2) with the SDC simulator for visualizing test scenarios.
- **GUI** *graphical user interface to execute commands* : Graphical user interface(GUI) for executing the commands of the pipeline.
- **Interaction** *Interaction between the user and car in real-time test execution* : User interaction with real-time test scenarios utilizing hotkeys such as identifying the relevant road segments that are hazardous, Starting/Stopping the Vehicles, change the viewpoint, and slider to access input allows the user to perceive safety in real-time.
- **Selection & Experiment** *select test cases and conduct a survey for the experiment* : Generate the test case dataset and select/set up the test scenario for both BeamNG and CARLA Simulate and set up Per-survey and On experiment survey.

[6] https://sbst22.github.io/

- **Conducted controlled** *experiment investigating the perception of safety by humans into the testing loop* (RQ3): Set up and conduct a controlled experiment to determine whether humans consider test case scenarios to be safe or dangerous.
- **Analysis** evaluation of the survey results (RQ4): Assessment of the user feedback from controlled experiment and Analyze results.

Chapter 2

Background and Related Work

Self-driving systems, also known as driverless cars, are vehicles that can sense their surroundings using sensors such as cameras, LIDAR, ultrasonics, electrics, and IMU sensors and move in defined lanes without human intervention [69]. Technologies for self-driving cars, primarily work with the computer system by automating vehicle control components. These technological components are capable of a variety of tasks, including fully automated driving, lane-keeping, adaptive drive control, forward collision warning, and anti-lock brakes. A variety of sensors, actuators, and cameras are combined in an autonomous vehicle.

2.1 Background on SDCs Simulator

Several simulation tools have been created to help developers in various phases of CPS design and validation. These methods offer varying degrees of precision and realism at varying execution costs, *i.e.*, simulations with a higher degree of precision typically demand more processing power. In the sphere of self-driving automobiles, engineers utilize abstract simulation models [26, 57], rigid-body simulations [41, 71], and soft-body simulations [21, 52] among others.

Basic simulation models, such as MATLAB and Simulink models, as well as abstract driving scenarios [3], have been deployed primarily for model-in-the-loop simulations, benchmarking of trajectory planners, and hardware/software co-design. They target largely non-real-time operations and lack photorealism, limiting their utility for evaluating SDC systems.

Rigid-body simulations approach the physics of bodies by representing things as indeformable bodies [1]. Rigid-body simulations adopt a very coarse approximation of reality and can only replicate fundamental object movements and rotations. Therefore, rigid-body simulations cannot correctly model actual and crucial circumstances (*i.e.*, vehicle accidents, inertia), even when integrated with rendering engines to generate photorealistic simulations [20].

Soft-body simulations are better than rigid-body simulations and are capable of simulating a vast array of simulation scenarios in addition to simple body movements and rotations. According to Dalboni and Soldati [17], soft-body simulations are capable of simulating body deformations, anisotropic mass distributions, and inertia, which are crucial in many CPS fields. Compared to rigid-body simulations, soft-body simulations are more suitable for replicating safety-critical driving conditions [21], and they may be paired with strong rendering engines to achieve photorealism (*i.e.*, [5]).

There are numerous simulators for self-driving cars, as shown and compared in section 2.1.4 but for our research, we focused on using the BeamNG simulator, which is the soft-body simulation widely used in academic research, as seen in the previous research on simulation-based testing for SDCs [8, 9, 21, 35].In addition, we use the CARLA simulator, which is a rigid-body simulation used more commonly in the industry than in academic research. We also compare the application of soft versus rigid bodies in simulation-based testing for SDCs.

This section provides a brief analysis BeamNG and CARLA simulators and compares them.

2.1.1 BeamNG

BeamNG.tech provides driving simulation software, virtual tests for the development and testing of autonomous vehicles, advanced driver-assistance systems (ADAS) and vehicle dynamics. This is possible thanks to our sensor suite, which is typical of autonomous driving and comprises. Each sensor can be customized to meet the individual needs of different applications [5].

The official open-source Python interface for BeamNG.tech is known as BeamNGpy[1]. The library uses a scenario-based strategy: The user configures vehicles and specifies the sensor configuration in a script. This makes it easier to collect data for learning-based systems and enables the validation and verification of autonomous driving software, such as BeamNG [5].

BeamNG Sensors

- **Camera**:
  Classic camera sensor to incorporate a variety of extra data. This enables it to get the maximum amount of information from the simulation. In addition to RGB photos, it gives pixel-perfect depth, object classes, and object instance information. It is also simply configurable in terms of traditional camera settings, such as field of view and quality, and may be fitted to any portion of your car to match your technology demonstration.

- **Lidar**:
  Lidar sensor implementation mimics the behaviour of real Lidar sensors. As a rotational Lidar, it generates a point cloud by relying on ray tracing. Just like any of our sensors, it is highly customizable and generates a perfect scan of the environment. Soon it will also provide ground truths along the generated point clouds, accelerating your research and product development.as illustrated in Figure 2.1

- **IMU**:
  IMU sensor captures all of the driving dynamics of the agent. While the other sensors provide data for intelligent decision-making, the IMU sensor is designed to account for the passenger's comfort. The successful deployment of non-emergency systems requires a steady driving style. Because our simulation generates precise vehicle dynamics, the IMU sensor provides the means to develop a product that takes the customer into account.

- **Ultrasonic Sensor**:
  Based on the simulation engine's graphical data, a special algorithm is developed to simulate the behaviour of actual ultrasonic sensors. The ultrasonic sensor is an integral part of our sensor suite since it is a standard component of any contemporary car.

[1] https://beamng.tech/

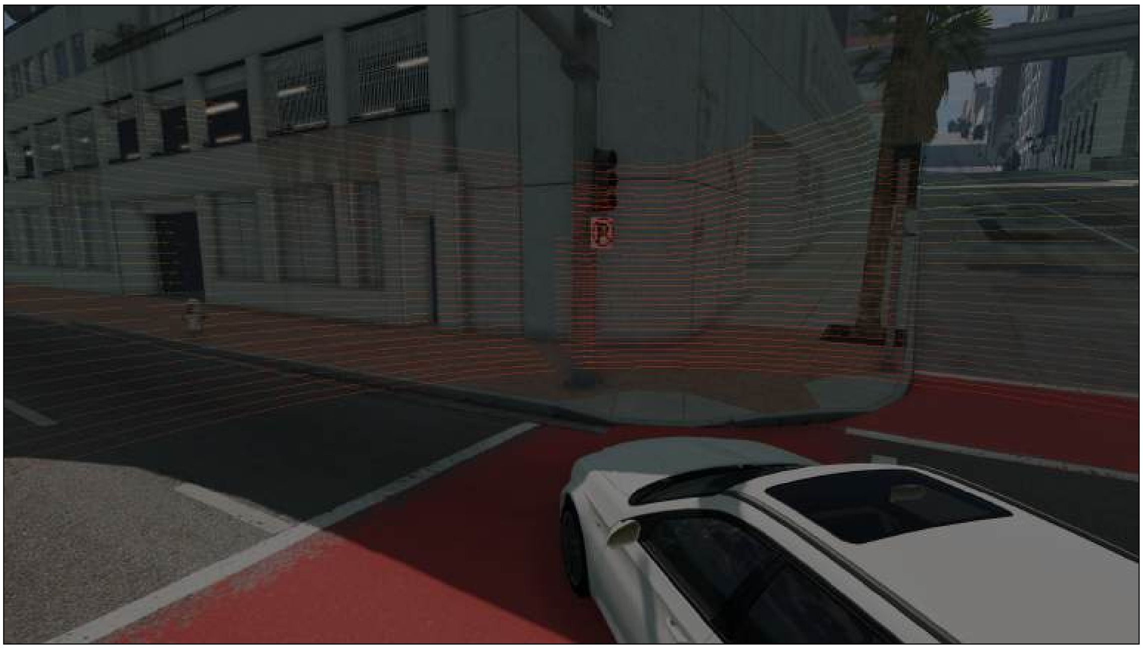


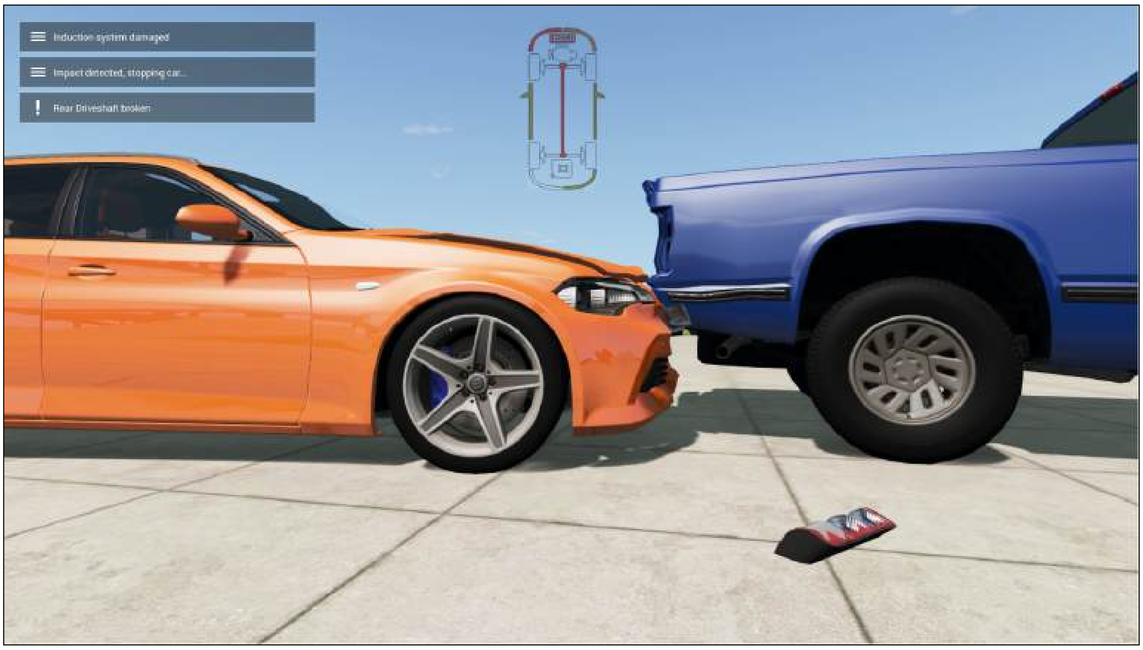
Figure 2.1: Lidar Sensor [5]

- **Damage Sensor**:
  The damaged sensor is exclusive to BeamNG.tech [5] and gives extensive information on the condition of the vehicle. With this sensor, it is possible to evaluate not only the condition of the outside vehicle components but also the condition of the individual engine components. Despite missing a real-world equivalent, this sensor is a vital instrument for evaluating the quality of any autonomous driving helper without endangering any real-world hardware. as illustrated in Figure 2.2

As deformable and breakable objects and fluids can be simulated using soft-body simulations, a variety of simulation scenarios can be modelled using these simulations. To be more precise, the finite element method (FEM) is the primary method for simulating solid bodies, while the finite volume method (FVM) and the finite difference method (FDM) are the primary methods for simulating fluids [43]. BeamNG makes use of soft-body simulations [10].

2.1.2 CARLA

CARLA [19] is an open-source simulator that democratizes autonomous driving research and allows everyone to extend and use this simulator. The simulator has been built for flexibility and realism in rendering and physics simulation. It is implemented as an open-source layer over the Unreal Engine [61]. It simulates a dynamic world and offers an interface for connecting the world with an agent that engages with it. It functions as a modular and adaptable tool with a strong API to support ADAS system training and validation. As a result, CARLA works to satisfy the needs of different ADAS use cases, such as developing perception algorithms or teaching driving rules. CARLA is built from the ground up using the Unreal Engine, and it



Induction system damaged

Impact detected, stopping car...

Rear Driveshaft broken

Figure 2.2: Damage Sensor [5]

makes use of the OpenDRIVE [2]standard to define roads and urban settings. Users can customize the CARLA API, which gives them simulation control. It is based on Python and C++ and is constantly expanding alongside the project, which is an ecosystem of projects created by the community around the primary platform [34].

CARLA Environment is built as a server-client system, where the server runs the simulation and renders the scene, to support this functionality. The interaction between the autonomous agent and the server via sockets is carried out by the client API, which is implemented in Python. The client requests commands and meta-commands from the server, and the server responds with a sensor reading. To generate realistic results, the server should run with a dedicated GPU. The client-side consists of some client modules that control the logic of agents appearing in the scene, including pedestrians, vehicles, bicycles, and motorcycles. [19].

Rigid-body simulations approximate the physics of static bodies (or entities), *i.e.*, by modelling. Basic simulation models implement fundamental signals, but they typically target non-real-time executions and are not very photorealistic. As a result, they are used for model-in-the-loop simulations and hardware/software co-design [10].

Every model is meticulously created to balance visual quality and rendering performance. We use low-weight geometric models and textures while maintaining visual authenticity by meticulously designing the materials and lighting [19].

CARLA proposes a safety assurance module based on the RSS library. The responsibility of this module is to put holds on the vehicle controls based on the sensor information. In other words, the RSS defines various situations based on sensor data and then determines a proper response according to safety checks. A situation describes the state of the ego vehicle in relation to an element of the environment. Leveraging the OpenDrive signals enables the RSS

[2] https://www.asam.net/standards/detail/opendrive/

module to take different road segments into consideration, which helps to check the priority and safety at junctions [34].

2.1.3 Compare Simulator

Table 2.1: Compare Simulator CARLA vs BeamNG

Attribute of comparison	**CARLA**	**BeamNG**
Open Source	CARLA is a fully open-source simulator that democratizes the autonomous driving research area [34]	BeamNG is a mix of commercial and open-source licenses. [5]
Free	Yes	No
Commercial	No	Yes
3D Rendering Engine	CARLA simulator is based on Unreal Engine that generates new maps by automatically adding stop signs based on the OpenDRIVE technology	The BeamNG simulator is built upon the Torque 3D engine, which serves as a backbone for a sandbox vehicle simulator
VR compatible	Yes	No
Pedestrian Simulation	Yes	No
Supported Weather	Yes	No
Recommended System	4GB GPU, 8GB RAM, 50-80GB	Radeon HD 7750 / Nvidia GeForce GTX 550 Ti, 8GB RAM, 15GB
Supported OS	Windows, Linux	Windows, Linux(recently released experimental version)
API support	Python,C++	Python
Architecture	Consists of a scalable client-server architecture.	Focuses on ground-based road-vehicle simulation.
Used	Industry	Academia,Industry
model-based	Rigid-body [34].	Soft-body [10]
Usage in Autonomous systems	Autonmous Driving Research, Synthetic Data Generation for Computer Vision and Machine Learning (Reinforcement Learnig)	Autonmous Driving Research, Synthetic Data Generation for Computer Vision and Machine Learning (Reinforcement Learnig)

2.1.4 Other simulators for SDCs

As outlined in the literature by Kaur et al. [34], we discuss additional potential simulators in this section, with a specific focus on the CARLA and BeamNG simulator comparison.

Matlab/simulink

Automated Driving ToolboxTM[3], which is a collection of tools from MATLAB/Simulink[4], makes it easier to design, simulate, and test automated and driver-assist systems. One of its key features is that High-Definition (HD) live map data and OpenDRIVE® road networks can be imported into MATLAB and used for various design and testing purposes. Users can also model various sensors and create 3D scenarios that are photorealistic. Because Simulink's logic blocks are presented in an understandable manner, MATLAB/Simulink is one of the best options for testing higher-level algorithms. It also has a quick plot function that makes it simpler to analyze the results [34].

CarSim

CarSim [15]is a vehicle simulator commonly used by industry and academia [34]. The newest version of CarSim supports moving objects and sensors that benefit simulations involving ADAS and autonomous vehicles (AVs). CarSim specializes in vehicle dynamics simulations because of its complete vehicle library and variety of vehicle parameters available to tune. However, it has limited ability to build customized upper-level algorithms in an efficient way [34].

Gazebo

Gazebo [36] is an open-source, scalable, flexible, and multi-robot 3D simulator. The physics, rendering, and communication libraries are the three main libraries that Gazebo depends on. First off, the physics library enables the user to specify the physical properties of the simulated objects, such as mass, friction coefficient, velocity, inertia, etc., so that they behave as closely to their real counterparts as possible. Although widely used, Gazebo is not the best option for testing complete self-driving car systems due to the time and effort required to create dynamic scenes [34].

LGSVL

LG Electronics America R&D Center (LGSVL) [54] is a multi-robot AV simulator. In order to test the algorithms for autonomous vehicles, it suggests an unconventional solution. The fact that it is integrated with some platforms makes it simple to test and validate the entire system. The simulator was created using the Unity Game Engine and is open source [34]. Between the AD stack and the simulator backbone, LGSVL offers a variety of bridges for message transmission.

High-quality simulation environments are provided by CARLA and LGSVL, which need a GPU computing unit to operate at a reasonable performance and frame rate. However, LGSVL lacks a built-in recorder, while CARLA does. The user can create a new map by manually importing various components into the Unity game engine, which powers the LGSVL simulator. Due to the numerous integrated automated features they support, LGSVL is most suitable for end-to-end testing of the unique functionalities that self-driving cars offer, such as perception, mapping, localization, and vehicle control, similar to CARLA [34].

[3] https://www.mathworks.com/products/automated-driving.html

[4] https://www.mathworks.com/products/simulink.html

2.2 Simulation-based testing

This section explores the literature on, (i) Simulation-based testing of SDCs, (ii) Simulation-based Testing of CPS in Virtual Environments (iii) VR in cyber-physical systems

2.2.1 Simulation-based testing of SDCs

Autonomous driving has the potential to significantly reduce the number of collisions, however, recently reported fatal collisions using self-driving cars reveal that this vital objective has not yet been reached [21]. This necessitates more thorough testing of the software that governs self-driving cars, which is problematic because it calls for creating complex driving scenarios. We suggest testing self-driving car software primarily in car crash situations, which are the most crucial kind of tests [21].

Virtual tests, which assess CPS like self-driving vehicle software in computer simulations, are more efficient and secure than real-world field operations tests. However, it is time-consuming and challenging to create adequate test cases. Gambi et al. [24] create challenging virtual scenarios for testing self-driving car software automatically by combining procedural content generation, a method commonly used in contemporary video games, and search-based testing, a method proven to be successful in many domains.

Gambi et al. [24] developed a tool called AsFault to automatically generate driving scenarios for SDC testing. AsFault generates virtual tests by procedurally generating road networks within a fixed-size map of configurable size to fit with the capabilities of current simulation software [23].

The cost of running numerous test-driving scenarios (test cases) that interact with simulation engines makes regression testing for self-driving cars (SDCs) particularly expensive. Birchler et al. [10] introduced two black-box test case prioritization strategies, SO-SDC-Prioritizer and MO-SDC-Prioritizer, to increase the cost-effectiveness of regression testing. In both of these methods, the test cases are prioritized using genetic algorithms , which are calculated using the suggested road features and test execution.

To specifically drive the ego-vehicle (the simulated automobile controlled by the SDC software under test) to deviate from the centre of the lane, AsFault uses a genetic algorithm to iteratively refine virtual road networks. A driving simulator called BeamNG [5] can create synthetic, photorealistic photographs of roads, which is how the virtual roads are produced. These qualities led to BeamNG [5] being utilized as the primary simulation platform for the 2022 SBST tool competition [49]. Advanced image processing, deep learning, or machine learning techniques are used by lane-keeping systems to continuously track the striped and solid lane markings of the road ahead and to activate the necessary control mechanisms (such as steering, braking, and speeding) to keep the car at the proper lane [10].

SDCs Safety and Fault Tolerance

Stolte et al. [58] presents a taxonomy that enables the definition of the fault tolerance regimes fail-operational, fail-degraded, and fail-safe in the context of SDCs. A steer-by-wire system, which is a crucial component of future automated vehicles and is used to demonstrate the taxonomy, is an example of a system that is highly safety-critical in general. When using a steer-by-wire system, the desired steering angle is determined and entirely managed by an electronic system.

According to SAE J3016 [11], a taxonomy for an SDC that is SAE Level 4 compliant. Since these SDCs determine how the vehicle will behave, they are extremely important for vehicle safety. The taxonomy and its derived definitions are consistent with the terms fail-operational, fail-degraded, and fail-safe as defined in the technical report ISO/TR 4804 [32]. However, unlike ISO/TR 4804, which is only applicable to automated driving systems at the vehicle level, our taxonomy permits applications at arbitrary system levels.

2.2.2 Simulation-based Testing of CPS in Virtual Environments

To effectively and efficiently test cyber-physical systems (CPS) several simulation environments have been developed, and these environments are nowadays critical for developers at various stages of the design and validation of CPS. Testing CPS in the real world is not only expensive but also dangerous, and has already caused fatalities [25].

CPS are systems that operate simultaneously in the physical and digital worlds [39]. Autonomous vehicles, including cars and trucks, have frequently made headlines in the automotive industry. Military, surveillance and shipping applications for drone-based systems are being developed. Additionally, CPSs are becoming more widespread in a variety of industries and research fields as a result of recent developments in artificial intelligence (AI) and the growing importance of the Internet of Things (IoT) [51].

Dohyeon et al. [67] compare six self-driving simulation platforms with varying levels of visual and motion input, ranging from a screen-based in-lab simulator to a mixed-reality on-road simulator. The simultaneous use of natural visual and motion experiences also increased the sense of presence.

2.2.3 Simulation-based Testing of Lane Keeping Systems

Lane Keeping Systems(LKS) are one of the fundamental features for testing autonomous driving. Simulation-based testing requires the creation of pertinent testing scenarios and the concretization of their executions [40]. Birchler er al. [8] lane-keeping system as the test subject for evaluating the driving agent and driving the car by computing an ideal driving trajectory to maintain lane centre while driving within a configurable speed limit. In accordance with current research on automated testing of LKS [49] [22], we consider scenarios that take place on a sunny day on single, flat roads surrounded by plain green grass. Consequently, tests take the form of the following driving tasks: driving without going off the lane from a given starting position, *i.e.*, the beginning of a road, to a target position, *i.e.*, the end of that road.

SDC-Alabaster relies on the open-source testing infrastructure developed for the Search-Based Software Testing (SBST) workshop's CPS testing competition [22]. This infrastructure can automatically implement executable simulations from road spines, run them, and collect their results (*i.e.*, pass/fail). This infrastructure was chosen for two primary reasons: (1) It employs the BeamNG.tech [15] simulator; consequently, it can conduct physically accurate and photorealistic driving simulations. (2) It has been used to benchmark a number of automatic test generators (see [49] [22]); therefore, it permits us to examine the generality of SDC-Alabaster. Birchler er al. [10] To evaluate the criticality of generated test cases, the road networks are instantiated in a driving simulation in which the ego-car is instructed to reach a destination by following an AsFault-selected navigation route. During the simulation, AsFault traces the ego-position cars at regular intervals in order to identify Out of Bound Episodes (OBEs), *i.e.*, lane departures.

Birchler er al. [8] on paper SDC-Scissor uses machine learning (ML) to identify SDC tests that are unlikely to detect faults in the SDC software under test, allowing testers to skip their execution and dramatically increasing the cost-effectiveness of simulation-based testing of SDC's software. In addition, SDC-Scissor successfully selected unsafe test cases across various driving styles and drastically reduced the execution time to dedicate to executing safe tests in comparison to the random baseline approachcitebirchler2022cost. The classification F1 score for SDC-Scissor was as high as 96% [8].

2.2.4 VR in Cyber-Physical Systems

The Virtual Reality Integrated Development Environment (VRIDE) for CPS could be advantageous to developers in a number of ways, including early and frequent design testing, collaboration with non-technical users, and bringing in developers from other fields to work alongside software developers on CPS development [46].

Although VR technology is most commonly used for gaming, it is also being used more and more in various robotics applications [62]. The Head Mounted Display (HMD) that is included with VR devices enables the user to become completely immersed in a virtual environment. The user can perform actions and manipulation tasks with a pair of hand controllers that can be used to control CPS in the real world. Additionally, the majority of commercial VR devices utilize potent gaming engines like Unreal and Unity that enable users to create a variety of realistic scenarios, intelligent characters, and objects with realistic dynamics and kinematics [62].

VR-based simulators are used to evaluate users' behaviour and subjective assessment in various scenarios. Simulators have the advantage of continually recreating the same complicated and risky scenarios without putting anyone at risk. Simulators frequently use motion platforms to provide motion cues and enhance a sense of motion. Motion platforms are the mechanism that creates the feeling of being in a real motion environment. But due to the constrained motion workspace, it is challenging to mimic realistic motion feedback [70].

2.2.5 VR and User perception

Yildirim et al. [68] demonstrate how virtual reality (VR) is a useful tool for visualizing how future mobility concepts could be implemented in the real world. A subsequent user study investigated how VR can affect attitudes and perceptions regarding such mobility concepts and technologies. Their research indicates that virtual reality is an effective way to quickly and intuitively explain complex ideas, and it may also play a role in broadening user perspectives.

2.3 Thesis Terminology

To avoid confusion in terminology, it is essential to note that throughout the remainder of the thesis, simulation-based test cases for **Self-driving car(SDCs)** are generated by SDC-Alabaster as **test cases**. Test cases are composed of virtual roads composed of sequences of multiple road segments. Formally, **road segments** refers to (parametric) portions of test cases' roads; therefore, they can be straight segments (no curvature), left turns (positive curvature), or right turns (negative curvature).

When a test case is executed in the CARLA or BeamNG simulators, it is called **test scenario**. The test scenario includes **static obstacles** such as road bumps, trees, and cylinders, as well as **dynamic obstacles** such as other vehicles, traffic, and pedestrians.

Test scenarios that have been executed and evaluated in the simulation are referred to as **executed test** cases. Then, if a test is passed successfully, we refer to it as a **PASS test**, and if it fails, revealing potential issues with the system being tested, we refer to it as a **FAIL test**.

Regarding the experiments to answer RQ3 and RQ$_4$ in section 5, we will discuss **safety perception**, which refers to how participants evaluate the test scenarios in terms of their safety. Also, we discuss the **level of realism**, which refers to how realistically the experiment participants relate to the real world. **Participants** are referring to candidates who were recruited for the experiment to assess safety perception and the level of realism of test scenarios.

In the experiment, we utilize **Virtual reality headsets**, also known as **VR**, to visualize test scenarios so that the user feels as though they are in the actual SDCs. To check how safety perception and realism are affected by different view angles of the car, we adjust the view to **Outside view**, which is the view of the car from the back top angle, and **Driver's view**, which is the view as a driver would experience it in real life.

In the experiment, we also provided participants with test scenarios in which they could interact with the vehicle. **Interaction** is an evaluation of how safe they felt on a particular road segment. For assigning the safety perception of the test scenario, participants can respond to the survey using the following metrics: **Very safe** and **Safe** when they feel extremely safe or safe by a significant margin, respectively. **Neutral** when participants do not feel safe or secure. **Unsafe** and **Very unsafe** when the test scenario includes SDC's dangers or is extremely dangerous, respectively.

Chapter 3

Approach

In this chapter, we aim to address the technical research questions RQ_1 and RQ_2. To answer RQ_1, which deals with test generation automation with various environmental conditions, we developed SDC-Alabaster. A tool that extends SDC-Scissor [8] to enable automated test generation and virtual reality (RQ_2) to assess the safety level of SDC test scenarios from a human's perspective.

- **Technological question - RQ_1**: How far can we automate test cases to generate a variety of scenarios with various environmental conditions and static and dynamic object placement, and safe and unsafe SDCs tests?
- **Technological question - RQ_2**: To what extent is it possible to integrate simulator scenarios into virtual reality (VR)?

First, we overview SDC-Alabaster's architecture and explain its components in more depth (Section 3.1). Then, in Section 3.2, we proceed with test case generation for SDCs in VR. To show how the generated tests are used, we explain the general workflow of SDC-Alabaster in Section 3.3. In Section 3.4, we explain in detail how to use SDC-Alabaster as a tool to follow the workflow. Finally, a summary of technical aspects that address RQ_1 and RQ_2 is provided in Section 3.5.

3.1 SDC-Alabaster Architecture Overview

SDC-Alabaster is based on SDC-Scissor and implements additional components (see Figure 3.1). Furthermore, the tool is dependent on external components, such as simulators and VR-related hardware. We provide an architecture overview of SDC-Alabaster with all its components and how they interact with their input and output as illustrated in Table 3.1.

3.1.1 Internal Components

SDC-Alabaster builds upon SDC-Scissor by extending it with additional independent components as shown in Figure 3.1. Specifically, SDC-Alabaster implements (i) a Human Component Interactor (HCI), (ii) an HCI Interpreter, and (iii) an HCI Actuator. Those components interact with each other over APIs by giving inputs and outputs. Below, we elaborate on the individual components in more depth:

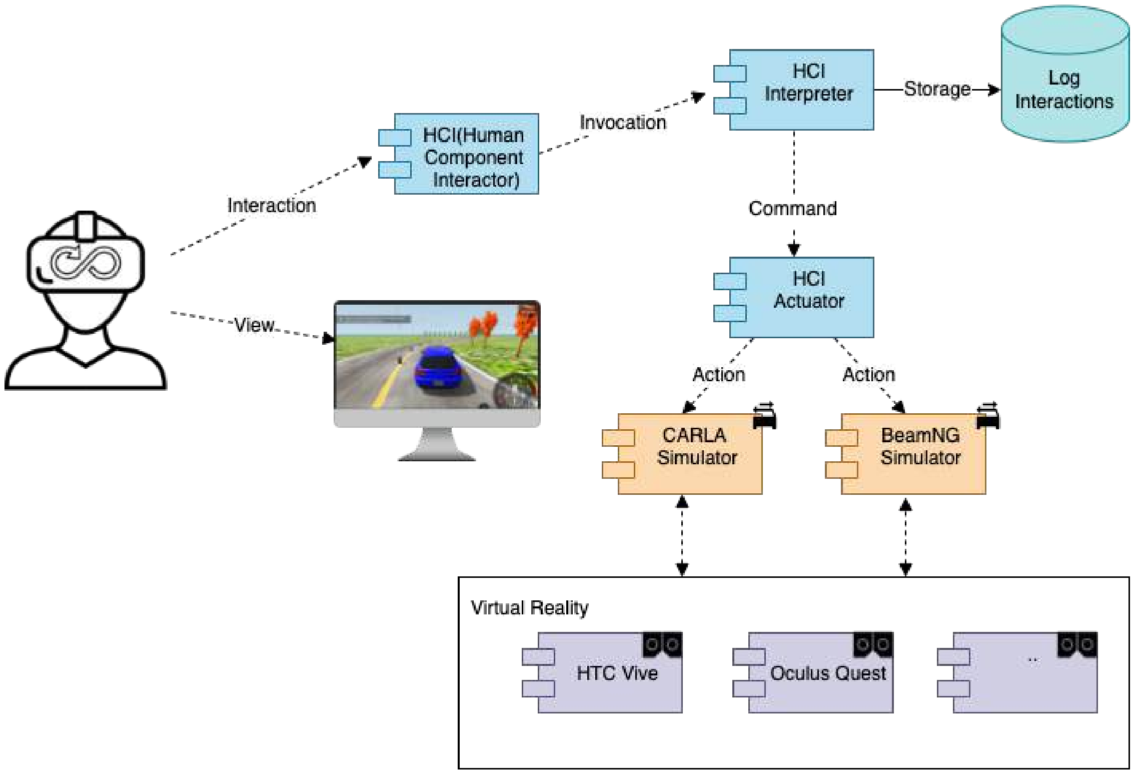


Figure 3.1: SDC-Alabaster Architecture Overview

Table 3.1: Overview of components' input and output

Component	Input	Output
HCI	User interactions	Input representation
HCI Interpreter	Input representation	Generic command representation for simulators
HCI Actuator	Generic command	Specific command for BeamNG or CARLA
Simulators	Specific simulator command	Video output
Virtual Reality	Video output	Immersive video output for VR headsets

Human Component Interactor (HCI). This component is the main component, which a user interacts with. It takes as input any user interaction from the keyboard and forwards it toward the HCI Interpreter component. The next component, the HCI Interpreter, will further process the interaction.

HCI Interpreter. The forwarded interaction from the HCI component is interpreted by the HCI Interpreter and produces a generic command for the simulators. Since not all forwarded interactions from the HCI components are valid (i.e., not supported keystrokes from the keyboard), the HCI Interpreter only allows certain pre-defined interactions for the user. We have to mention that due to technical limitations such as the fact that an AI is driving the car a user can not specify the vehicle's speed manually; for instance, the user is only allowed to set the vehicle's maximum speed with specific keystrokes on the keyboard. The component analyzes the type of interaction and its values so that it can generate a generic command for the simulators.

HCI Actuator. The HCI Actuator gets as input a generic command for the simulators. This component processes the generic abstract command and produces specific commands for each simulator (*i.e.*, BeamNG and CARLA) since the API for the simulators are different and therefore require different implementations. The concrete simulators will be invoked with the generated commands from this component.

3.1.2 External Components

Next to the internal components described in Section 3.1.1, SDC-Alabaster relies also on external components, which were developed by third parties. They are represented as orange and violet components in Figure 3.1. Concretely, SDC-Alabaster requires simulators, such as BeamNG and CARLA as well as VR-related hardware to immerse the user into the simulator.

Simulators

As indicated in Figure 3.1 with orange components, SDC-Alabaster relies on simulators. We use two simulators, namely BeamNG, and CARLA since they implement fundamentally different physics behavior.

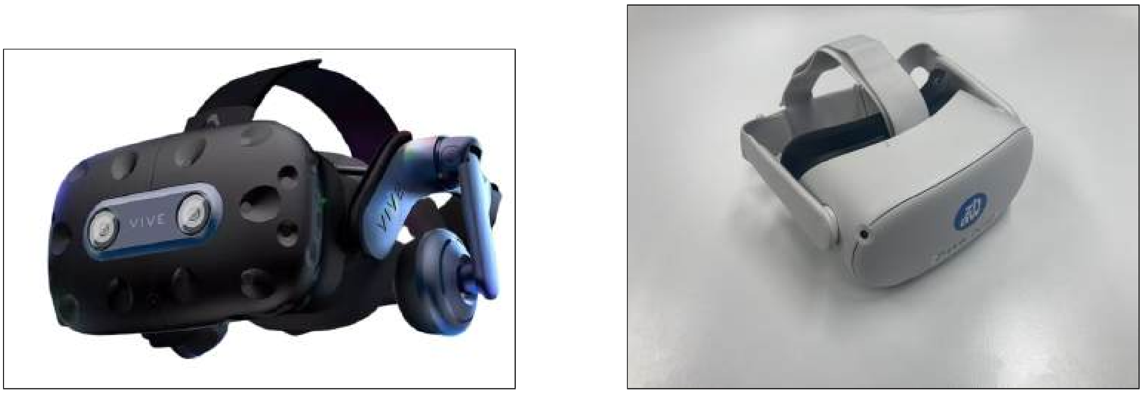
BeamNG. We use BeamNG since this simulator is represented in recent years in academic publications and workshops on SDC testing [7, 8, 10, 22, 35, 49]. The BeamNG simulator comes along with a soft-body physics engine. This type of physics engine allows the simulation of body deformation and therefore more realistic simulations. On the other hand, BeamNG does not allow headless simulations which means that all simulations must be rendered. More details are already elaborated in Section 2.1.1. However, BeamNG provides a Python API called *beamngpy* [4] so that SDC-Alabaster can interact with the simulator and send the actions.

CARLA. Another widely used simulator in industry and academia is CARLA [20, 28, 30, 47, 71, 72]. The differences between CARLA and BeamNG are twofold. On one hand, CARLA comes with a rigid-body physics engine, which works differently than the soft-body physics engine of BeamNG. On the other hand, the test specifications and concepts of these simulators are different. For more details, refer to Section 2.1.2 and Section 2.1.1. Despite those differences, CARLA also provides a Python API [12] to manage the simulations. This allows easy integration into SDC-Alabaster since all components are written in Python.

Virtual Reality

Another external component is Virtual Reality (VR), which is the framed part in Figure 3.1. This component is mainly about VR hardware such as VR headsets to immerse users into the simulation environment. We implemented SDC-Alabaster so that it works with two VR technologies namely (i) HTC Vive Pro 2, and (ii) Oculus Quest 2.

HTC Vive Pro 2. The VR headset HTC Vive Pro 2 (Figure 3.2a) immerses the user in a virtual environment. HTC Vive Pro 2 has no onboard GPU for simulation but a wired connection to an external device with a dedicated GPU.



(a) HTC Vive Pro 2 [29]

(b) Oculus Quest 2

Figure 3.2: Virtual Reality (VR) technologies

Table 3.2: Feature overview of HTC Vive Pro 2 and Oculus Quest 2

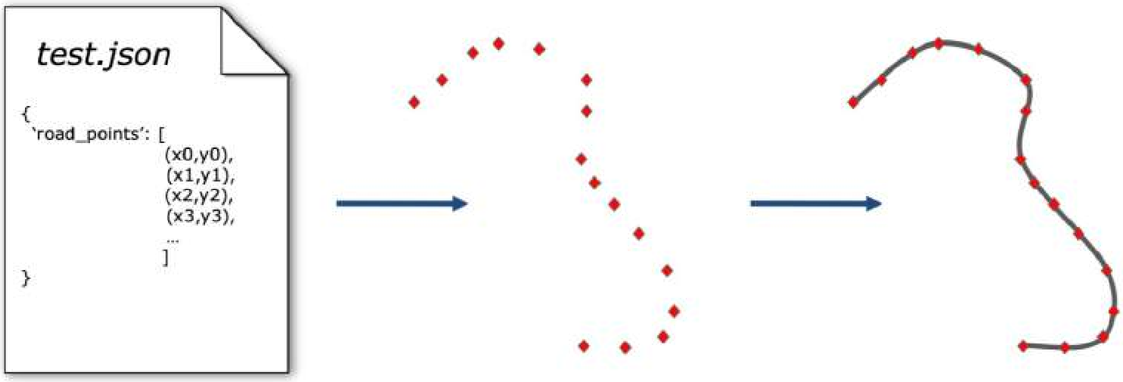
Feature	**HTC Vive Pro 2**	**Oculus Quest 2**
Onboard GPU	✘	✔
Wired connection	✔	✔
Wireless connection	✘	✔
Peripherals	✔	✔
Android OS	✘	✔
Windows OS	✔	✘

Oculus Quest 2. Another popular VR technology is the Oculus Quest 2 (Figure 3.2b). In contrast to HTC Vive Pro 2, the Oculus Quest 2 is able to operate wired and wireless. For the wireless operation of the device, Oculus Quest 2 has an onboard GPU for rendering. However, we suggest using a wired connection to an external, more powerful GPU for better simulation performances.

Both VR technologies come with different features, which enable SDC-Alabaster to be used in different use cases. Depending on the user's needs, the appropriate VR technology can be selected. An overview of the features provided by HTC Vive Pro 2 and Oculus Quest 2 is shown in Table 3.2.

3.2 SDC-Alabaster Test Case Scenarios Generation and Selection

SDC-Alabaster automatically generates a variety of test scenarios for BeamNG and CARLA using the test generators (RQ_1). As already mentioned, SDC-Alabaster is based on SDC-Scissor and applies the same concept for specifying tests for SDCs. Specifically, a test is simply specified in a JSON file by a sequence of XY-coordinates, which are referred to as road points. The actual road in the virtual environments is the result of interpolating the road points as illustrated in Figure 3.3. Creating the tests manually by testers is not feasible since specifying the



```
{
 'road_points': [
                (x0,y0),
                (x1,y1),
                (x2,y2),
                (x3,y3),
                ...
                ]
}
```

Figure 3.3: Road points as SDC test specication

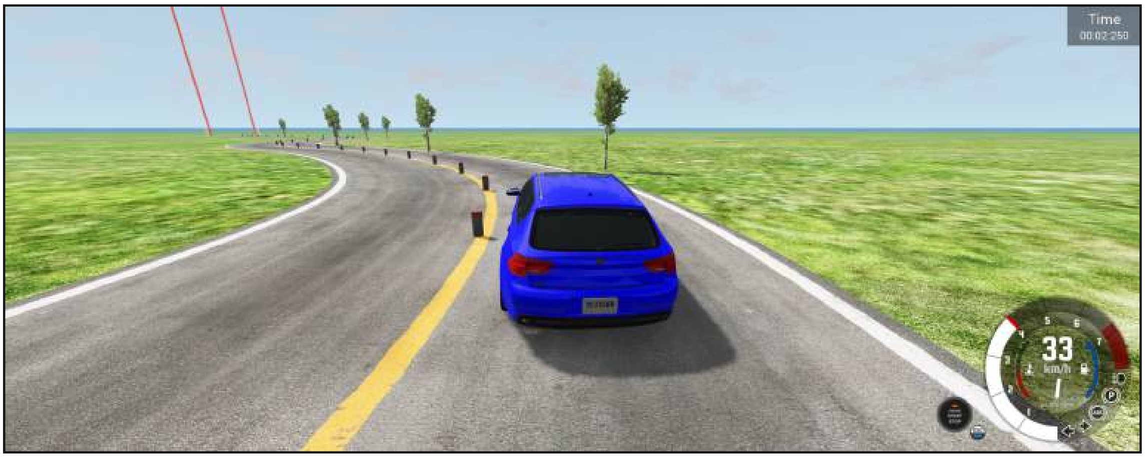


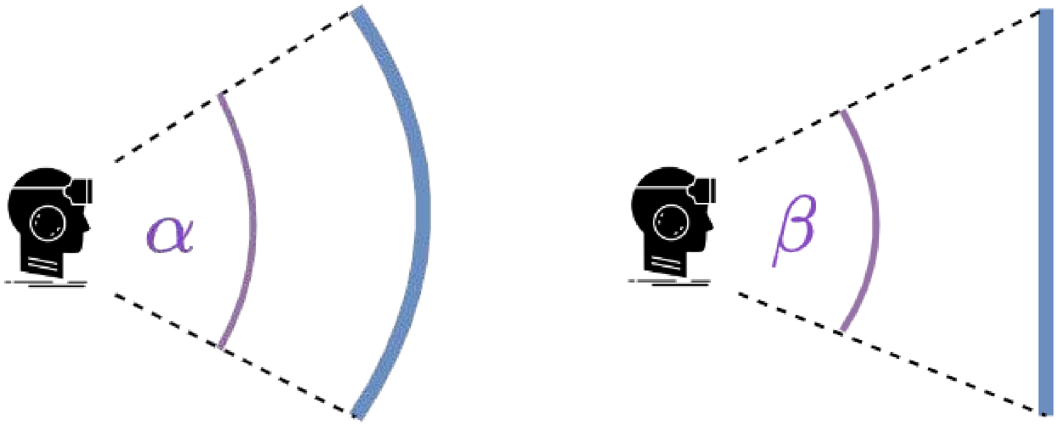
Figure 3.4: BeamNG with static objects

sequences of road points can be a cumbersome task. To overcome this issue, SDC-Alabaster leverages state-of-the-art test generators for SDCs [13, 14, 22, 31, 49] that automatically generate those road points. SDC-Alabaster integrated these tools into its framework and make them applicable for BeamNG and CARLA although, their implementations need to be adapted for each simulator separately since CARLA and BeamNG have different APIs. The following sections provide more details of the implementation of the test generators into SDC-Scissor and eventually into SDC-Alabaster. Furthermore, we will elaborate on the technical use of VR with the different simulators for immersing the users into the virtual environments (RQ_2).

3.2.1 BeamNG

The test generators from the SBST [22, 49] tool competition for CPS are developed by using their own platform [1] in combination with the BeamNG simulator. Since SDC-Alabaster's test specification is based on the SBST tool competition platform, there is a straightforward

[1] https://github.com/se2p/tool-competition-av



(a) Top view with angle α

(b) Side view with angle β

Figure 3.5: View angles in VR with vorpX

integration of the test generators into our framework. More challenging is to enable the immersion of the user into the virtual environment of BeamNG with VR technologies and the integration of static and dynamic objects into the virtual environment.

With SDC-Alabaster, the user can modify the test scenes by adding different types of objects (*i.e.*, static and dynamic). When using SDC-Alabaster with the BeamNG simulator, the user has the option to add trees, speed bumps, and delimiter to the road as illustrated in Figure 3.4. These objects can be placed by specifying the parameter in the test configuration (see Listing 3.2). The integration of objects into the virtual environment should give a different perception of the scenario to the user who is immersed in the simulation scenario.

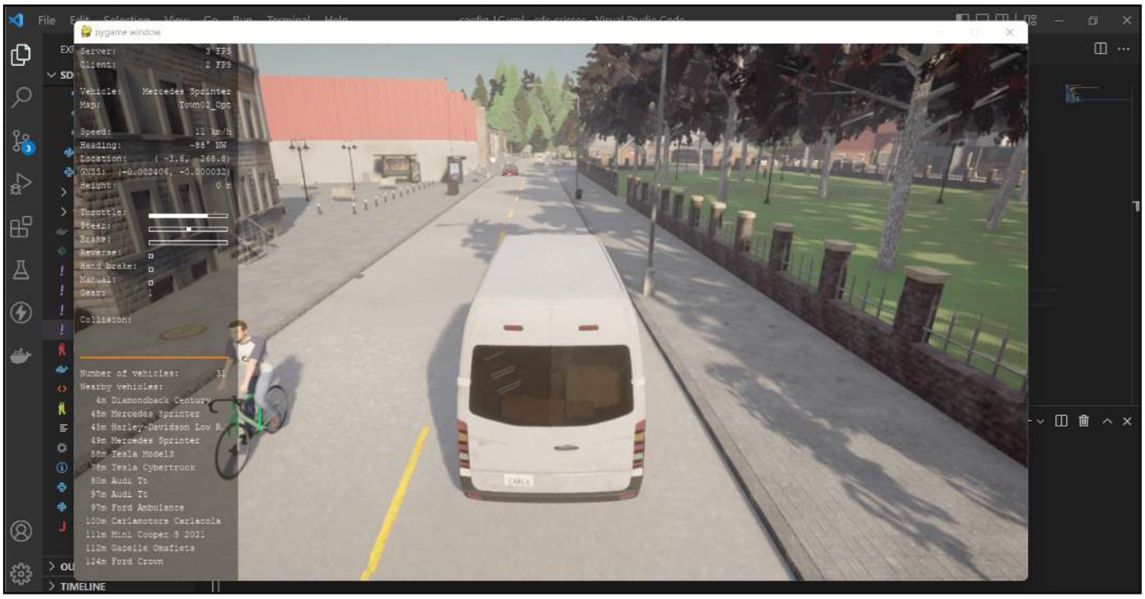
BeamNG has not a built-in solution for using the simulator with VR technologies, thus third-party tools are required to bridge the gap. We used the VR driver vorpX [2], a specialized tool to transform any visual output to the screen to a compatible input for VR headsets so that it gives to a certain extent an immersive feeling for the user. As illustrated in Figure 3.5, the vorpX software gives a broader view angle when wearing a VR headset. The user can move the head and can explore the virtual environment according to its head movement. However, the view is still limited and does not provide a 360° round view for the user.

3.2.2 CARLA

In the case of the CARLA simulator, there is no difference in the use of the test generators. However, tests generated are processed differently compared to the case of the BeamNG simulator since CARLA defines the virtual environment slightly differently. For instance, an automatically generated test contains as mentioned before a sequence of XY-coordinates specifying the road points. The CARLA simulator, however, does not need all the road points defined in the test. Instead, SDC-Alabaster segments the road definition and only uses the start and end points of the segments to declare the beginning and end of the scenario in CARLA.

The integration of additional objects (static and dynamic) into the CARLA simulator allows a more comprehensive understanding of the level of safety perception of the user. With

[2] https://vorpx.com/



```
Server:                 3 FPS
Client:                 2 FPS

Vehicle:     Mercedes Sprinter
Map:                Town02_Opt

Speed:                 11 km/h
Heading:               -88° NW
Location:        ( -3.6, 268.8)
GNSS:  (-0.002406, -0.000032)
Height:                    0 m

Throttle:
Steer:
Brake:
Reverse:       □
Hand brake:    □
Manual:        □
Gear:          1

Collision:

Number of vehicles:         31
Nearby vehicles:
    4m Diamondback Century
   45m Mercedes Sprinter
   45m Harley-Davidson Low R
   49m Mercedes Sprinter
   58m Tesla Model3
   78m Tesla Cybertruck
   80m Audi Tt
   97m Audi Tt
   97m Ford Ambulance
  100m Carlamotors Carlacola
  111m Mini Cooper S 2021
  112m Gazelle Omafiets
  124m Ford Crown
```

Figure 3.6: CARLA with static and dynamic objects

CARLA, the SDC-Alabaster framework can add easily objects to the environment. This modification of the environment can even be done at runtime, *i.e.*, during the test execution. To enable the immersion of the user and manual safety evaluation of the scenario SDC-Alabaster adapts the test specifications for CARLA automatically and uses VR technology to immerse the user into CARLA's virtual environment.

An extra feature of CARLA enables the use of VR for its simulations. SDC-Alabaster utilizes the HARPLab [3] extension project for CARLA to enable the VR integration. The HARPLab project contributes to the development of the VR environment. When you launch the CARLA application, passing the `-VR` flag will put the simulator into VR mode. VR mode is in its experimental phase, so it can only be used for one Carla map, and integrating VR controllers requires modifying the simulator's core and cannot be done through the client API. Details information on setup can be found on the dedicated repository of HARPLab [4].

3.3 SDC-Alabaster Workflow

The general workflow of SDC-Alabaster is illustrated in Figure 3.7. First, the user invokes the test generation process of SDC-Alabaster by choosing a state-of-the-art test generator for SDCs. Secondly, by selecting a simulator, the generated tests are processed according to the chosen simulators. In the last step, the user immerses himself in the virtual environment with VR technology and labels the tests as safe or unsafe.

Test generation. The first step of the general workflow of SDC-Alabaster is to automatically generate test cases with state-of-the-art test generators for SDCs. All test generators that SDC-Alabaster uses come from the SBST [22, 49] CPS tool competitions. SDC-Alabaster

[3] https://github.com/HARPLab/DReyeVR

[4] https://github.com/HARPLab/DReyeVR/blob/main/Docs/SetupVR.md

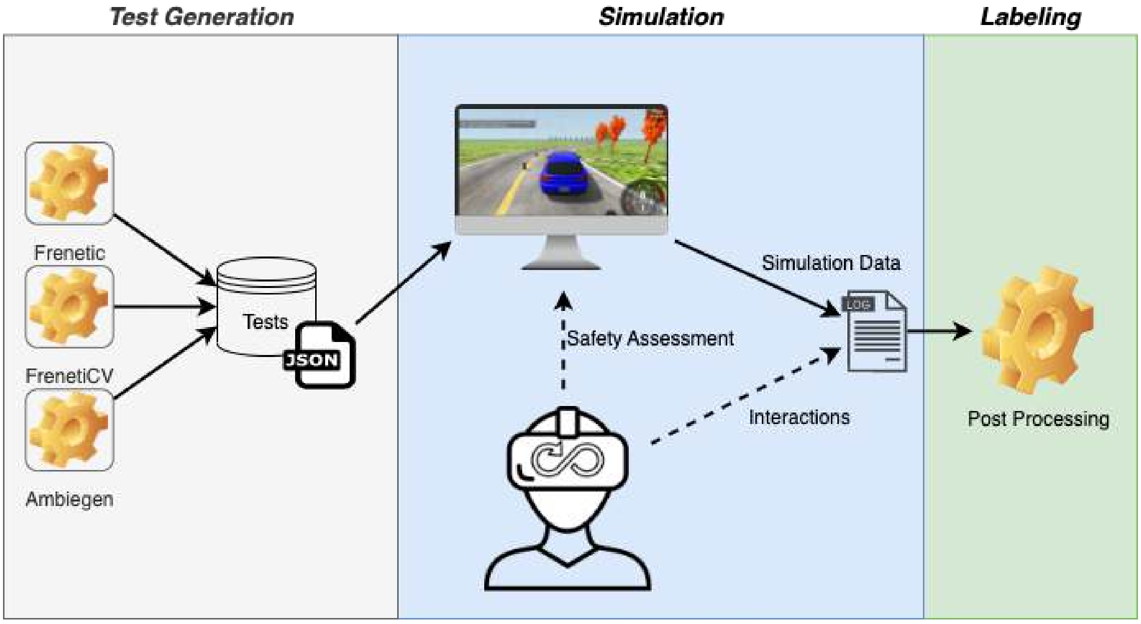


Figure 3.7: SDC-Alabaster's workflow overview

persists all generated tests as JSON files containing the road specifications which will be further processed in the next step for the simulation environments of BeamNG or CARLA.

Simulation. Depending on the actual simulator that will be used, SDC-Alabaster translates the road specifications to the simulator-specific environment. As input, SDC-Alabaster takes the road specifications obtained from the previous phase, and an option specifies which simulator (BeamNG or CARLA) should be used for running the tests. Furthermore, the simulation process is run with VR so that the users can immerse themselves in the environment.

Labeling. The last phase of the workflow is the actual labelling of the tests depending on the user's level of safety perception. On a Likert scale, the user classifies the test into different levels of safety. Furthermore, any interaction a user does (*i.e.*, lower the maximum speed of the SDC) will be logged for further analysis after the test execution.

In summary, SDC-Alabaster will produce a set of tests that are labelled based on the user's perception of safety. These data will be used for further analysis and future research on investigating safety-critical scenarios of SDC test cases in virtual environments.

3.4 SDC-Alabaster Tool

SDC-Alabaster uses APIs written in Python only, and therefore the tool itself is also written in Python. Access to SDC-Alabaster is granted by applying to the owner of the repository [5] on GitHub. This section provides low-level guidance on how to install and use SDC-Alabaster. Furthermore, screenshots of the tool's interface in Figure 3.8 show precisely what input SDC-Alabaster for different use cases need.

[5] https://github.com/ChristianBirchler/sdc-alabaster

Requirements. In order to use SDC-Alabaster, some requirements need to be fulfilled. The following software need to be installed first:

- Windows 10 [6]
- BeamNG.tech [7] v0.24
- CARLA [8] v0.9.13
- Unreal Engine [9] v4.26
- Visual Studio [10] v2019
- Python 3.9 [11]
- Git [12]
- Poetry [13]

We developed and tested SDC-Alabaster with the mentioned versions of the requirements. There is no guarantee that newer versions of the requirements are compatible.

Installation. The installation of SDC-Alabaster consists of two steps; (i) cloning the repository, and (ii) installing the necessary dependencies. Listing 3.1 illustrates the commands to install SDC-Alabaster.

```
~$ git clone https://github.com/ChristianBirchler/sdc-alabaster.git
~$ cd sdc-alabaster
~/sdc-alabaster$ poetry install
```

Listing 3.1: Cloning and installing SDC-Alabaster

[6] https://microsoft.com/en-us/software-download/windows10ISO

[7] https://beamng.tech

[8] https://carla.org/

[9] https://unrealengine.com/en-US/ue-on-github

[10] https://visualstudio.microsoft.com/

[11] https://python.org/downloads/release/python-3915/

[12] https://git-scm.com/

[13] https://python-poetry.org/

Generation and labelling of test cases with YAML file. All the command line options can also be specified in a dedicated configuration file. Listing 3.2, represents an example YAML file that is used to configure the execution.

```
command: ’label-tests’
options:
  home: ’/path/to/beamng/executable’
  user: ’/path/to/beamng/user/folder’
  tests: ’destination’
  rf: 1.5
  oob: 0.5
  max_speed: 50
  interrupt: true
  obstacles: false
  bump_dist: null
  delineator_dist: null
  tree_dist: null
  field_of_view: 120
```

Listing 3.2: Example YAML configuration file

SDC-Alabaster reads all the configs directly from the file as illustrated in Listing 3.3. By enabling the -c flag, a concrete configuration file can be used by SDC-Alabaster.

```
~/sdc-alabaster$ poetry run sdc-alabaster -c file.yml
```

Listing 3.3: SDC-Alabaster with YAML configuration file

GUI to generate test cases. SDC-Alabaster can generate test cases according to parameters, *i.e.*, the test generator and the number of tests that need to be specified. A Graphical User Interface (GUI) helps the user to enter the parameter (see Figure 3.8a). For example, the GUI for generating test cases can be launched by running the command in Listing 3.4.

```
~/sdc-alabaster$ poetry run ./gui/generate-tests.py
```

Listing 3.4: GUI for generating test cases

GUI to label test cases with BeamNG. The labelling process of SDC-Alabaster on BeamNG can also be configured over a GUI (see Figure 3.8b). As demonstrated in Listing 3.5, a separate script starts the GUI so that the tests execute with BeamNG.

```
~/sdc-alabaster$ poetry run ./gui/label-tests.py
```

Listing 3.5: Label test cases with BeamNG

GUI to label test cases with CARLA. The same labelling process can also be performed with the CARLA simulator but with slightly different parameters, as shown in Figure 3.8c. For that purpose, a script specific for CARLA is invoked as illustrated in Listing 3.6.

```
~/sdc-alabaster$ poetry run ./gui/label-tests-carla.py
```

Listing 3.6: Label test cases with CARLA

Run CARLA instance. The use of VR requires additional configuration of the CARLA simulator. An additional `-vr` flag must be set on the command line (see Listing 3.7) so that all VR-related rendering features are enabled. Since CARLA's underlying physics behavior is determined by Unreal Engine 4 [14], a separate process must be run.

```
# -vr flag for to run on VR
~/sdc-alabaster$ ./CarlaUE4.exe -vr
```

Listing 3.7: Run CARLA instance

In summary, we developed SDC-Alabaster user-friendly by making the installation process and the use of the tool as straightforward as possible by guiding by a GUI for each use case. More detailed instructions for the tool can be found in the repository, and support is given over the GitHub platform from the developers. We aim to actively maintain SDC-Alabaster to enhance the research in the SDC domain.

3.5 Technical Aspects

In the context of human-based test assessment for SDCs, SDC-Alabaster focused on using VR to immerse the users into the virtual environments for a more realistic evaluation of safety-critical test scenarios. The following two research questions guided the development of SDC-Alabaster:

- **Technological question - RQ$_1$:** How far can we automate test cases to generate a variety of scenarios with various environmental conditions and static and dynamic object placement, and safe and unsafe SDCs tests?

- **Technological question - RQ$_2$:** To what extent is it possible to integrate simulator scenarios into virtual reality (VR)?

[14] https://unrealengine.com

3.5.1 Technical Aspect of RQ_1

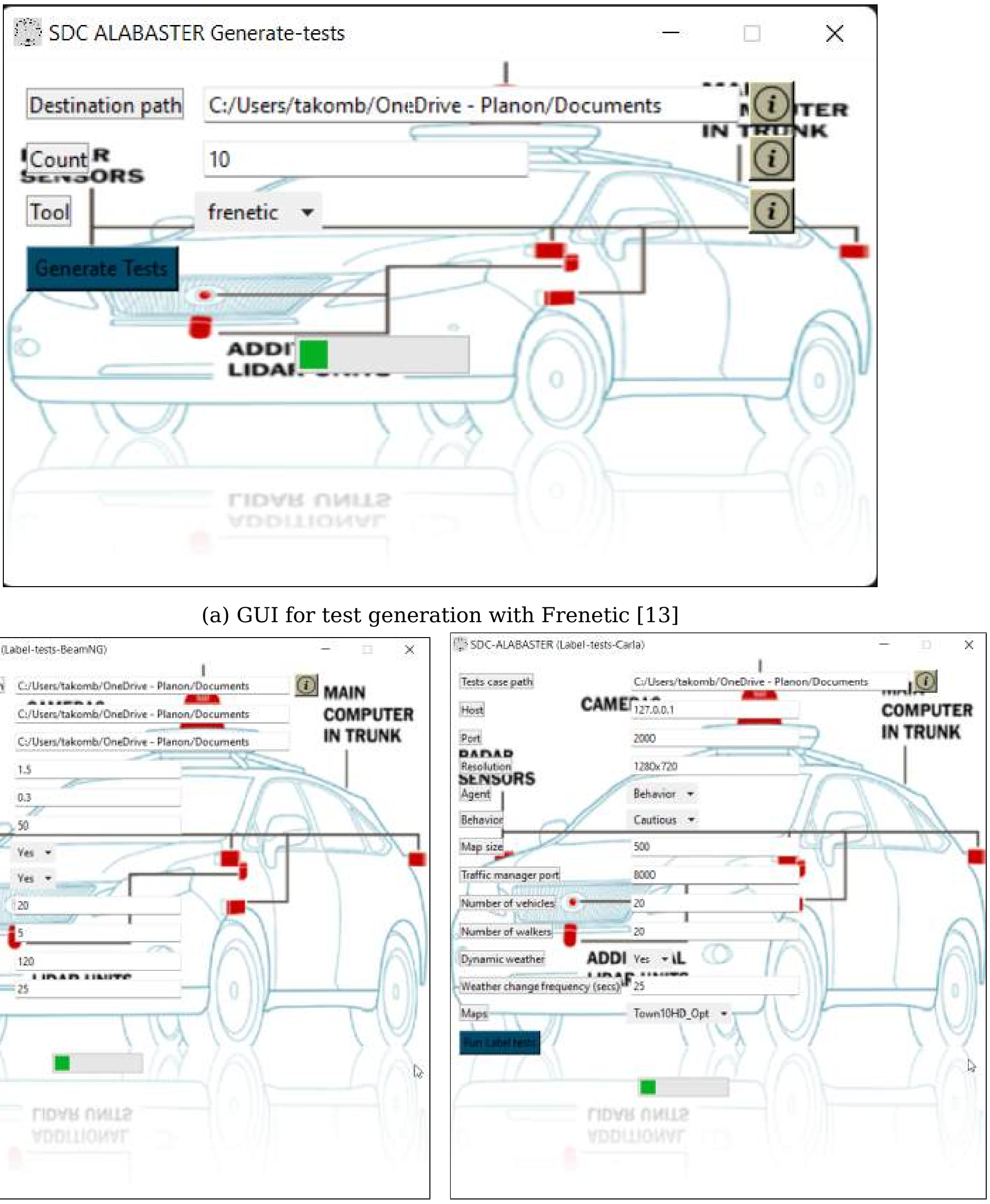
SDC-Alabaster uses state-of-the-art test generators to generate a variety of road specifications (*i.e.*, sequences of XY-coordinates). In addition, SDC-Alabaster can generate different types of objects to place into the virtual environments. However, the generation and placement of these objects highly depend on the simulator in charge of running the test cases.

In the case of BeamNG, static objects can be placed explicitly into the virtual environment by defining some placement parameters (*i.e.*,., XYZ-coordinates and rotation in quaternion). When enabling BeamNG with SDC-Alabaster, the user has the option to place trees, delimiter of the road, and speed bumps in the center of the road. The use of dynamic objects like pedestrians is not supported, but it is possible to introduce other vehicles next to the virtual environment in BeamNG. However, due to the nature of the soft-body physics engine of BeamNG, the interaction between the car and the objects (*i.e.*, a crash) is more realistic compared to simulators with static-physics engines such as CARLA, which do not simulate the dynamic deformations of the objects as a result of an interaction. So far, SDC-Alabaster supports only static objects when using the BeamNG simulator, whereas the CARLA simulator supports both types of objects.

For the CARLA simulator, SDC-Alabaster has the option to add dynamic objects. However, running a test scenario in CARLA with additional static and dynamic objects requires more computational overhead (e.g., running driver agents of other cars) compares to the BeamNG simulation with SDC-Alabaster' framework. Furthermore, the test execution with CARLA takes more time than BeamNG due to the aforementioned computational overhead. For executing a test with CARLA, the simulation needs around 200 seconds whereas BeamNG requires only around 30 seconds. In contrast to the BeamNG simulator, when using CARLA, the user has the possibility to change the environmental conditions at runtime (*i.e.*, during the test execution). Thus, using SDC-Alabaster with the CARLA simulator allows having more complex contexts in the simulation environment. For instance, the vehicle can drive in a city with many other cars and pedestrians, which is not possible yet with BeamNG.

3.5.2 Technical Aspect of RQ_2

For both simulators (BeamNG and CARLA), SDC-Alabaster can translate to video output in a VR-compatible format. However, the approaches differ in how the users are immersed in the virtual environments. In the case of BeamNG, there is neither a built-in solution nor a plugin for the simulator to make the simulator compatible with VR technologies. To overcome this limitation, we used vorpX, a general-purpose video output to VR translator, for immersing the user into the virtual environment. The vorpX tool is a third-party tool that simply makes the video output compatible with VR but with the limitation that the user does not have a full 360° immersive feeling. As reported in Figure 3.5, the view angle by using vorpX is limited. On the other hand, the CARLA simulator does not depend on the vorpX tool. For the specific use case to enable VR with CARLA there is an open-source extension project HARPLab (see Section 3.2.2). This extension allows the user to have a full 360° immersion into the virtual environment of the CARLA simulator. In summary, the VR support for both simulators is limited. However, in the case of CARLA the simulator gives a better immersive feeling whereas, for BeamNG, SDC-Alabaster has to make use of vorpX, a commercial third-party tool for simple video output translation for VR without providing a 360° immersive view of the virtual environment.



SDC ALABASTER Generate-tests

Destination path: C:/Users/takomb/OneDrive - Planon/Documents

Count: 10

Tool: frenetic

(a) GUI for test generation with Frenetic [13]

SDC-ALABASTER (Label-tests-BeamNG)

BeamNG home path: C:/Users/takomb/OneDrive - Planon/Documents

BeamNG user path: C:/Users/takomb/OneDrive - Planon/Documents

Tests path: C:/Users/takomb/OneDrive - Planon/Documents

1.5

oob: 0.3

Max speed: 50

Interrupt: Yes

Obstacles: Yes

Bump dist: 20

Delineator dist: 5

Field of view: 120

Tree dist: 25

SDC-ALABASTER (Label-tests-Carla)

Tests case path: C:/Users/takomb/OneDrive - Planon/Documents

Host: 127.0.0.1

Port: 2000

Resolution: 1280x720

Agent: Behavior

Behavior: Cautious

Map size: 500

Traffic manager port: 8000

Number of vehicles: 20

Number of walkers: 20

Dynamic weather: Yes

Weather change frequency (secs): 25

Maps: Town10HD_Opt

(b) GUI for test labeling with BeamNG

(c) GUI for test labeling with CARLA

Figure 3.8: SDC-Alabaster’s Graphical User Interface (GUI) for different use cases

Chapter 4

Methodology

In this thesis, we investigate how closely the SDC test case resembles real-world driving scenarios, as well as how humans perceive the level of safety.

The first two challenges of this thesis (RQ_1 and RQ_2) are to implement SDC test scenarios with a variety of objects (*i.e.*, trees, cars, etc.), as well as to integrate simulator scenarios into virtual reality, which is already discussed and addressed in Section 3.

We also investigate whether the SDC test case resembles real-world driving scenarios (RQ_3 and RQ_4) and human perception of safety test cases. Also, we investigate human perception from different viewpoints (VR, outside viewpoint, drivers' view) illustrated in Figures 4.1b, 4.1a, 4.1d, 4.1c and assess the safety from different viewpoints. In RQ_3 we only study the test scenario without any interaction, whereas in RQ_4 human will have the possibility to interact with the self-driving car.

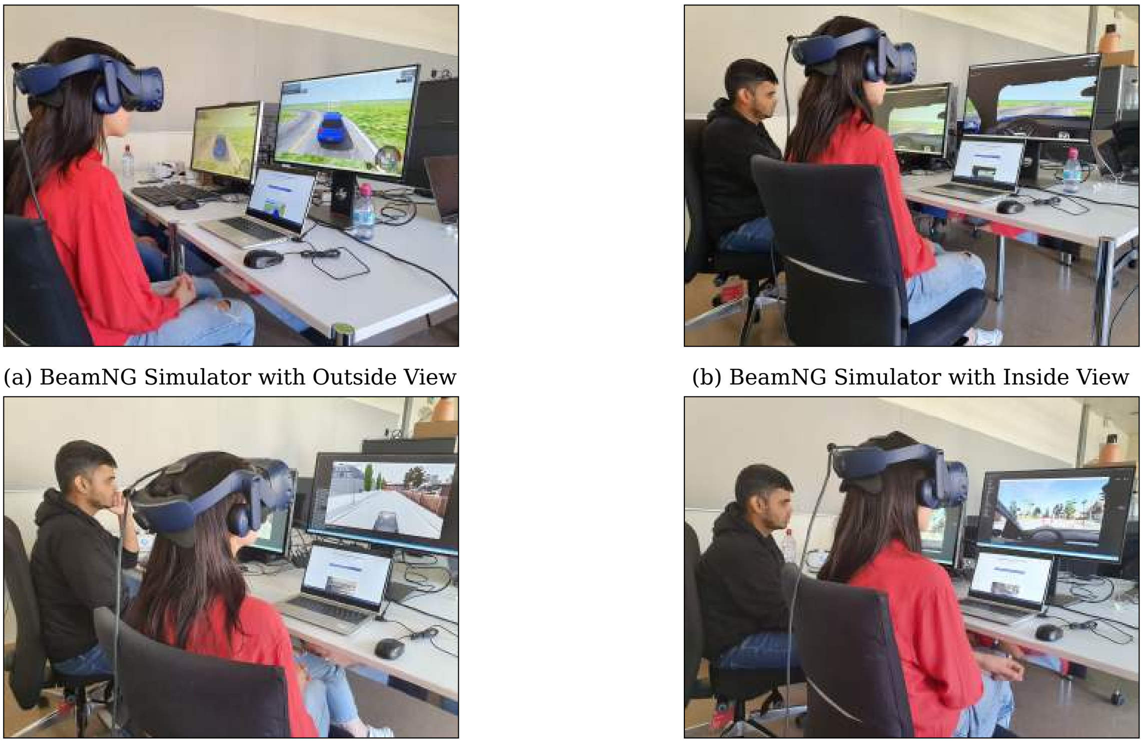
In the following sections, we describe our study's design and the steps we followed in answering RQ_3 and RQ_4 .

4.1 Research questions

We designed experiments to answer our remaining research questions:

- **RQ_3** *How closely does the SDC test case resemble a real-world driving scenario, and what is the human perception of SDCs test failures/safety?*
  This research question addresses the main goal of the study, which is to find out how closely SDC-generated test cases resemble real-world driving situations and how people view SDC test failures and level of safety. We focused on this research question to determine to what extent individuals perceive the test scenarios as safe or unsafe, so we conducted an in-person controlled experiment with a survey to determine safety. No interaction will occur with the car in this research question.

- **RQ_4** *What is the human perception of SDC's test failures/safety when humans can interact with the car?*
  In the preceding research question, we focused on test failures and realism. In this research question, we look at how people's perceptions change with or without interaction. Important differences between RQ_3 and RQ_4 are that in RQ_4 we examine how humans perceive safety when interacting with a vehicle, as well as how they feel from the previously described different perspectives.



(a) BeamNG Simulator with Outside View

(b) BeamNG Simulator with Inside View

(c) CARLA Simulator with Outside View

(d) CARLA Simulator with Inside View

Figure 4.1: Participants with BeamNG and CARLA Simulator

4.2 Five steps of research methodology

Figure 4.2 depicts our research methodology for answering our research questions. We describe the research methodology from experiment setup to survey and log analysis.

1. **Experiment setup**: First, we set up the experiment by planning the overall process and generating a dataset of test scenarios for unbiased sampling. We conducted a pilot experiment with one participant who is a virtual reality expert researcher (age: 23; gender: female). and research with VR, and refine the experiment based on their feedback.

2. **Recruiting**: After setting up the experiment, we recruited study participants. We used various recruiting methods. We targeted dedicated mailing lists from different institutions and organizations (*i.e.*, including non-computer science organizations) . In addition, physical and digital flyers were used to attract more diverse participants. We recruited 41 for our experiment.

3. **Pre-survey**: Once we identified the participants, we sent them a pre-survey a day before the experiment. The pre-survey provided a high-level introduction to our experiment and a disclaimer clarifying that we will proceed with anonymizing data considering and that virtual environment accidents can be experienced during the experiments. To understand participants' backgrounds, we also collected demographic, driving, and VR experience data.

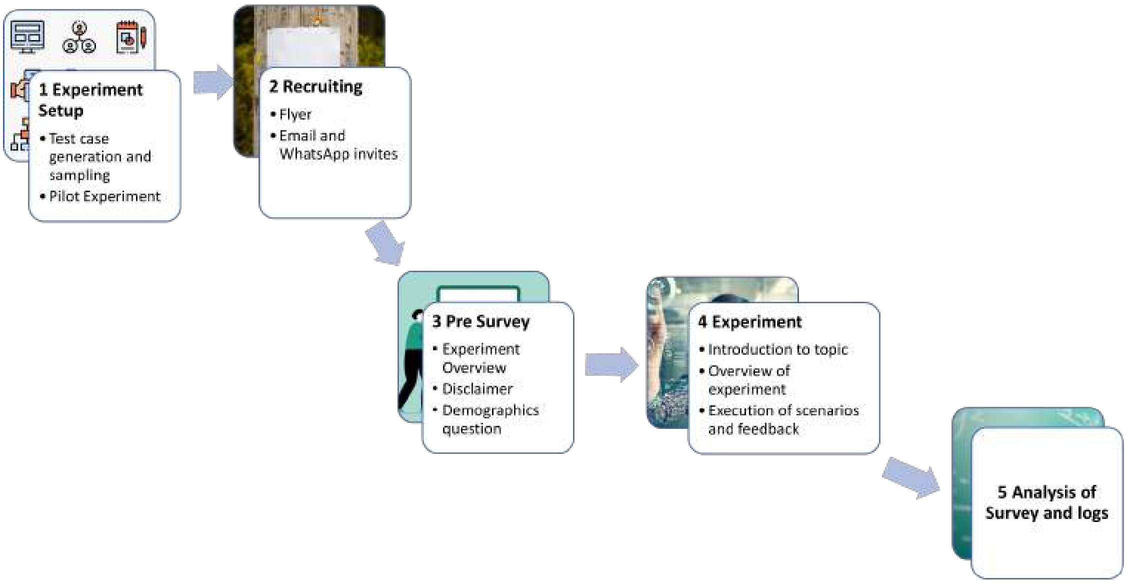


Figure 4.2: Overview of research approach

4. **Experiment**: In the experimentation phase, we gave a brief introduction to the study and a detailed overview of the experiment. We executed various experiments with various perspectives (outside view, driver's view, etc.) and collected feedback via survey with Likert scale and qualitative inputs.

5. **Analysis of survey and logs**: We analyzed the survey and participant logs after the experiment to answer questions RQ_3 and RQ_4.

The subsequent sections elaborate on each of the aforementioned steps.

4.3 Test case generation and dataset preparation

In relation to RQ_3 and RQ_4, we chose a test case dataset and a sampling strategy to select three test cases for the experiment to determine whether SDC test cases are safe or unsafe. The SDC-Alabaster generated test case contains the attributes listed in Table 4.1 *i.e.*, a road point is an array of coordinates of road points, and interpolated road points are the path the vehicle has to take in the test case. We can see an example of a selected test case from SDC-Alabaster in Listing 4.1. We applied probability sampling with a script to select three test cases from a dataset of 1,000 test cases, which was used with both the CARLA and BeamNG simulator test scenarios, as manual sampling can result in bias in dataset selection and reduce sample bias. We used stratified random sampling[1], in which test cases are divided

[1] https://www.investopedia.com/terms/stratified_random_sampling.asp

into smaller, non-overlapping groups. When sampling, it is possible to organize these groups and then draw samples from each group separately.

```
{
  "test_id": 0,
  "test_outcome": "NOT_EXECUTED",
  "predicted_test_outcome": null,
  "test_duration": null,
  "road_points": [
      [
       46.46775185638596,
       75.47341331328971
      ]...
   ],
   interpolated_road_points": [
      [
       46.467751856385966,
       75.47341331328973
      ]...
   ],
   "simulation_data": []
}
```

Listing 4.1: Selected test case

Table 4.1: Test Case

Attribute	description	Example
test_id	Identification of the test case	0
test_outcome	Results of the test execution	NOT_EXECUTED
predicted_test_outcome	prediction of the test execution	null
test_duration	Time taken to execute the test case	null
road_points	Coordinates of road points	[[100,100]]
interpolated_road_points	Driving Path Road Coordinates	[[100,100]]
simulation_data	sensor and simulation data after execution	[]

The test scenario generated by SDC-Alabaster contains roads, whereas in CARLA we use a selection of maps. We divided the generated test case into multiple segments and set a car's destination until it reaches the generated test case's final destination point.

4.4 Study Procedure, Material, and Test Case Environmental setting

Through various channels, including flyers posted on university bulletin boards and email/online chat platforms, we recruit participants. As summarized in Table 4.2, we recruited 41

Table 4.2: Summarizes participants *(*Higher Professional education includes researchers, professor)*

Field of study or profession	**Education level**					**Total**
	*Higher Professional**	*Postdoc*	*PhD*	*Masters*	*Bachelor*	
Computer Science	3	2	5	22	3	35
AI ethics / Political science	-	-	1	-	-	1
Artificial Intelligence	-	-	-	1	-	1
Biology	-	1	-	-	-	1
Robotics	-	-	1	-	-	1
Business administration	-	-	-	-	2	2
Total	3	3	7	23	5	**41**

participants selected entirely voluntarily, *i.e.,* using a convenience sample primarily of researchers and students from universities in Zurich (Zurich universities of applied sciences, University of Zurich, and ETH Zurich) with diverse backgrounds and varying levels of education.

As you can see in Figure 4.3, the majority of the participants were between the ages of 26 and 30. There were 10 women and 31 men in our study. To diversify the results, we will focus on recruiting more participants from fields other than computer science after the completion of the thesis.

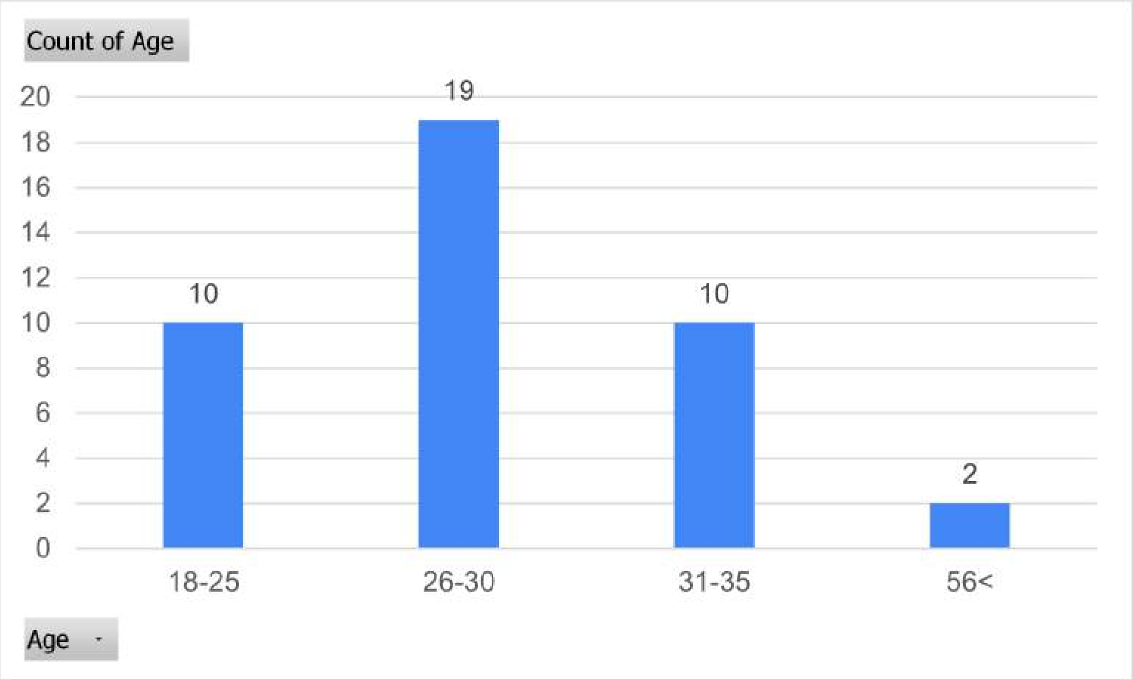


Figure 4.3: Participants' age distribution

Once interested participants contacted us via email, we sent them a link to book a time slot for the experiment. A day before the experiment, we sent an email with location details, time, and a pre-survey to the participants. The following subsection will provide details on the pre-survey.

4.4.1 Pre-survey

The pre-survey contains a high-level introduction to the topic and additional information such as research group collaboration and EU Project COSMOS [2] affiliation, as well as an experimental overview (approximate time, location of experiment, traveling expenses, and recommendation to wear contact lenses). In addition, we described the type of simulator and virtual reality headset used for the experiment. The pre-survey also contains a disclaimer regarding confidentiality and anonymity of personal information, as well as a warning to be prepared for fatalities or accidents.

> **Disclaimer**:
> All the information that you provide will be treated as confidential and will only be used for research purposes. We will not disclose your personal information to third parties. To simulate a real-world setting, some driving simulations may lead to fatalities or accidents. Please consider your readiness for participating to these experiments.

Following the summary information and disclaimer section of the pre-survey, a yes/no question was asked to agree to the disclaimer (**Question:** *Would you accept the above terms and continue participating in the study?*) As soon as they agreed, we moved on to the second section of the survey, which consisted of collecting profile information and experimental questions.

Table 4.3: Pre-survey questions. (MC: Multiple Choice, OA: Open Answer)

Section	ID	Summarized Question	Type	# Responses
Background	Q1.1	Full Name?	OA	41
	Q1.2	Email?	MC+OA	41
	Q1.3	You are a(Student/Professional/Both Student and Professional)?	MC+OA	41
	Q1.4	Education (Currently doing / Completed) ?	MC+OA	41
	Q1.5	Age?	MC	41
	Q1.6	What is your field of study or profession ?	MC+OA	41
Experimental Evaluation	Q2.1	Do you have any programming experience ?	MC	41
	Q2.2	Do you have any programming experience ?	MC	41
	Q2.3	Do you drive motorbikes, cars etc?	MC+OA	41
	Q2.4	What kind of vehicle do you drive ?	MC+OA	41
	Q2.5	How many years of driving experience do you have?	MC+OA	41
	Q2.6	How would you rate your driving skills?	MC+OA	41
	Q2.7	Have you ever use virtual reality(VR) headset (like Oculus Quest, HTC VIVE, HoloLens etc)?	MC+OA	41
	Q2.8	Have you ever been in a self-driving car?	MC+OA	41
	Q2.9	Do you play PC/Consoles Games ?	MC+OA	41

We have grouped the questions reported in Table 4.3 into two topics: (i) *Background*, (ii) *Experimental Evaluation*.

The *Background* questions provided us with demographic information, and the *Experimental Evaluation* questions identified the expertise of participants on various aspects of the experiment, such as testing, driving experience, and the application of virtual reality, which

[2] `https://www.cosmos-devops.org/`

helped us provide a guide to evaluating results and provided a background for safety and realism in virtual self-driving cars.

4.4.2 Environmental setup

As mentioned in the preceding section, participants will receive an email containing the study's location. We conducted this research in a soundproofed, separate room. The configuration of the computing platform included a high-definition Graphical User Interface (GUI) and two monitors (one for participants to view the non-VR test scenarios and the other for organizers to execute the experiment), and for VR scenarios, we used in HTC Vive Pro 2[3], as visualized in Figure 4.4 . After the execution of each test scenario, study participants were able to provide feedback on the experiment in the survey we prepared to guide the whole experiment.

Figure 4.4: Participant in experiment with VR(P25)

The overall experimental setup is illustrated in Figure 4.5, where participants received a brief introduction to the experiment, including a high-level explanation of the study, an explanation of main points such as simulator and VR, and an overview of the experiment. After a brief introduction, we started with the BeamNG simulator's on-screen test scenario (scenarios without VR), which includes three test cases with warmup tasks (no evaluation is required to familiarize oneself with the environment), consisting of observing test scenarios without obstacles and test scenarios with obstacles. This preliminary activity was performed to allow participants to get familiar with the technologies that will be used in the study. Then, as shown in figure 4.6, participants had the possibility to use a VR with an outside view and then a VR with a driver's view.

[3] https://www.vive.com/us/product/vive-pro2/overview/

After a BeamNG simulator, we start with CARLA test scenarios similar to BeamNG. We start with on-screen test scenarios (scenarios without VR), which include three test cases with warm-up tasks (no evaluation is required to familiarize oneself with the environment), consisting of observing test scenarios without obstacles and test scenarios with obstacles. The participants were then given the option of using a VR with an outside view and then a VR with a driver's view to see if their perception of safety changed with a different view from the VR, as shown in Figure 4.7. In addition, experiment with additional test cases in which participants use a keyboard to interact with the vehicle.

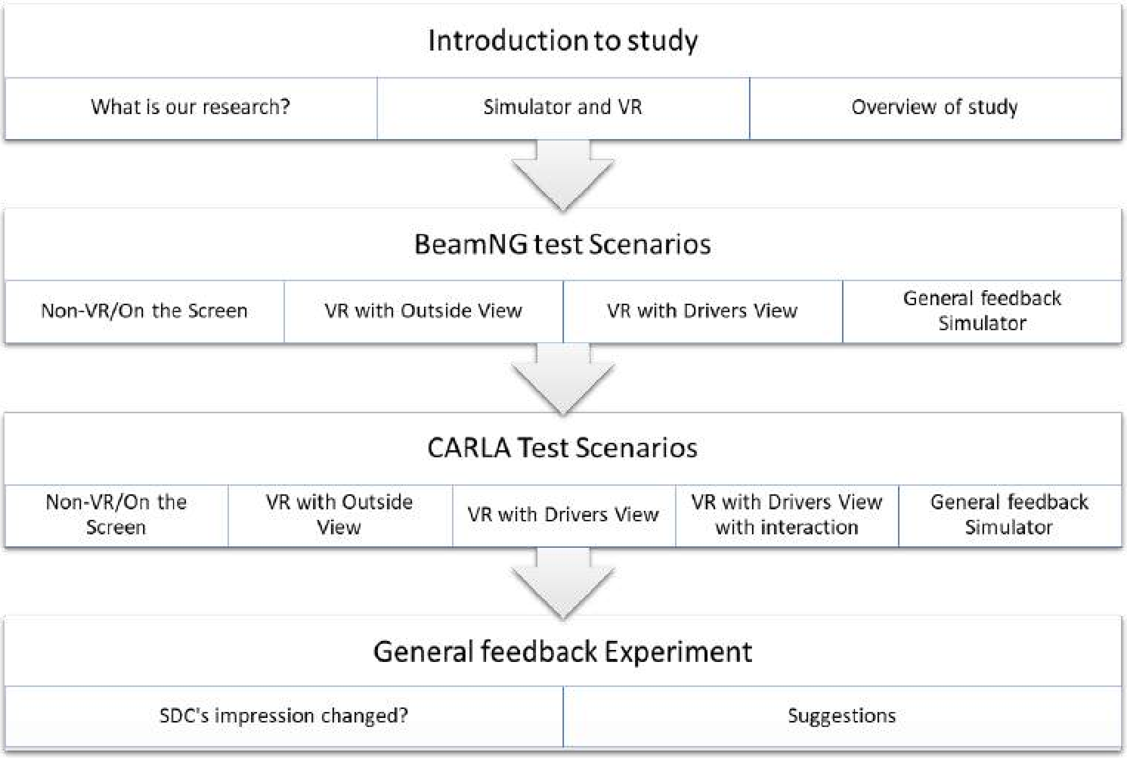


Figure 4.5: Overall study setup

4.4.3 Log & Survey data collection and analysis

For each test scenario, we collected data in a variety of formats, including logs generated by the simulators BeamNG and CARLA and saved them as JSON files. These logs were stored locally on the experiment's computing infrastructure. In addition to the SDC-Alabaster also classifies the test case as safe or unsafe, the framework also records the execution time in the JSON log file.

You can view an example of BeamNG logs by following Listing 4.2. In the log example, one can see the time of logging, the coordinates of the vehicle's position, and sensor data such as fuel, gear, and wheel speed, among others. On CARLA, the structure is slightly different, where you can find timestamps and coordinates, details of weather conditions, and if the car crossed the line. On CARLA, the format is slightly different, with timestamps and coordinates, weather conditions, and whether or not the car crossed the line. The example is found in the following listing: 4.3.

After each test scenario has been executed, we collected feedback information through surveys. We used the Likert scale to assess the level of safety of each scenario (*What is the perceived safety of Scenario 1?*) with options as in Table 4.4 and we asked them to justify their response (*Justify perceived safety of Scenario 1?*).

Table 4.4: Safety levels (Likert-scale intensity)

Safety options	**Example justification**
Very Safe	when the passengers feel it is a dangerous threat
Safe	when the passengers feel it is a marginally less risk
Neutral	when the passengers feel it is a normal situation
Unsafe	when the passengers feel it is risk-free
Very Unsafe	when passengers feel extremely safe in the car

```
{
    "time": 0.03287863731384277,
    "position": [
      49.1695671081543,
      77.75309753417969,
      -27.797107696533203
    ],
    "sensors": {
      "_data": {
        "tcs": 0,
        "fuel": 0.999969815320138,
        "gear": 1,
        "wheelspeed": 0.00015000228758618706,
        ....
      }
    }
  }
```

Listing 4.2: Example BeamNG logs

For both BeamNG and CARLA simulators, we ask the participant to evaluate the simulator based on a number of constraints as reported in Table 4.5. The participants' experimental procedure is as follows:

- An introduction to the topic of research and explanation of VR and simulators, followed by an overview of exploration with the help of survey sections, is used to direct the

```
{
   "simulation_data":[[
     "NO_ACTION",
     46.66442942619324,
     230.62664794921875,
     36.942161560058594,
     0.0016762923914939165
   ],
   [
     "CHANGE_WEATHER",
     46.66442942619324,
     "Sun(alt: 4.73, azm: 191.29) Storm(clouds=48%, rain=8%, wind=40%)"
   ],
   [
     "CAR_CROSSED_LINE",
     46.66442942619324,
     230.62664794921875,
     36.942161560058594,
     0.001712379395030439
   ]]
   }
```

Listing 4.3: Example CARLA logs

participant throughout each experimental session.

- After each testing session, scenarios are visualized and evaluated using post-session questions.

- Each section of the simulator concludes with a final question based on the simulator as shown in Table 4.5.

- We request feedback on the study's impression and feedback as a final step as shown in Table feedback section 4.5 .

4.5 Research method

Each participant completed the experiment by producing two artefacts for each task: (i) BeamNG and the CARLA simulator generated logs. The logs include sensor data like wind speed and car runtime coordinates; in the CARLA simulator, we also log information on dynamic weather conditions. We also log the users' interactions with the SDC into the logs in interaction scenarios; (ii) Level of safety evaluation with both close-ended and open-ended questions for statistical and qualitative analysis.

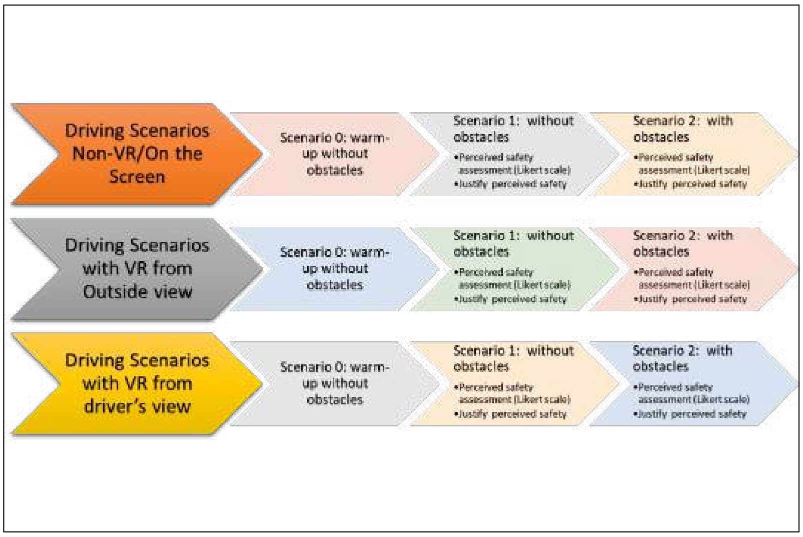

Driving Scenarios Non-VR/On the Screen
Scenario 0: warm-up without obstacles
Scenario 1: without obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Scenario 2: with obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Driving Scenarios with VR from Outside view
Scenario 0: warm-up without obstacles
Scenario 1: without obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Scenario 2: with obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Driving Scenarios with VR from driver's view
Scenario 0: warm-up without obstacles
Scenario 1: without obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Scenario 2: with obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety

Figure 4.6: BeamNG study setup

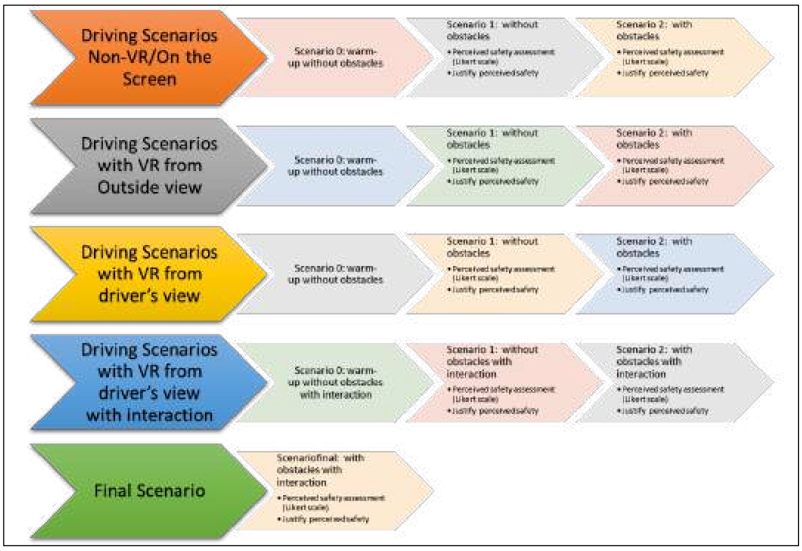

Driving Scenarios Non-VR/On the Screen
Scenario 0: warm-up without obstacles
Scenario 1: without obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Scenario 2: with obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Driving Scenarios with VR from Outside view
Scenario 0: warm-up without obstacles
Scenario 1: without obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Scenario 2: with obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Driving Scenarios with VR from driver's view
Scenario 0: warm-up without obstacles
Scenario 1: without obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Scenario 2: with obstacles
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Driving Scenarios with VR from driver's view with interaction
Scenario 0: warm-up without obstacles with interaction
Scenario 1: without obstacles with interaction
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Scenario 2: with obstacles with interaction
•Perceived safety assessment (Likert scale)
•Justify perceived safety
Final Scenario
Scenariofinal: with obstacles with interaction
•Perceived safety assessment (Likert scale)
•Justify perceived safety

Figure 4.7: CARLA study setup

4.5.1 RQ$_3$: How closely does the SDC test case resemble a real-world driving event, and what is the human perception of SDCs test failures/safety?

A participant's experience with the SDC-Alabaster can be analyzed by understanding the resemblance that the users felt between the real world and the SDC test scenarios. In order for us to better understand this resemblance we analyze the participant survey scores and the qualitative feedback that was collected during the experiments. The primary focus would be the safety perception of the users i.e how safe the users felt while driving in the virtual test

Table 4.5: Simulator Based Survey questions. (MC: Multiple Choice,OA: Open answer, LS: Likert scale (1-5 were 1 is worst and 5 is best))

Section	ID	Question(s)	Type	# Responses
Evaluation based on BeamNG	Q1.1	How would you scale the level of realism of scenarios generated test cases in the BeamNG simulator?	LS	41
	Q1.2	Justify the level of realism of scenarios generated by test cases.	OA	41
	Q1.3	How would you scale driving of AI BeamNG Simulator?	LS	41
	Q1.4	How would you scale overall experience with BeamNG Simulator?	LS	41
	Q1.5	Justify overall experience with BeamNG Simulator?	OA	41
Evaluation based on CARLA	Q1.1	How would you scale the level of realism of scenarios generated test cases in the CARLA simulator?	LS	41
	Q1.2	Justify the level of realism of scenarios generated by test cases.	OA	41
	Q1.3	How would you scale driving of AI CARLA Simulator?	LS	41
	Q1.4	How would you scale overall experience with CARLA Simulator?	LS	41
	Q1.5	Justify overall experience with CARLA Simulator?	OA	41
	Q1.6	How do you compare safety with Interaction and without interaction?	OA	41
Feedback Section	Q3.1	Did this experiment change the way you thought about the Self-driving Cars safety?	MC	41
	Q3.1	Please write in a few words on your experience and suggestions.	OA	41

scenario. In order to further enhance our findings, we also analyze the qualitative feedback provided by the users while trying out the experiment WITH or WITHOUT virtual reality. The former is a case where the user partakes in the experiment primarily conducted on the screen of a computer while the latter focuses on the user experience when he/she is wearing a VR headset.

To understand the users' perception of safety and realism in the second scenario i.e with Virtual reality, we compare two pioneering simulators namely BeamNG and CARLA. The responses that are recorded in the experiment were both close-ended as well as open-ended. While the close-ended questions are used to derive statistical inferences about the users' safety and realism perception, the open-ended answers are helpful in getting an in-depth qualitative analysis about the same. For the research question in discussion *i.e.,* RQ_3, we mainly analyse responses that were recorded in the non-interactive sessions which had almost no user interventions.

A variety of visualization techniques were used to analyse and compare the safety and realism perceptions with and without virtual reality Stacked histograms were used to understand the spread and structure of the dataset that we condensed using the responses acquired from the users. To get the statistical analyses for the data, we make use of boxplots to visualize the statistical variables like mean and quartiles across the different attributes of the dataset. However, these boxplots were also helpful in determining if the difference between the scores that depicted the safety perception of users with and without virtual reality was significant.

We compare safety perception with and without virtual reality using stacked histograms, and also we used statistical tests using boxplots and visualize and determine whether the difference between the scores is statistically significant (e.g Safety perception with and without virtual reality).

In order for us to understand the degree of resemblance between the SDC test case and real-world driving, we thoroughly examine the results of the survey for each simulator section. Once again, visualizations played a vital part in helping us to arrive at a conclusion about the difference in the results between CARLA and BeamNG simulators. Stacked histograms were used to understand the spread of data across the two simulators and box plots were used to understand the distribution of the logs recorded in the SDC-Alabaster in terms of measures of central tendency like mean and Inter-Quartile range. This aided us in verifying whether the test case was classified as a failure or a success. In addition to this, box plots were utilized by us to determine the extent to which the results from SDC-Alabaster matched the level of perceived realism.

The safety perception and level of realism between the CARLA and BeamNG simulators had to be compared and analyzed. To achieve this, stacked histograms once again played a vital role along with the various statistical tests, which include the Shapiro Wilk test to verify normality, the Wilcoxon rank sum test, and Vargha- Delaney statistic to determine the effective size. As expected, we found that the safety and realism for all participants differed as the complexity of the scenario also increased. To account for this variation, we divide the results into two groups. The first group refers to the test case which has no obstacles in the environment (cars, bumps, pedestrians). The second one is the scenario where the obstacles mentioned above are all included in the scenario. This gives us a clearer picture of how aspects of safety and realism were perceived by different users under different circumstances. Visualizing the results using a box plot, helps us further understand the differences with greater clarity.

4.5.2 RQ$_4$: What is the human perception of SDC's test failures/safety when humans can interact with the car?

While testing Self Driving Cars in a simulated environment which was unfamiliar to many people, we wanted to verify the effect that interactions have on the users' perception of safety and failures that they witness in the simulator.

In order to analyze this difference in perception of safety among the users, we analyze both the close-ended questionnaire *i.e.,* participant survey scores and open-ended questions including qualitative feedback. Since we performed tests with active interaction with the user on the CARLA simulator we compare data obtained for the test cases on that respective simulator. Histograms are used again to understand the spread of data across the two categories under consideration *i.e.,* test cases with interaction and test cases without interaction. To analyse the measures of central tendency across the two categories, box plots are used along with the statistical tests discussed in RQ$_3$ to establish and prove the statistical significance of the difference between the two classes under question.

Chapter 5

Results

In this section we report all the findings from the experiments with the aim of answering the research questions formulated in Section 4.

5.1 RQ$_3$: How closely does the SDC test case resemble a real-world driving event, and what is the human perception of SDCs test failures/safety?

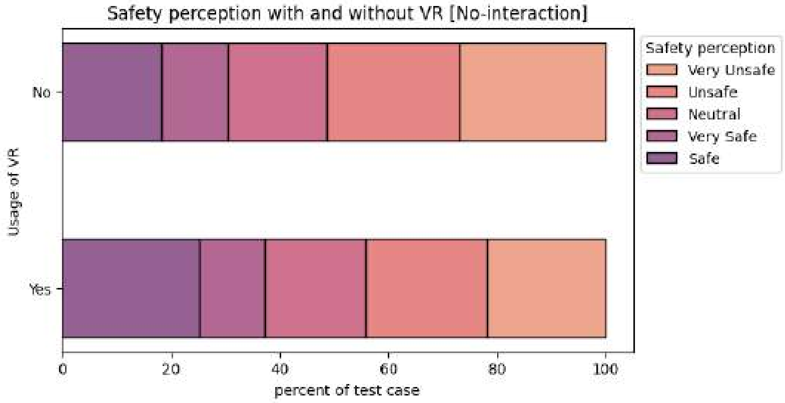
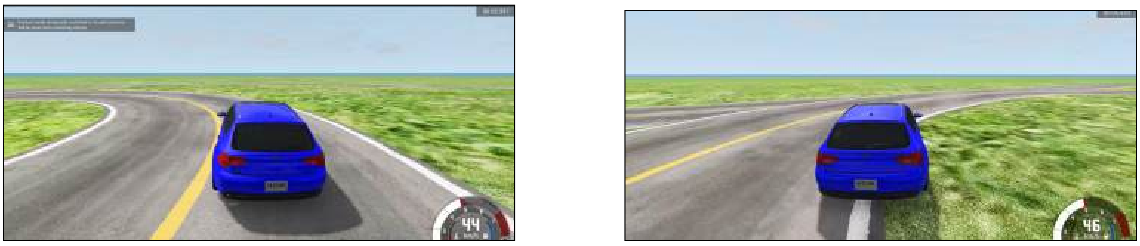


Figure 5.1: From the graphs, the safety perception with and without VR has an almost similar distribution.

As mentioned in section 4.5.1 we made use of proper visualizations to get a better understanding of the results. Figure 5.1 depicts the stacked histogram of the proportion of test cases with different safety perceptions WITH and WITHOUT a VR headset. We observed that the participants' perception of safety was higher WITH the VR headset than they had WITHOUT it. It is also depicted in the histogram that is seen in Figure 5.1, a greater proportion of participants felt safe or Very safe in the same scenario WITH VR than in those WITHOUT VR. Judging from the comment, *"It felt very unsafe, with the car crashing with multiple objects. Again, the car didn't even stop accelerating after having severely crashed and being stuck with one of the ob- stacles"*(P4), *"I Felt more unsafe and the VR view made me more worried about the incident than without it"*(P1) who were wearing a Virtual reality headset in 5.3 we can see that they felt extremely uncomfortable in the given environment which can be attributed to the level of realism in the VR environment. We argue that the users when not wearing VR headsets felt safer than the situations where they had to since VR simulates a scenario which is much closer to reality than a typical monitor screen. This might also be attributed to some users not being very familiar with the environment simulated by Virtual Reality.



(a) Car about to crossing the Line (P27)

(b) Car out of the road (P30)

Figure 5.2: Figure of BeamNG simulator from participants' experiment

Table 5.1: Interpretation of the Vargha-Delanay effect size

Significant	**effect size** $\hat{A}_{12}$
negligible	2|$\hat{A}_{12}$-0.5| <0.147
small	2|$\hat{A}_{12}$-0.5|>=0.147 & <0.33
medium	2|$\hat{A}_{12}$-0.5|>=0.33 & <0.474
large	2|$\hat{A}_{12}$-0.5| >0.474

Afterwards, we performed a statistical analysis of the data. Statistical tests play a major role in understanding and analyzing the behavioural patterns found in the dataset. This also helps us establish evidence, which helps us prove our hypothesis. The first phase of performing statistical analyses was to understand if the number of safe or unsafe test cases would follow a normal distribution $p < 0.01$. To this end, we made use of the Shapiro-Wilk test of normality, which revealed that neither of the parameters of interest followed a normal distribution. The p-value threshold in this approach was set to 0.05 (as a rule of thumb), which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As we see in Table 5.2 The p-value of the level of safety WITH and WITHOUT VR is 0.15, which is greater than 0.05, which means there is no statistically significant difference between the perception of safety between the two simulations used, which are WITH and WITHOUT VR.

Finding 1. From the results obtained we observe that the users' perception of Safety and realism varies with the simulating environment (*i.e.,* with or without VR). Analysing these experiences recorded in the data we can conclude that the users tend to be safer and close to the real world when equipped with the VR headset. These results were not statistically significant.

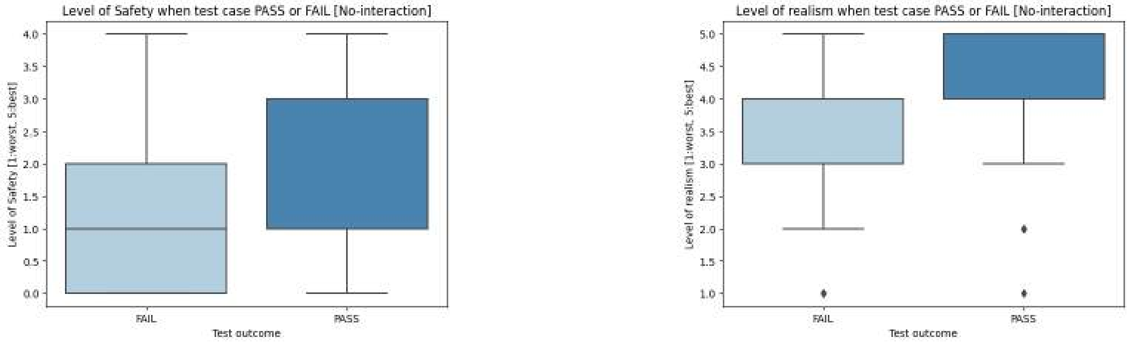
Table 5.2: Statistics for the test scenario WITH or WITHOUT Virtual reality

Variable	**Factor**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety	WITH WITHOUT	0.0 0.0	1.82 1.64	4.0 4.0	0.01e-12 (non-gausian) 0.01e-8 (non-gausian)	0.15	-
Test outcome	WITH WITHOUT	0.0 0.0	0.37 0.45	1.0 1.0	0.01e-24 (non-gausian) 0.01e-18 (non-gausian)	0.12	-

Table 5.3: Qualitative comments on the safety perception

Participants Code	**VR**	**Safety perception**	**Simulator**
P7	NO	Very Safe	*"The car followed traffic signal and was almost all the time within the line. Only lane changing looks bit weird."*
P7	YES	Very Safe	*"It was all safe until it onto the pavement of the roundabout."*
P4	No	UnSafe	*"While the car managed to detect the road correctly, it deviated a bit from it on a couple occasions. Even though it rightly reached the end, in a real-life scenario deviating from the road could cause a fatality."*
P4	YES	Very UnSafe	*"It felt very unsafe, with the car crashing with multiple objects. Again, the car didn't even stop accelerating after having severely crashed and being stuck with one of the obstacles."*
P1	YES	Very UnSafe	*"I Felt more unsafe and the vr view made me more worried about the incident than without it"*

Further, we analyse the relationship between the perceived safety of the participants and the actual outcome of the test case. Figure 5.3a illustrates a box plot which shows the distri-



(a) From the graphs, it seems clear that the user perception of safety resembles the SDC-Alabaster test outcome (pass/-fail)

(b) From the graphs, failure cases, which are the most important to test, are regarded as less realistic than success cases.

Figure 5.3: Graphs visualizing Safety and realism with test outcome from SDC-Alabaster

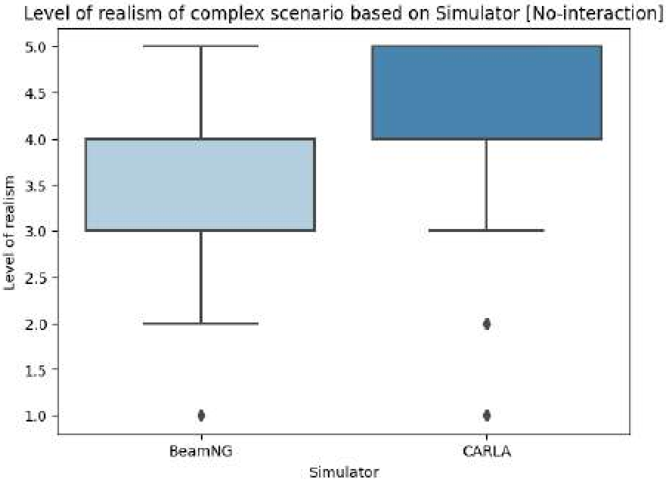
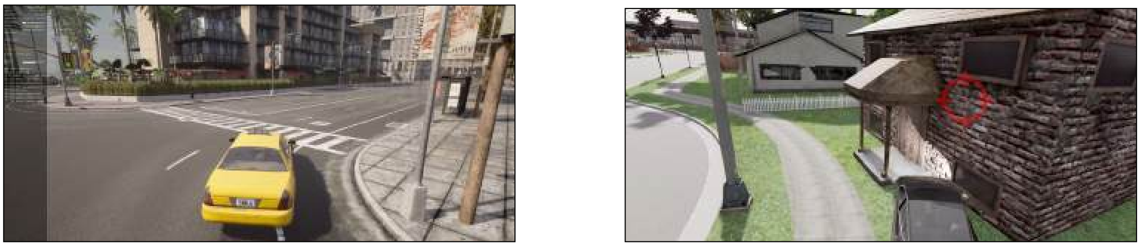


Figure 5.4: From the graphs, it seems clear that the CARLA simulator is perceived as more realistic compared to BeamNG

bution of data according to the results obtained, i.e PASS or FAIL. In the figure, we see that there is a clear difference in distribution between the two categories under consideration. As we expected, in the distribution depicting the FAIL categories, we see that the perceived safety of the users is also low, which directly communicates that whenever the vehicle in the scenario fails to complete the path successfully, the users perceive it to be unsafe. This proves that the human perception and outcome of the scenario are in sync with each other, which confirms the validity of our experiments. The same can be said about distributions that indicate successful scenarios, i.e the users feel safer when the vehicle successfully completes the scenario.



(a) Car is out of line on the turn (P23)

(b) Car crashed to the house (P7)

Figure 5.5: Figure of CARLA simulator from participants' experiment

Finding 2. The prominent finding in this scenario was that participants' perception of safety (safe/unsafe) was inline with the test cases presented by the SDC-Alabaster. This proves that the test case classification (pass/fail) is corresponding to user perception. These results were not statistically significant.

Afterwards, we try to analyse the correspondence between the outcome of the scenario and the level of realism in the scenario. The results obtained can be seen in Figure 5.3b. Interestingly, the level of realism perceived by the user corresponds to the outcome of the test conducted by SDC-Alabaster. This can be seen from the fact that whenever the test cases passed the level of realism experienced by the user is much higher than what he experiences when the test case fails. Although the results seem resounding between the two attributes, no statistical evidence can be attributed to explain this phenomenon due to the lack of available data. However, we plan to conduct further experiments in order to obtain the data which would help us the better understand this finding.

Finding 3. This analysis yielded one of the most important findings of the research under question. We found that in general, the failure cases which are most important to test are also regarded as less realistic by the users in their feedback. which helps future research on SBST and SDC to more focus on how realistic was the failing scenarios. This enables future researchers in SBST and SDCs research to be more focused on how realistic the failing scenarios were to optimize the test selection and minimization.

To analyze the distribution of the perceived levels of realism between the users, we use the Shapiro-Wilk test of normality and verify that the distributions we want to verify are mostly non-gaussian in nature as can be seen in Table 5.4. Due to this observation, we use the unpaired Wilcoxon test, which yields a significance threshold (p-value) of 0.2e-37, indicating that the distribution of the level of realism is statistically significant.

In order to further enhance our understanding of the Wilcoxon test conducted, we compute the effect size of the observed differences. Here we make use of the Vargha-Delaney ($\hat{A}_{12}$) statistic [60]. The Vargha-Delaney ($\hat{A}_{12}$) statistic also classifies the obtained effect size values as in Table 5.1 that are easier to interpret. As a further step in the analyses, we make use of the Vargha-Delaney effect size metric, $\hat{A}_{12}$ has an effect size of 1. This result reveals to us that the effect size is largely significant which shows that the perception of safety among users is dependent on the obstacles in the environment in the sense that users feel safer in scenarios where there are no obstacles compared to the ones with obstacles, as seen in the Table 5.1. Further, we can see feedback from participants on test scenarios with obstacles, as one participant said, quoting *"The scenario was very real along with traffic lights, day and night, foggy and it looked like high quality graphics."*(P7) and another participant felt unsafe with the other bike traffic, quoting *"for the majority of the ride, the drive was safe. It was responding well to the abnormal behaviour of the motorbike applying brakes abruptly. Towards the end, it again hit the sideways at the roundabout. This made it unsafe. "* (P19).

Table 5.4: Statistics for the test scenario WITH or WITHOUT obstacle

Variable	**Factor**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety	WITH	0.0	1.39	4.0	0.01e-13 (non-gausian)	0.64e-8	0.56
	WITHOUT	0.0	2.13	4.0	0.01e-9 (non-gausian)		
Level of realism	BeamNG	0.0	3.75	5.0	0.01e-13 (non-gausian)	0.2e-37	1
	CARLA	0.0	3.75	5.0	0.01e-9 (non-gausian)		

Finding 4. From the experiment, we found that the effect size is due to the difference in the data acquired for scenarios involving obstacles and those not involving obstacles (largely different). The perception of realism among users is significantly dependent on the presence of obstacles in the given scenario.

Finally, we compare the realism of the simulators and compare BeamNG and CARLA in Figure 5.4 and 5.6. This research was conducted as a sort of groundwork to find the major differences between CARLA which is the industrial standard and BeamNG which is the academic standard as we strongly believe this will pave the way for future research. Interestingly, participants felt the CARLA simulator was more realistic than the BeamNG. According to participants, the BeamNG simulator and the environment were not realistic, quoting *"It was not too realistic, but it was also not too shabby. The roads specifically felt very real, but the environment itself did not feel too polished"*(P8). CARLA simulator was perceived as more realistic, quoting "This level was more realistic compared to the previous simulator"(P8).

To analyze the distribution of the perceived levels of realism between the simulators, as seen in Table 5.7, we use the Shapiro-Wilk test of normality to verify that the distributions we want to verify are mostly non-gaussian in nature. which yields a significance threshold (p-value) of 0.1E-16, indicating that the level of realism is statistically significant. Due to this observation, we use the unpaired Wilcoxon test, and as a further step in the analyses, we make use of the Vargha-Delaney effect size metric, $\hat{A}_{12}$ has an effect size of 0.85. This result reveals to us that the effect size is largely significant, which shows that the perception of realism among Simulator has huge differences and users felt CARLA Simulator to be more realistic than BeamNG, as seen from the quote made by the participant, which says *"best than BeamNG t generating realistic environments and scenarios, as well as a better graphical aspect."*(P20). This might be attributed to CARLA having more realistic scenarios that depict a lot of obstacles and area maps that can be found by users in everyday scenarios.

Further to show we analysed the comments from participants and see in Table 5.8. as a participant said CARLA was a city map and it had (pedestrians, cars, other vehicles, traffic signs, etc.) which made participants more realistic to the real world, Quoting "*it was possible to observe almost anything you see in a City (pedestrians, cars, other vehicles, traffic signs, etc.). It was also a way more realistic driving style*"(P1). In the following Table 5.5 and 5.6 we can categorize aspects (or factors that) contribute to high and low safety perception for both CARLA and BeamNG simulators.

Finding 5. The VR view improves the CARLA simulator's safety; vehicles, pedestrians, traffic rules, safety on curves, and the driver's perspective.

Complementarily, VR view and Driver view contribute to the enhancement of BeamNG safety.

Table 5.6: What makes the safety perception low for CARLA and BeamNG

Category "others"	**Description**	**Level of Safety**	**Comments**	**BeamNG**	**CARLA**	**Participants**
Car was very fast	Vehical drow very fast	Unsafe	*"I felt like the start of the self-driving car was very fast compared to a normal driving..."*(P1)	YES	-	P1,P2,P15, P31,P34,P37
		Unsafe	*"The car drove too fast over the speed bumps"*(P2)			
		Unsafe	*"Crashed and also too fast for the bumps."*(P15)			
		Very Unsafe	*"the car was going pretty fast."*(P31)			
		Unsafe	*"exited lane, too fast"*(P34)			
		Unsafe	*"the car went faster and drifted a little bit in the curve"*(P37)			
VR view felt more unsafe	The view of the VR makes the level of perceived safety low.	Safe	*"I Felt more unsafe and the vr view made me more worried about the incident than without it."*(P1)	YES	-	P1
Lane keeping	the car always follows the lane and keep track of traffic rules on lane keeping .	Unsafe	*"The car did abrupt changes in steering and throttle although it kept most of the time the lane."*(P2)	YES	-	P2,P16,P23, P29,P30,P34

		Unsafe	*"the car is not lane keeping properly and even seems to steer off the road at the end.*(P16)			
		Very unsafe	*"Ran off the roads multiple times and did not follow safety lines in curves"*(P23)			
		Very unsafe	*" outside the lines in the curves"*(P29)			
		Very unsafe	*" Drive out of the road and not in the middle of the lane"*(P30)			
		Unsafe	*"exited lane, too fast"*(P34)			
Curves	Steering actions during curves was very unsafe	Unsafe	*"The car cut some curves and it was too fast."*(P1)	YES	YES	P1,P4,P5,P19, P17, P23,P29, P30,P31,P36, P37
		Unsafe	*"the car still went a bit off-road."*(P4)			
		Unsafe	*"driving too fast during the curves"*(P5)			
		unsafe	*"turning was not good. the first right turn was very off. .*(P19)			
		Unsafe	*"Cannot term it safe as car drove off the road"*(P17)			
		Very unsafe	*"Ran off the roads multiple times and did not follow safety lines in curves"*(P23)			
		Unsafe	*"The car drove outside the lines and was too fast in the curves "*(P29)			
		Very unsafe	*" Drive out of the road and not in the middle of the lane"*(P30)			
		Very unsafe	*" When the car starts to go off the road when driving in a curve it feels pretty unsafe. "*(P31)			
		Unsafe	*"car went out of the road partially on curves"*(P36)			
		Unsafe	*"the car went faster and drifted a little bit in the curve"*(P37)			
Crossed a STOP sign	Car crossed crossed a STOP sign	Neutral	*"the car crossed a STOP sign without stopping.."*(P2)	-	YES	P2,P5,P14, P15, P16,P1, P24,P26,P29, P31
		Unsafe	*"it doesnt recognized the stop signal"*(P5)			
		Neutral	*"Safe- maintains a speed limit and traffic limitations but fails to stop at the stop signs. This makes it unsafe"*(P14)			
		Unsafe	*"Goes towards the edge of the road"*(P15)			
		Unsafe	*"the car was not so smooth at the turns and the speed was high at the turns. "*(P16)			

		Unsafe	*"the car didnt stop on STOP sign and red light "*(P1)			
		Unsafe	*"didnt stop at a stop sign, otherwise pretty safe "*(P24)			
		Unsafe	*"skip a stop sign "*(P26)			
		Unsafe	*"Car didn't stop at stops "*(P29)			
		Unsafe	*"car ignores stop signs"*(P31)			
Car abrupt frequently	Car stops frequently.	Safe	*"safe but instead of going slow when a car is slow it accelerates and stops too many times."*(P20,29)	-	YES	P20
		Very unsafe	*"brakes were very abrupt"*(P29)			

Finding 6. The factor limiting CARLA's safety is that the VR view felt more dangerous; CARLA frequently crossed a STOP sign and stopped abruptly. Complementarily, the reason for limiting the safety of BeamNG is that the car was traveling at a high rate of speed, and the VR view felt more dangerous.

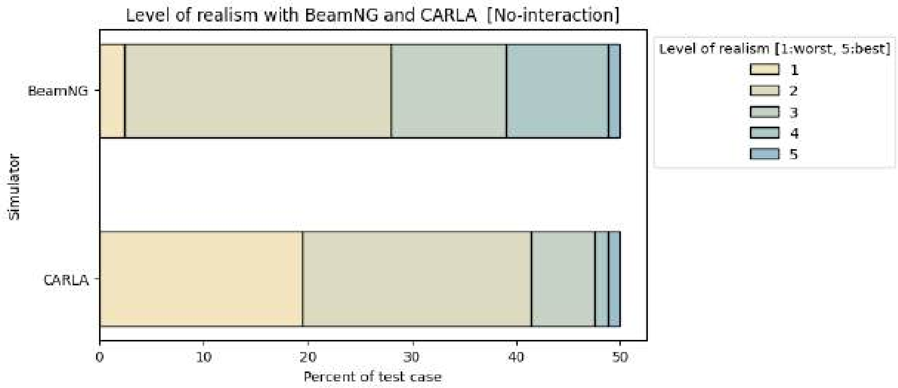


Figure 5.6: From the graphs, It shows CARLA slightly more realistic than BeamNG

Further, we compare how participants perceive safety in both the CARLA and BeamNG simulators. In contrast to the results observed in Figure 5.7 it's clear that participants felt safer in the CARLA simulation than in the BeamNG. This was an expected result, as we saw in the previous result, where CARLA was considered more realistic than BeamNG Simulator, so it was obvious that participants felt safer because the scenario was more realistic. As per participants said, they felt CARLA was better than BeamNG environments and scenarios, quoting "best than BeamNG generate realistic environments and scenarios, as well as a better graphic aspect" (P20). Another P35 said the CARLA test case was a city, which helped them

Table 5.5: What makes the safety perception high for CARLA and BeamNG?

Category	Description	Level of Safety	Comments	BeamNG	CARLA	Participants
VR view	The view of the VR makes the level of perceived safety high.	Safe Very unsafe	*"i felt safer with the vr view"*(P1) *" I felt as unsafe as without VR glasses."*(P8)	YES	YES	P1,P8
Vehicles pedestrian and traffic rules	Other vehicles,pedestrian and and traffic rules improved the safety perception	Very Safe Safe Very safe safe safe	*" impressive considering that now there were other cars, obstacles and people in the scene.."*(P1) *" it stoped when the traffic light was red. "*(P5) *"following the traffic light and very safe. "*(P6) *"then during the whole trip it respected the lights, the speed and the road limits"*(P22) *"car was following rules and drived carefully"*(P39)	-	YES	P1,P5,P6, P22,P39
Safety on curves	Steering actions during curves was very safe.	Safe Safe Neutral Very safe Very safe	*"I Felt more unsafe and the vr view made me more worried about the incident than without it."*(P1) *"The car followed traffic signal and was almost all the time within the line."*(P7) *"the car followed the traffic signals speed limit and lanes"*(P11) *"The car seems to be able to lane keep almost perfectly... "*(P17) *"The car follows the traffic rules and does a good job keeping track of vehicles and pedestrians. "*(P18)	-	YES	P1,P7,P11, P17,P18
Drivers view	drivers view make more safer.	Very Safe safe safe	*"From inside the car, it felt very safe and smooth. ."*(P16) *"From inside feels more safe, since I can check speed its going and we dont really see thats its going ouside the road"*(P2) *"Being inside the car gives a full perspetive of the scenario and it felt very safe"*(P5)	YES	YES	P6,P2,P5

feel more realistic, quoting, "it was more realistic than BeamNG since it was an actual city. "The car was also driving smoother, which helped with the realism." (P35).

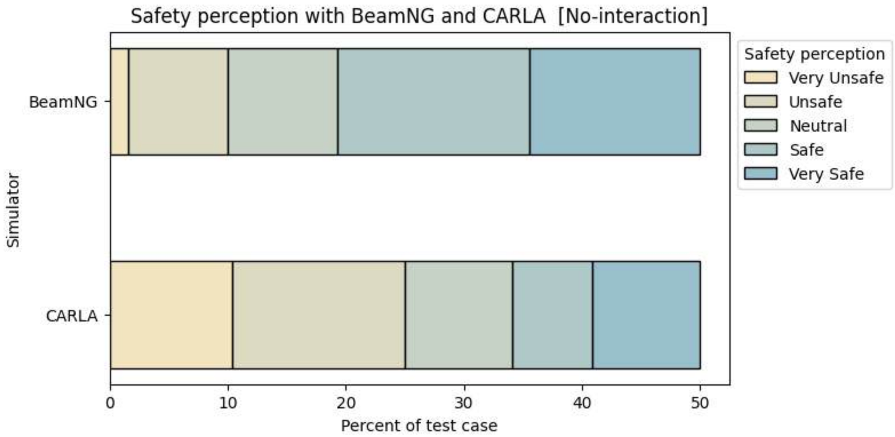


Figure 5.7: From the graphs, it seems clear that the safety perception CARLA simulator is higher than simulator BeamNG

Table 5.7: Statistics for the test scenario of CARLA and BeamNG simulator

Variable	**Factor**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety	BeamNG CARLA	0.0 0.0	1.32 2.20	4.0 4.0	0.01e-11 (non-gausian) 0.03e-11 (non-gausian)	0.05e-10	0.68
Level of realism	BeamNG CARLA	0.0 0.0	3.36 4.14	5.0 5.0	0.01e-13 (non-gausian) 0.01e-15 (non-gausian)	0.1e-16	0.85

To analyze the distribution of the perceived levels of safety between the simulators we use the Shapiro-Wilk test of normality as seen in Table 5.7 and verify that the distributions we want to verify are mostly non-gaussian in nature. which yields a significance threshold (p-value) of 0.05e-10, indicating that level of safety is statistically significantly higher in CARLA is than BeamNG. Due to this observation, we use the unpaired Wilcoxon test, we make use of the Vargha-Delaney effect size metric, $\hat{A}_{12}$ has an effect size of 0.68. This result reveals to us that the effect size is largely significant, which shows that the perception of safety among Simulator is high and users feel safer in BeamNG scenarios compared to CARLA test scenarios, as seen in the Table 5.1.

Finding 7. According to the results of the experiment, The CARLA simulator was more realistic than the CARLA simulator because the participants found the CARLA scenarios more realistic. These results were statistically significant with p-value> 0.1e-16.

When we check the results between complex scenarios (scenarios with obstacles such as bumps, other cars, and pedestrians) and non-complex (without any obstacle). As we expected we can see in Figure 5.9 participants felt safer in simple test scenarios compared to a complex

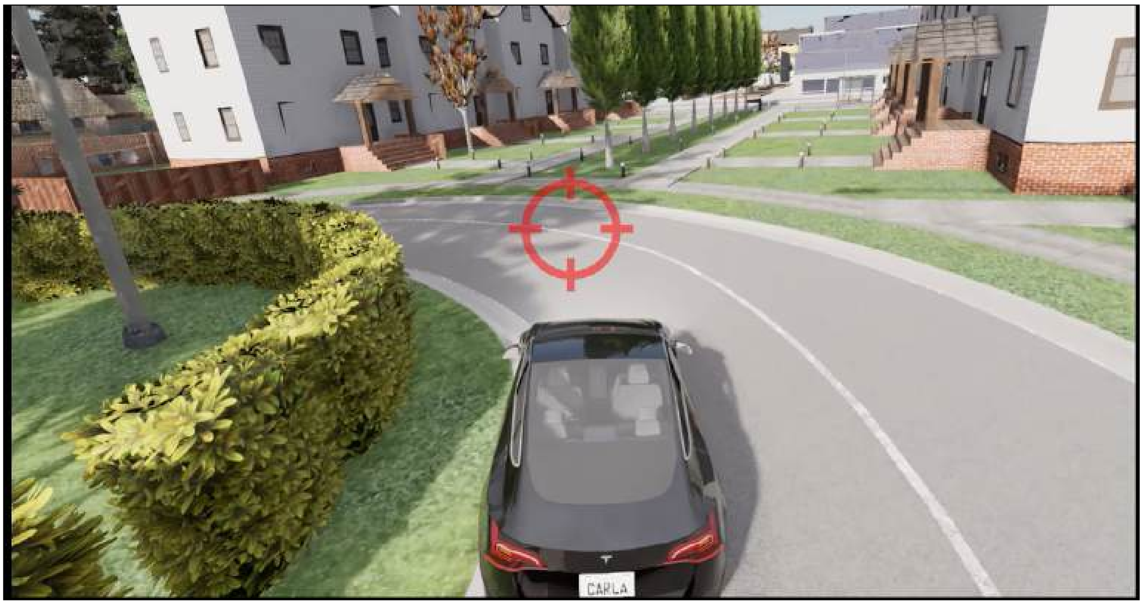


Figure 5.8: Participants' P17 outside view car crash to the roundabout.

scenario. Which has been proved by analysing the distribution of the perceived levels of safety between complex and simple scenarios. We use the Shapiro-Wilk test of normality and verify that the distributions we want to verify are mostly non-gaussian in nature as seen in Table 5.4, which yields a significance threshold (p-value) of 0.64e-8, indicating that level of safety is statistically significantly higher in complex scenarios than simple scenarios. Due to this observation, we use the unpaired Wilcoxon test, we make use of the Vargha-Delaney effect size metric, $\hat{A}_{12}$ has an effect size of 0.56. This result reveals to us that the effect size is largely significant, which shows the perception of feeling more unsafe with the complex scenario, as seen in Table 5.1.

Finding 8. Analyzing the results of the experiment, we can see that the users' perception of the complexity of the scenario also affects how safe they feel in that particular scenario. From the tests, we found that the users felt less safe when the environment grew complex.

Overall, 5.3b and 5.3a clearly show that humans' perception of SDCs' test failures/safety resembles that of humans. Therefore, it provides evidence that the SDC-Alabaster classification of the test case scenarios closely resembles human perception, which answers one part of RQ$_3$ is *"How closely does the SDC test case resemble a real-world driving event?"*. The other part of RQ$_3$ is *"Does the Test case resemble a real-world driving event?"* When we observe Figure 5.4 we can observe that CARLA is more realistic compared to BeamNG but still both simulators do not show their close resemblance. In general, we see in Table 5.9 we observe the feedback from participants on the simulators, as one participant said *"the road was not*

Table 5.8: Some qualitative comments on the CARLA simulator's Quality

Participants	**Simulator**	**Simulator**
P1	CARLA	*"It was possible to observe almost anything you see in a city (pedestrian, cars, other vehicles, traffic signs, etc.). It was also way more realistic driving style."*
P5	CARLA	*"Very good actually. I don't put a 5 because there is always room for improvement and I've seen game engines with more realistic results, but I was positively surprised. While the car was designed as a single box, the landscape was much more realistic, which made you more immerse in the scenario. Also the fact that it uses full VR (3D) makes a big difference."*
P20	CARLA	*"best than BeamNG t generate realistic environments and scenarios, as well as a better graphical aspect."*
P35	CARLA	*"It was more realistic than BeamNG since it was an actual city. The car was also driving smoother which helped for the realism."*

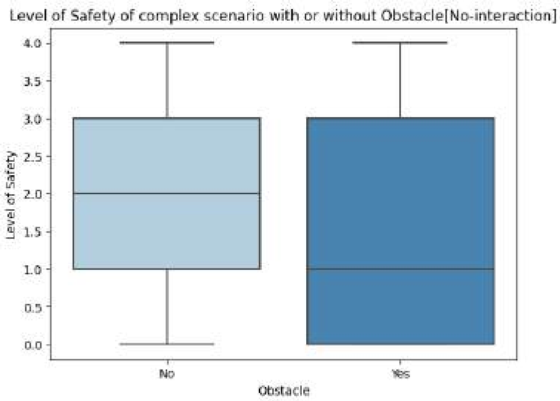


Figure 5.9: From the graphs, it seems clear that when test scenarios are complex safety perception is low compare with simple scenarios(Without obstacles)

realistic in some cases, like a very small side road in warmup scenario."(P36) for BeamNG simulator, *" The whole simulation is haltingly and felt very artificial."* (P2) for CARLA, which is surprising given that simulators like CARLA and BeamNG are used widely used in the industry and academia. Considering the state of the art for testing SDCs testing but still we see the participants felt less realistic. Simulator BeamNG used in SBST [1] and CARLA is an industrial standard simulator that is used in various domains of vehicles [19]. But when we look at the statistical analysis Table 5.5 and the Figure 5.6 we can see no simulator has any high perception of level realism. This could be because of limited intelligence of BeamNG toward obstacles and the speed of the car. In CARLA, there were dynamic obstacles that were

[1] `https://sbst21.github.io/`

too close, and the driving was poor.

Table 5.9: Some qualitative negative comments on the simulator's realism

Participants	**Simulator**	**Simulator**
P25	BeamNG	*"The scenario felt ok, especially the one with the obstacles and the bumps of the car on the obstacles on the road. Same for the quality of the perception in the simulation inside the car. It would be nice to add the sound to the scenario to increase realism."*
P28	BeamNG	*"It looks like a computer game to me and I can differentiate clearly between reality and fiction as far as I know."*
P36	BeamNG	*"the road was not realisitic in some cases, like a very small side road in warmup scenario."*
P40	BeamNG	*"Low graphics from the environment, no other cars or people around, car feels quite regid and going over the obsticles is not bumpy."*
P2	CARLA	*"The whole simulation is haltingly and felt very artificial."*
P33	CARLA	*"The graphics were not clear sometime, also the car drive in a nonuniform way which makes sometime the dizzy experience."*
P10	CARLA	*"I do not feel self driving cars safe especially with dynamic obstacles and changing driving conditions. It is very difficult for self driving car to tackle with real world drivers as they can change their decision."*

5.2 RQ$_4$: What is the human perception of SDC's test failures/safety when humans can interact with the car?

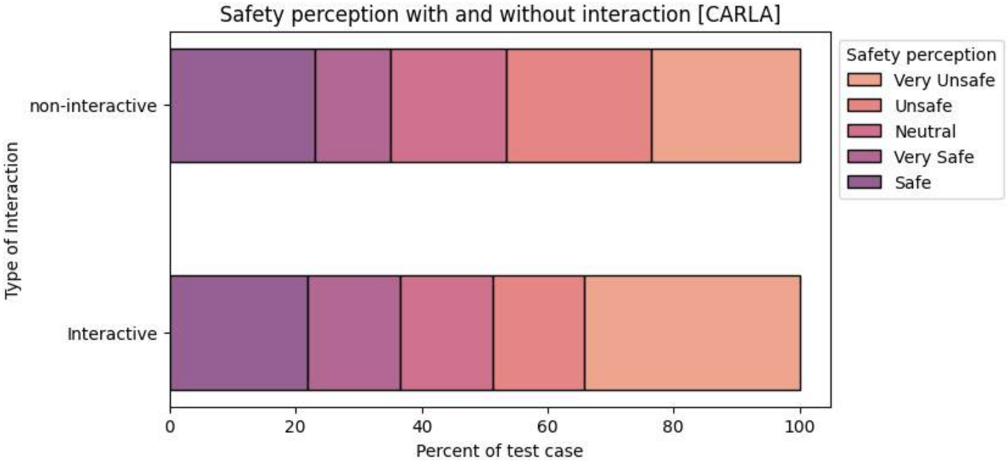


Figure 5.10: From the graphs, the perception increases when participants can interact with the car.

As mentioned in section 4.5.2 we made use of proper visualizations to get a better understanding of the results. When we see the Figure 5.10 clearly shows that the proportion of test cases with varying safety perceptions, with and without interaction has a huge difference. when we check further with participants' feedback we can see that participants felt safer when had some interaction, quoting *"it is safer when controlling safety zones"*(P20), *"Having the control is better for safe perception"* (P26). When we compare the same test case with similar attributes (obstacles, traffic, etc.) and compare the result we see participant P35 felt safer when he/she could control the care. Without interaction *"No slowing down for the intersection and later crashed in the roundabout"* (P35) and with interaction, *"I was able to slow down the car in front of the intersection and front of the roundabout"* (P35), *"Could have a higher level of control over the interactions* "(P12). Participant feedback included how we can utilize interaction until the AI-agent model is fully trained quoting *"The interactive experiments made me see that a combination of self-driving AI with human control could be a way to solve the initial phases of self-driving cars, until cars are properly trained."*(P5). This could be useful not only for testing autonomous vehicles but also for boosting user confidence.

We did further analyze the distribution of the perceived levels of safety with and without interaction as you can see in Figure 5.11, We use the Shapiro-Wilk test of normality and verify that the distributions we want to verify are mostly non-gaussian in nature as seen in Table 5.10, which yields a significance threshold (p-value) of 0.001, indicating that level of safety is statistically significant. Due to this observation, we use the unpaired Wilcoxon test, we make use of the Vargha-Delaney effect size metric, $\hat{A}_{12}$ has an effect size of 0.36. This result reveals to us that the effect size is medium significant, which shows that the perception of safety

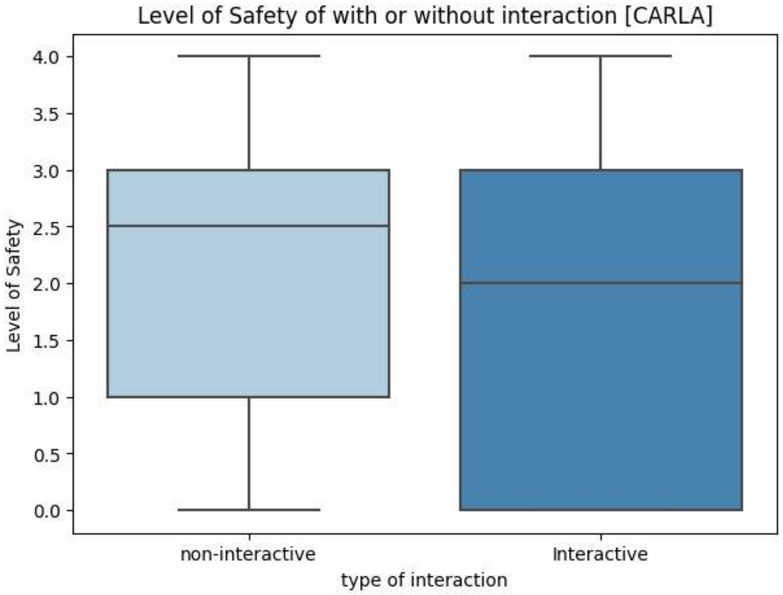


Figure 5.11: From the graphs, participants can interact with the car, their perception improves.

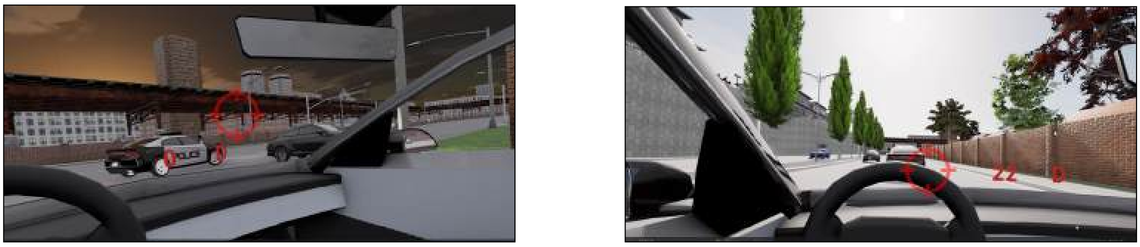
Table 5.10: Statistics for the test scenario WITH or WITHOUT interaction

Variable	**Factor**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety	WITH WITHOUT	0.0 0.0	1.68 2.20	4.0 4.0	0.06e-10 (non-gausian) 0.03e-11 (non-gausian)	0.001	0.36

when the participant can interact with the car is statically higher than without interaction, as seen in Table 5.1.

Finding 9. Our experiment with the participants being able to interact with the cars provided evidence that with a slight interaction, humans felt more safe compared to the results without any interaction. This demonstrates the importance of having a human in the loop of testing self-driving cars. These results were statistically significant with a p-value > 0.05e-10.

Unexpectedly the slight interaction with the car improved the safety perception of participants significantly. This is a novel finding that we have in our use case scenario. We usually come across experiments where that do not take into account the human interaction with the vehicle [1]. However, according to our findings, we observe that human interaction plays a significant part in the safety perception of the users. We believe that this finding would play a prominent role in future research as it allows experiments to include user interaction with the



(a) Participants' control when there were accidents (P4)

(b) Participants' control with interaction (P23)

Figure 5.12: Figure of CARLA simulator with interaction from participants' experiment

Table 5.11: Qualitative comments on the CARLA simulator's safety with or without interaction

Interaction	Participants Code	Simulator
NO	P1	*"similar level of unsafety. main difference was that the not so smooth behavior of the car happens also in the proximity of other cars (fast restarts followed by rapid stops, a more safe driving would be ideal). However, without obstacles i felt in some cases the car was too slow compared to the one of BeamNG "*
	P7	*"It was very unnsafe. The vehicle was not able to take turn on the roundabout and there was accident."*
	P35	*"No slowing down for the intersection and later crashed in the roundabout"*
	P41	*"started okay, but the unsteady behaviour of other vehicles made me feel uncomfortable, late stopping of ego vehicle also, very unsafe behaviour around roundabout"*
YES	P1	*"the fact i could control the car when needed, gave me a safer perception of the driving experience. Moreover, i could speed up the car when i wanted to."*
	P4	*"With a bit of control it feels safer, especially being able to adjust the speed in dangerous situations. However, it is still not safe since the car ends up going off road at the end of the scenario."*
	P20	*"it is safer when controlling safety zones"*
	P26	*"Having the controll is better for safe perception"*
	P35	*"I was able to slow down the car in front of the intersection and in front of the roundabout"*
	P1	*"with interaction, I got the feeling that was slightly more easy to feel more safe, still with unpredictable behaviors of the car, the perception of safety was very low."*
	P2	*"With interaction I felt safer since I had still the final control to stop the car although I could not steer it."*
	P4	*"With interaction it feels much safer, since you can control speed in case of emergency."*

vehicles at centre of the testing automation. This is an unexplored dimension of simulating autonomous cyber-physical systems in which we believe a lot of more diverse studies need to be conducted.

Table 5.12: What makes the interaction test case for the CARLA simulator safer?

Category "others"	Description	Comments	Participants
Unpredictable behaviors	Car was still unpredictable behaviors with interaction	*"I got the feeling that was slightly more easy to feel more safe, still with unpredictable behaviors of the car, the perception of safety was very low."*(P1)	P1

Late reaction	Car responded very slowly to the participant's action.	*"However, the car didnt react to my orders as I would have expected at the moment I wanted to avoid the accident "*(P22) *"In the end the perception of the safety was quite comparable with and without the interaction, maybe because on of the simualation with the interaction ended up anyway with an accident but it was nice to have the possibility to interact with the simulation."*(P25) *"but after I did not manage to stop the common crash, I felt more insecure. I would still prefer to have a say and give feedback to control the speed to at least reduce the intensity of the crashes."*(P40)	P22,P25,P40
interaction has no effect on the participants	Didn't find any impact with interaction	*"I did not feel a large difference between the two as I did not feel like I had a big impact on the cars behavior"*(P3) *"not very effective while interacting"*(P6) *"doesnt make a difference as the responsiveness was slow and also couldnt perceive much change in the way the ai functioned"*(P13) *"It would be the same. The car was slowing down automatically without me pressing the unsafe key during the interactive session."*(P14) *"not of much use I would say. At slow speeds within the city it is still useful."*(P19) *"I did not feel any difference."*(P2) *"With interaciton is more safety. "*(P23) *"I did not feel like it made that much of a difference with or without interaction."*(P27) *"no difference basically"*(P28) *"Does not change much, since there is no instant feedback. Since one does not have direct control over the brakes"*(P31)	P3,P6,P13,P14,P19 P2,P23,P27,P28,P31
Avoid accident	Felt safer so they can avoid accidents	*"It help that i have a little control but still could do much when the accident happend so maybe even better."*(P21) *"With interaction, I could control speed and avoid accidents. So I would prefer that."*(P32)	P21,P32
Stop car	Felt safer so they can stop car	*"With interaction I felt safer since I had still the final control to stop the car although I could not steer it."*(P2) *"I think it feels much safer with interactions because I know I can intervene if something unexpected happens. "*(P37)	P2,P37
Slow down	Felt safer so they can slow down at critical moments	*with interaction it feels much safer, as I could tell the car to slow down in critical moments"*(P24) *"With interaction I was able to slow down the car beofre the intersection and the roundabout which helped alot to feel safer."*(P35)	P24,P35

Perceived safer	participatns felt more safe with interaction	*"Safety with Interaction made me feel I had a bit more control of the situation compared to without interaction "*(P22) *"Much safer with interaction"*(P29) *"I feel much safer when I can interact."*(P30) *"With interaction, it was a better experience."*(P33) *"A bit better, and the tool gets direct input which most probably will be helpful"*(P34) *"with the controll, i felt more safe, even though i could not do exactly what i wanted to do with the controll "*(P36) *"It is a little bit saver to have the possibility to interact, but that is the drivers habit"*(P38) *"Safety with interaction was better"*(P39) *"I was happier and feeling safer that I can control through my feedback at the begining"*(P40)	P22,P29,P30,P33,P34,P35,P36,P37,P38,P39,P40

Chapter 6

Discussion

This section discusses additional factors that might impact the results of the research questions and provides more insights and insights surrounding them.

In addition, it includes an overview of summaries of feedback and lessons learned from the participant on the conducted study.

6.1 Safety perception and realism of test cases with various participants' factors (RQ_3)

As we have seen in Section 5 we observed that the participants' perception of safety and level of realism varied with factors like WITH or WITHOUT VR, the difference between the simulators, and compared the results between the complexity of test scenarios and with interaction. Here, we will examine how the participant's background influences the outcomes.

6.1.1 Does the participants' age affect their perception?

Figure 6.1 depicts the stacked histogram of the proportion of test cases of different ages. We observed that the participants' perception of safety was slightly higher in people in the age group 18–30 years compared to the people with an age above > 30. We can conclude that age doesn't influence a lot on the level of safety, and for further investigation, we analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.1. The p-value threshold in this approach was set to 0.05 (as a rule of thumb) which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.1 the "p-value" parameter has a value of 0.35, which reveals that there is no statistical evidence that there is a difference between the distribution.

Further, we analyze the level of realism perceived by the corresponding participants' age, as you can see in Figure 6.2a. Surprisingly, the level of realism perceived by participants aged 18 to 30 is significantly higher than that of those aged 30 and older. We speculate that younger generations (such as Generation Z and Millennials[1]) are very accustomed to the new graphic user interface, VRs, and digital media, which makes them more familiar with the technology than older age groups, which have higher expectations. also, because younger generations

[1] http://tony-silva.com/eslefl/miscstudent/downloadpagearticles/defgenerations-pew.pdf

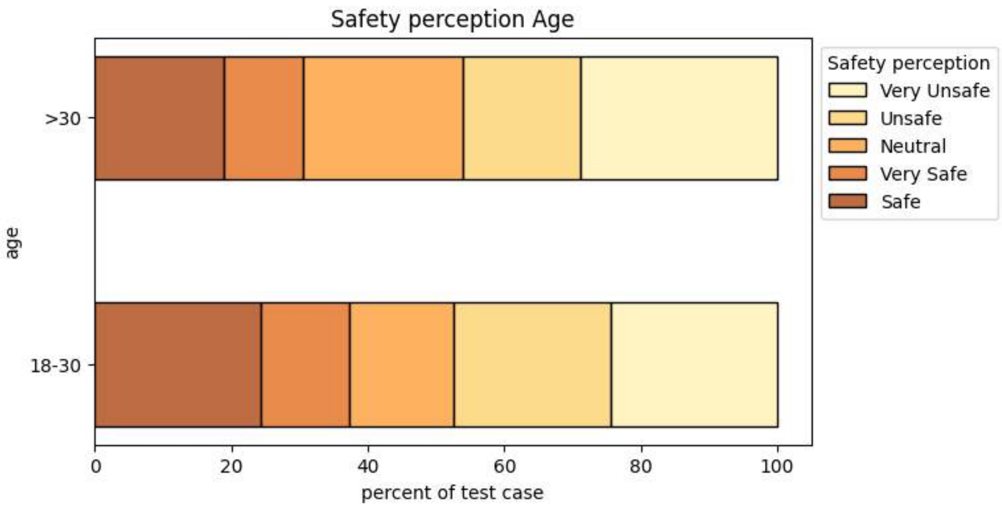


Figure 6.1: From the graphs, it is evident that participant age has no effect on safety perception, as the distributions are almost identical.

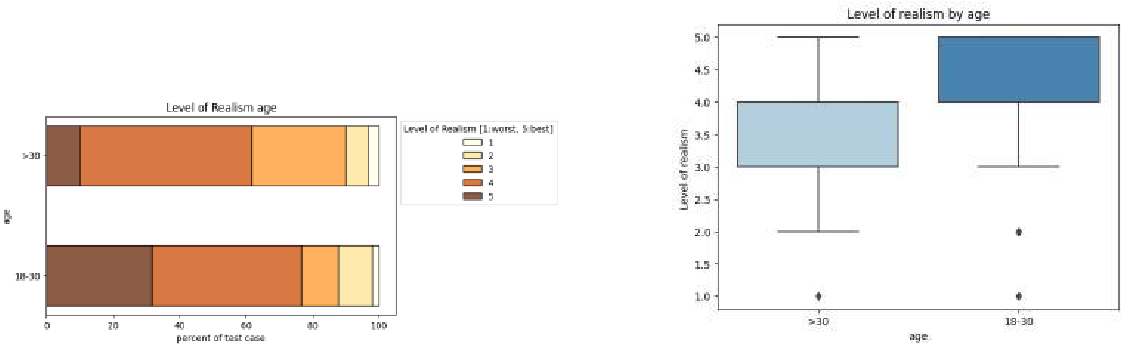
Table 6.1: Statistics for the test scenario based of different age group.

Variable	Age	Min	Mean	Max	Distribution	p-value	$\hat{A}_{12}$
Level of safety	>30 18-30	0.0 0.0	1.67 1.78	4.0 4.0	0.08e-9 (non-gausian) 0.09e-16 (non-gausian)	0.35	-
Level of realism	>30 18-30	0.0 0.0	3.58 3.93	5.0 5.0	0.07e-11 (non-gausian) 0.01e-18 (non-gausian)	0.01e-04	0.37

are more likely to be exposed to newer forms of media at an earlier age in life. In addition, as seen in prior research [16] young drivers' attitudes toward the most important road safety measures and their perceptions of their effectiveness are very low.

To analyze the distribution of the perceived levels of realism between the participants' age groups, we use the Shapiro-Wilk test of normality and verify that the distributions we want to verify are mostly non-gaussian in nature, as can be seen in Table 6.1. Due to this observation, as we can see in Figure 6.2b, which yields a significance threshold (p-value) of 0.1e-1, indicating that level of realism is statistically significant, As a further step in the analyses, we use the unpaired Wilcoxon test and make use of the Vargha-Delaney effect size metric, $\hat{A}_{12}$ has an effect size of 0.37. This result reveals to us that the effect size is moderately significant, as seen in Table 5.1. which proves that participants in the age group 18–30 have a significantly higher perception of realism than participants in the age group 30 and older.

Finding 10. According to the results of the experiment, age plays an important role in how realistic the simulators are. Participants in the age group 18–30 felt simulators were more realistic than those who were older than 30 years. We hypothesize that



(a) The graphs make it very clear that one of the most important factors in determining how realistic test scenarios are is the participant's age.

(b) From the graphs, it seems clear that different age groups have impacts on the level of realism of the test scenarios.

Figure 6.2: Graphs visualizing level of realism based on age of the participants'

younger generations are more habituated to modern graphical user interfaces, virtual reality, and digital media, making them more comfortable with technology than older age groups, who have greater expectations.

6.1.2 Does the participants' gender play any role in safety perception and realism in SDC test cases?

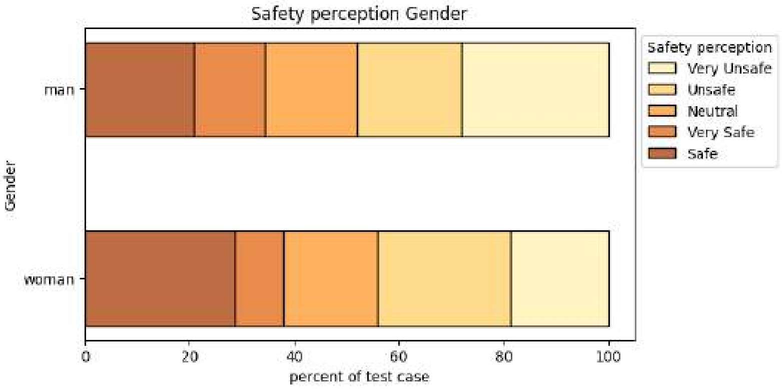


Figure 6.3: From the graphs, the Gender has an effect on the safety perception.

Further, we analyze if the gender of the participants has any effect on the perceived safety and realistic nature of the test scenarios. Figure 6.3 depicts the stacked histogram of the proportion of test cases of different Gender (Male/Female). When we compared, we found that female participants had a more positive perception of safety than male participants. As seen in previous research [42], men are more accustomed to injuries and accidents than women, so male participants felt more insecure. We analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.2. The p-value threshold in this approach was set to 0.05 (as a rule of thumb) which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.2 the "p-value" parameter has a value of 0.27, which reveals that there is no statistical evidence that there is a difference between the distributions.

Table 6.2: Statistics for the test scenario based of different gender.

Variable	**gender**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety	Man Women	0.0 0.0	1.72 1.84	4.0 4.0	0.07e-18 (non-gausian) 0.06e-7 (non-gausian)	0.27	-
Level of realism	Man women	0.0 0.0	3.77 4.02	5.0 5.0	0.07e-11 (non-gausian) 0.01e-18 (non-gausian)	0.02	0.26

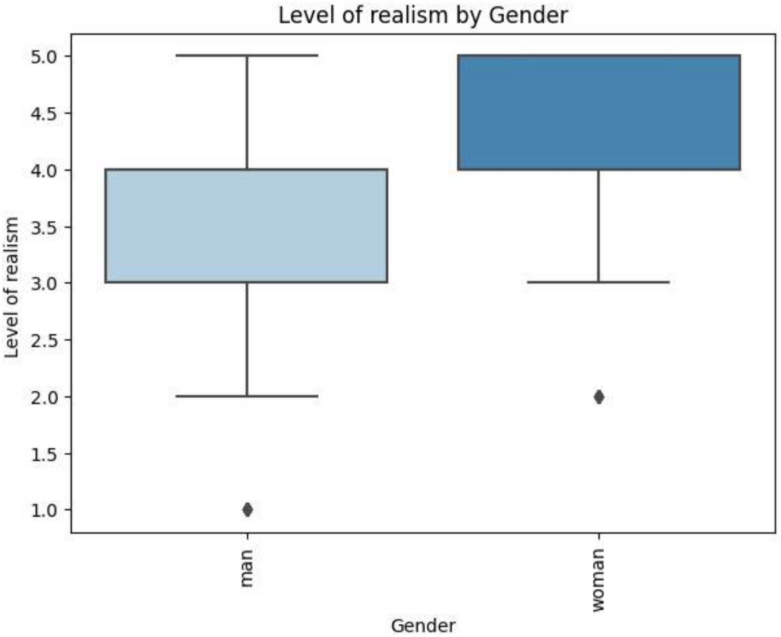


Figure 6.4: The graphs make it very clear that one of the women felt test scenarios were more realistic than men

Further, we analyze the level of realism perceived by the corresponding participant's gender, as you can see in Figure 6.5. Astonishingly, the level of realism perceived by the female participants was significantly higher than males. As we see the feedback from a female partic-

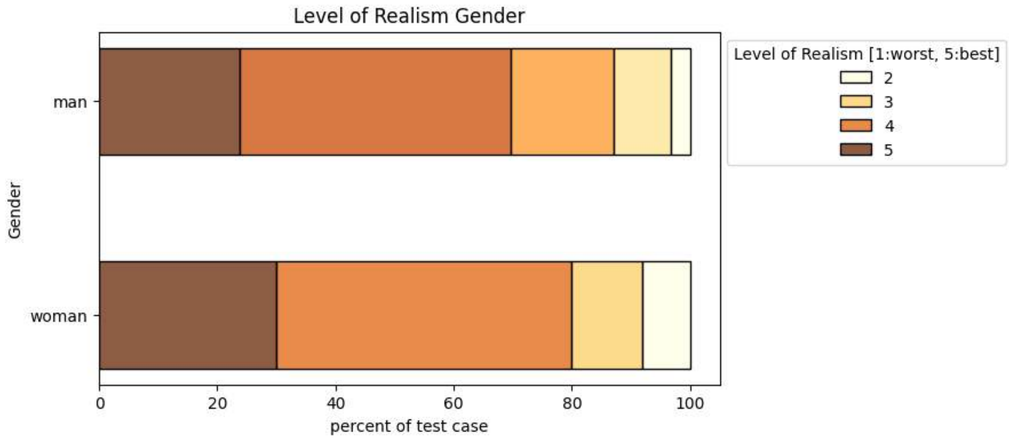


Figure 6.5: From the graphs, it seems clear that gender has different impacts on the level of realism of the test scenarios.

ipant quoting *"The scenario it quite realistic probably also because it considers the simulation inside a city in which the car has to deal with traffic signs and other cars, but also people."* and when we compare with male participant feedback *"The movements were very abrupt, the physics did not feel too realistic"* which shows with female participants took the test scenarios more realistic than males. This would cause a higher level of expectation in male participants compared to female participants. Also, as we can observe in Figure 6.6, men played PC games more than women, which would make the perceived scenarios more realistic. We will compare the effects of playing PC games on perceived safety and level of realism in section 6.1.5.

To analyze the distribution of the perceived levels of realism between the participants' genders, we use the Shapiro-Wilk test of normality to verify that the distributions we want to verify are mostly non-gaussian in nature, as can be seen in Table 6.2. Due to this observation, as we can see in Figure 6.4, which yields a significance threshold (p-value) < 0.02, indicating that the level of realism is statistically significant. As a further step in the analyses, we use the unpaired Wilcoxon test and use of the Vargha-Delaney effect size metric, $\hat{A}_{12}$ has an effect size of 0.26. This result reveals to us that the effect size is small and significant, as seen in Table 5.1 because of the unequal distribution of participants (men: 31, female: 10). But still, it proves that participants of the female gender have a significantly higher perception of realism than participants of the male gender.

6.1.3 Does the participants' field of expertise or the studies (IT v.s non-IT experts) they've done affect the level of perceived realism and safety of test scenarios?

Next, analyze if the expertise of the participants (IT versus non-IT experts) has any effect on the perceived safety and realistic nature of the test scenarios. Figure 6.7 depicts the stacked histogram of the proportion of test cases for a different group of participants who study or work in computer science and participants from another background (i.e business

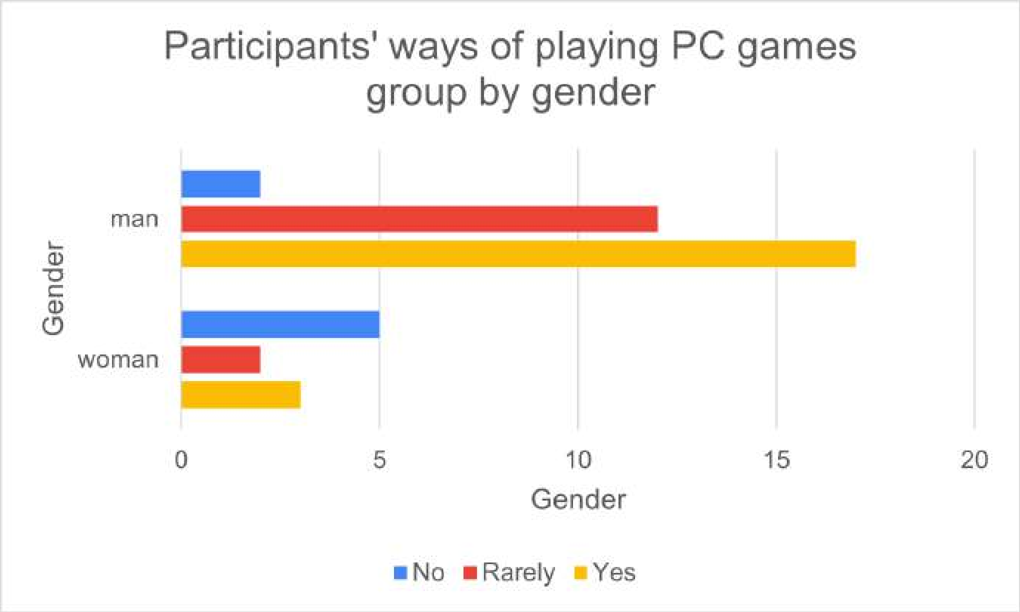


Figure 6.6: This graph clearly shows that male participants played more PC games compared to female participants.

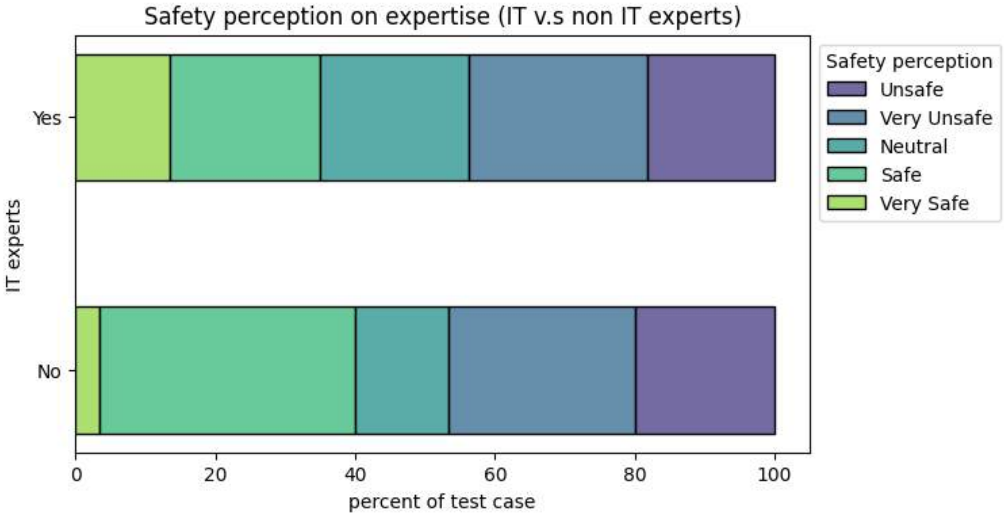


Figure 6.7: From the graphs, it is evident that non-IT participant safety perception higher than IT experts.

administration, banking and finance, economics, AI ethics, political science, and biology). When comparing the IT expert participants, we found that non-IT expert participants had a more positive perception of safety than IT expert participants. This was to be expected given that IT experts have a greater amount of experience with graphical user interfaces and simulators than those who are not experts in IT. We analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.3. The p-value threshold in this approach was set to 0.05 (as a rule of thumb) which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.2 the "p-value" parameter has a value of 0.81, which reveals that there is no statistical evidence that there is a difference between the distributions.

Table 6.3: Statistics for the test scenario based on IT versus non-IT experts

Variable	**IT Experts**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety	Yes No	0.0 0.0	1.75 1.7	4.0 4.0	0.04e-18 (non-gausian) 0.01e-21 (non-gausian)	0.81	-
Level of realism	Yes No	0.0 0.0	3.80 4.1	5.0 5.0	0.07e-11 (non-gausian) 0.01e-8 (non-gausian)	0.02	0.26

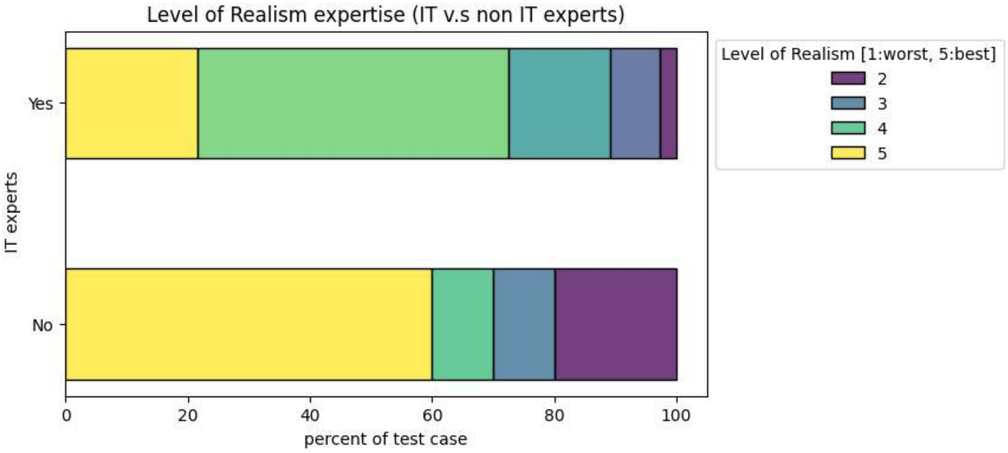


Figure 6.8: The graphs make it very clear that non-IT experts felt test scenarios were more realistic than IT experts

As you can see in Figure 6.8. Surprisingly, the level of realism perceived by the non-IT expert participants was significantly higher than that of the IT experts. We speculate that IT experts are more used to virtual environments compared to non-IT experts. This would cause a higher level of expectation among IT expert participants compared to non-IT expert participants. As we can see from the feedback of P12, who is from *Business administration* background quoting *"Was not very realistic; Poor graphics and over unrealistic* "(P12) This demonstrates that non-IT participants had a unique level of perception, prompting researchers to conduct experiments with participants from diverse backgrounds in the future.

To analyze the distribution of the perceived levels of realism between the participants' IT versus non-IT experts, we use the Shapiro-Wilk test of normality to verify that the distributions we want to verify are mostly non-gaussian in nature, as can be seen in Table 6.3. Due to this

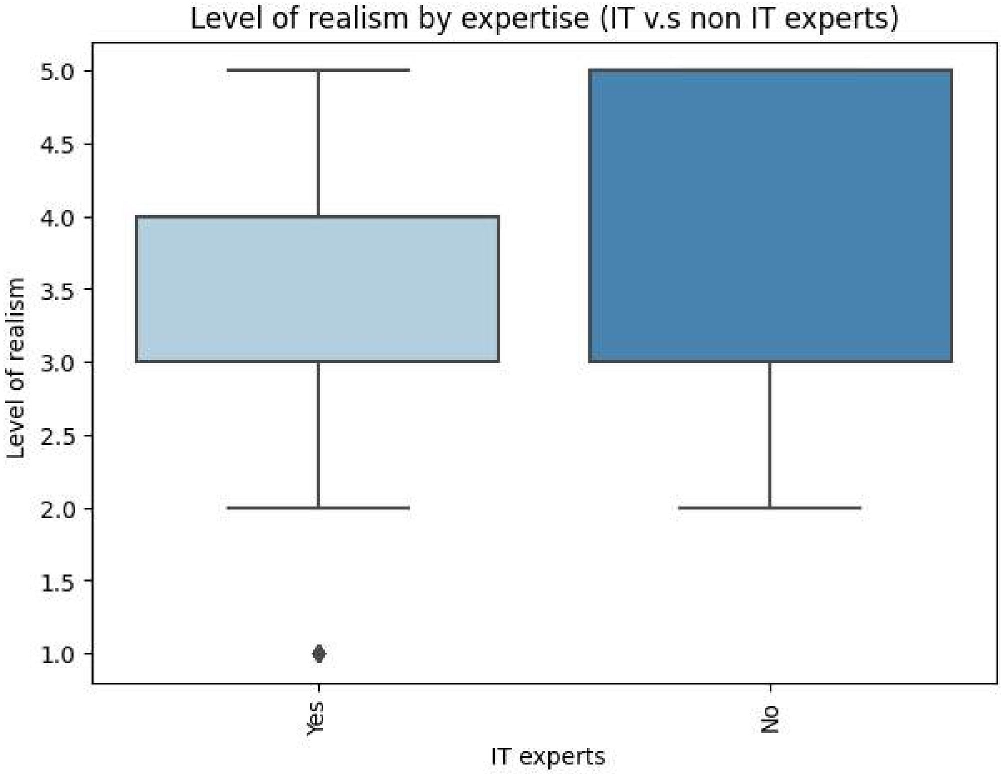


Figure 6.9: From the graphs, it seems clear that IT versus non-IT experts has major impacts on the level of realism of the test scenarios.

observation, as we can see in Figure 6.9, which yields a significance threshold (p-value) < 0.05, indicating that the level of realism is statistically significant. As a further step in the analyses, we use the unpaired Wilcoxon test and make use of the Vargha-Delaney effect size metric, $\hat{A}_{12}$ has an effect size of 0.26. This result reveals to us that the effect size is small and significant, as seen in Table 5.1 because of the unequal distribution of participants on the dataset(IT: 30, non-IT: 11). But still, it proves that participants of the non-IT experts have a significantly higher perception of realism than participants of the IT experts.

6.1.4 Does the prior experience with VR participants impacting their perceived level of safety and realism?

When we analyze the participants' who already used VR, do they have any effect on the perceived safety and realistic nature of the test scenarios? Figure 6.10 depicts the stacked histogram of the proportion of test cases for the level of safety of the participants who already used VR. When comparing the results, we found that participants who already used VR and who didn't use VR same similar safety perceptions. We analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.4. The p-value threshold in this approach was set to 0.05 (as a rule of thumb) which indicates that if the p-values obtained during the test were less than

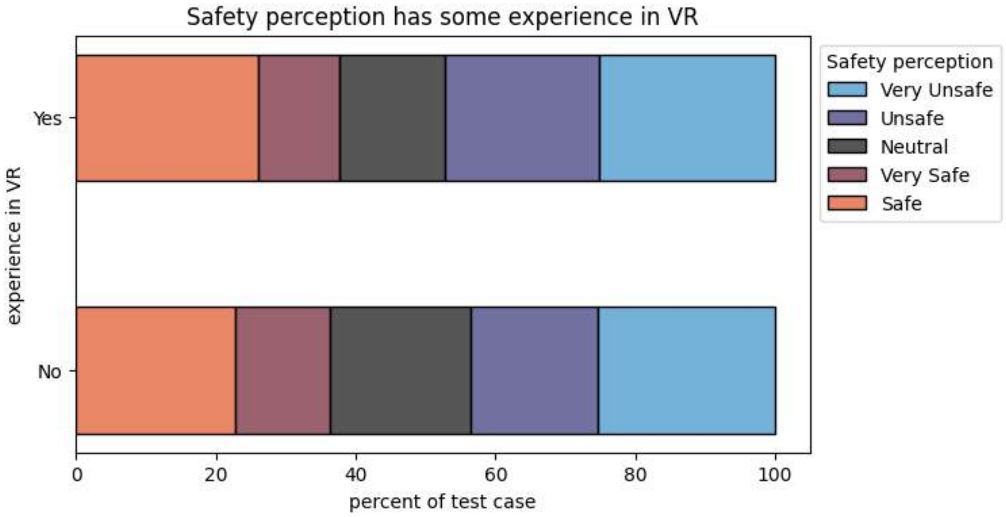


Figure 6.10: From the graphs, it is evident that participant who already used VR has no effect on safety perception.

0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.4 the "p-value" parameter has a value of 0.80, which reveals that there is no statistical evidence that there is a difference between the distributions.

Table 6.4: Statistics for the test scenario based on participants who already used VR.

Variable	**Used VR**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety	Yes No	0.0 0.0	1.77 1.80	4.0 4.0	0.01e-10 (non-gausian) 0.06e-10 (non-gausian)	0.80	-
Level of realism	Yes No	0.0 0.0	3.82 3,9	5.0 5.0	0.04e-14 (non-gausian) 0.03e-13 (non-gausian)	0.60	-

Further, we analyze the level of realism perceived by the corresponding participants who have already used VR, as you can see in Figure 6.11. As speculated, the level of realism perceived among participants who already used VR was lower than that of participants who hadn't used VR before, because a first-time user's expectations would be lower compared to those of participants who had already used VR, this is consistent with the results of previous studies involving newcomers and experts of VR [14]. To analyze the distribution of the perceived levels of realism among the participants who had already used VR, we analyzed the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.4. The p-value threshold in this approach was set to 0.05 (as a rule of thumb), which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.4 the "p-value" parameter has a value of 0.60, which reveals that there is no statistical evidence that there is a difference between the distributions.

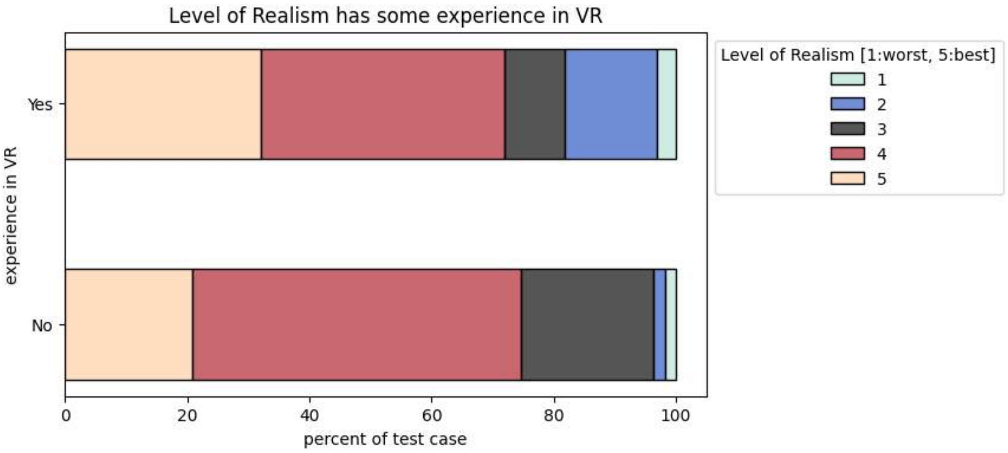


Figure 6.11: From the graphs, it is evident that participants who already used VR felt test scenarios were less realistic.

6.1.5 Do the participants' who play pc games as any impact on the level of safety perception and realism in SDC test cases?

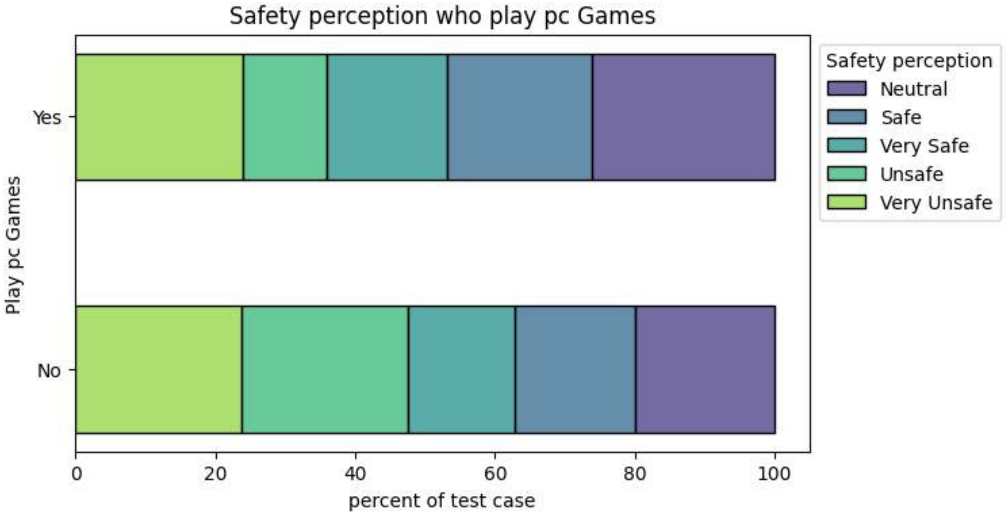


Figure 6.12: From the graphs, it is evident that participant who play pc Games has effect on safety perception.

When we analyze the participants who play PC games, do they have any effect on the perceived safety and realistic nature of the test scenarios? Figure 6.12 depicts the stacked histogram of the proportion of test cases for the level of safety of the participants who play pc games. When comparing the results, we found that participants who play PC games have a more positive perception of safety than participants who don't play any PC games. This was the expected result, as virtual simulators are based on engine simulators similar to PC games i.e, CARLA is based on Unreal Engine [2] which is also used in games like Fortnite [3] etc, participants who already played PC games felt more familiar with the environment.

We analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.5. The p-value threshold in this approach was set to 0.05 (as a rule of thumb) which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.5 the "p-value" parameter has a value of 0.93, which reveals that there is no statistical evidence that there is a difference between the distributions.

Table 6.5: Statistics for the test scenario based on participants who play PC Games.

Variable	**PC Games**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety	Yes	0.0	1.74	4.0	0.02e-16 (non-gausian)	0.93	-
	No	0.0	1.76	4.0	0.01e-5 (non-gausian)		
Level of realism	Yes	0.0	3.74	5.0	0.09e-19 (non-gausian)	0.04-e4	0.65
	No	0.0	4.28	5.0	0.01e-10 (non-gausian)		

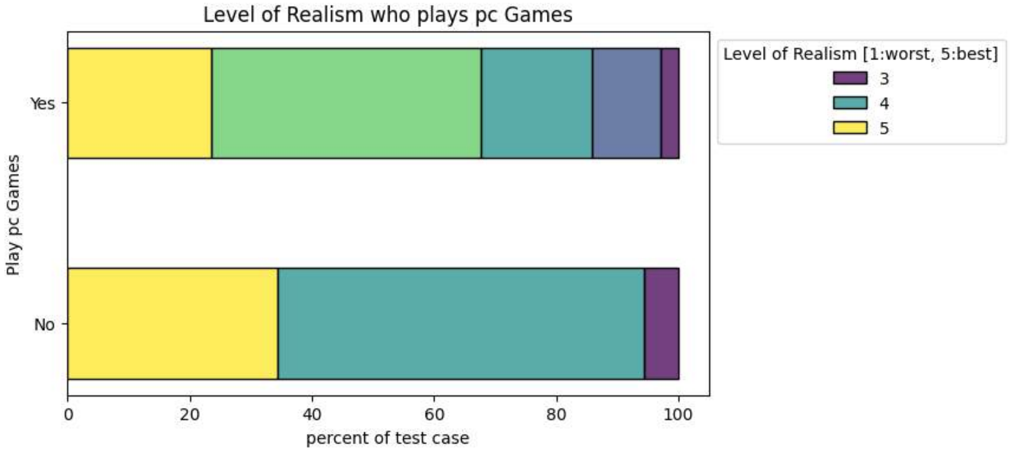


Figure 6.13: The graphs make it very clear that participants who has experience on PC games felt test scenarios are more realistic than those who don't play PC games.

Further, we analyze the level of realism perceived by the corresponding participants who play PC games, as you can see in Figure 6.13. As expected, the level of realism perceived among participants who play PC games is higher than that among participants who don't

[2] https://www.unrealengine.com/en-US
[3] https://www.epicgames.com/fortnite/en-US/home

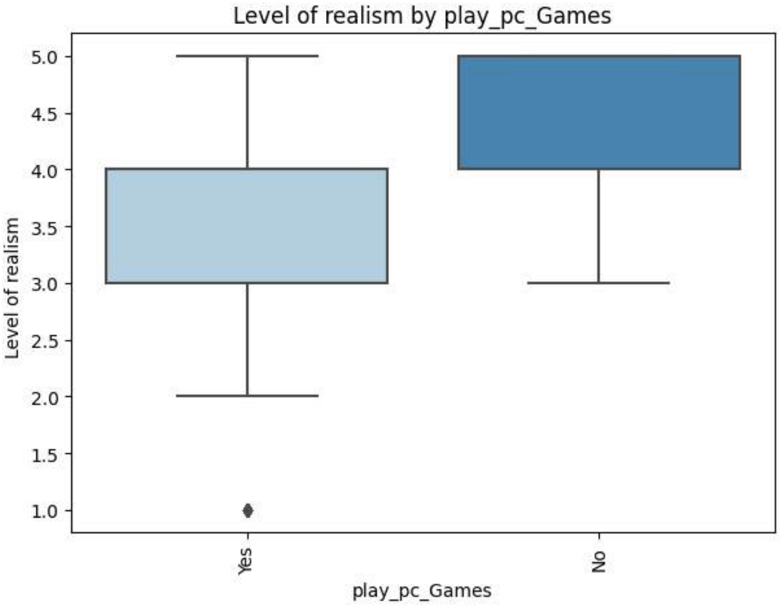


Figure 6.14: From the graphs, it seems clear that participants who has experience on PC Games impacts the level of realism of the test scenarios.

play PC games. To analyze the distribution of the perceived levels of realism among the participants' who play PC games, we use the Shapiro-Wilk test of normality to verify that the distributions we want to verify are mostly non-gaussian in nature, as can be seen in Table 6.5. Due to this observation, as we can see in Figure 6.14, which yields a significance threshold (p-value) < 0.04e-4, indicating that the level of realism is statistically significant. We use the unpaired Wilcoxon test and make use of the Vargha-Delaney effect size metric, $\hat{A}_{12}$ has an effect size of 0.65. This result reveals to us that the effect size is largely significant, as seen in Table 5.1. It proves with concrete evidence that participants who play PC games have a significantly higher perception of realism than participants who don't play PC games.

6.1.6 Does the participants' years of experience in testing as any impact on level of safety perception and realism in SDC test cases?

When we analyze the participants' years of experience in testing, do they have any effect on the perceived safety and realistic nature of the test scenarios? Figure 6.15 depicts the stacked histogram of the proportion of test cases for the level of safety of the participants' years of experience in testing. When comparing the results, we found that there was no difference in the proportion. Further, we analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.6. The p-value threshold in this approach was set to 0.05 (as a rule of thumb), which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.6 the "p-value" parameter has a value of 0.23, which reveals that there is no statistical evidence

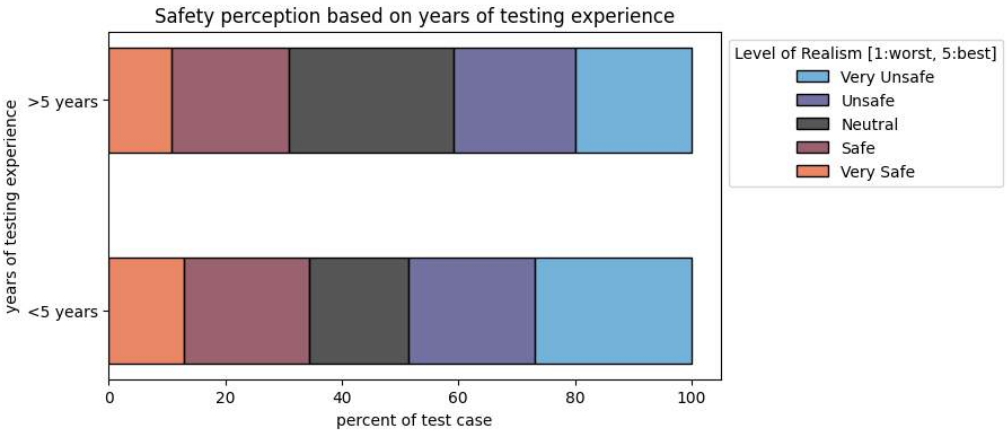


Figure 6.15: From the graphs, it is evident that participant years of experience in testing has no effect on safety perception.

Table 6.6: Statistics for the test scenario based on participants years of experience in testing.

Variable	used VR	Min	Mean	Max	Distribution	p-value	$\hat{A}_{12}$
Level of safety	Yes No	0.0 0.0	1.88 1.71	4.0 4.0	0.08e-6 (non-gausian) 0.03e-8 (non-gausian)	0.23	-
Level of realism	Yes No	0.0 0.0	3.87 3.82	5.0 5.0	0.01e-8 (non-gausian) 0.01e-19 (non-gausian)	0.87	-

that there is a difference between the distributions.

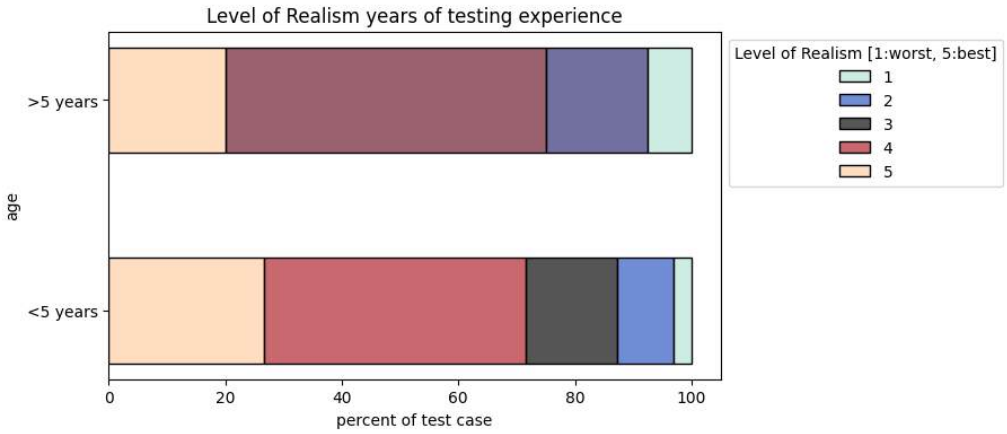


Figure 6.16: From the graphs, it is evident that participants who had 5 years and more experience in testing felt the test scenarios were more realistic.

Further, we analyze the level of realism perceived by the corresponding participants' years of experience in testing, as you can see in Figure 6.16. As expected, the level of realism perceived among participants who had 5 years or more participants years of experience in testing felt more realistic compared to participants with testing experience of fewer than 5 years. To analyze the distribution of the perceived levels of realism among the participants' years of experience in testing, we analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.6. The p-value threshold in this approach was set to 0.05 (as a rule of thumb) which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.4 the "p-value" parameter has a value of 0.87, which reveals that there is no statistical evidence that there is a difference between the distributions.

Finding 11. This analysis revealed that the backgrounds of participants can have a significant impact on the results of SDCs test cases. Participants with prior gaming experience perceived scenarios to be safer. Participants who were IT experts versus non-IT experts had already experienced VR, and their level of experience did not affect their perception of safety.

However, participants with IT expertise, experience playing games or using VR, and the testing experience felt the test scenarios were more realistic than those who did not.

6.2 Lesson learned on comparing the safety perception of the two control groups of participants with one group WITHOUT interaction and another WITH interaction (RQ$_4$)

As can be seen in the Figure 4.7 we also conducted an additional test scenario in the CARLA simulator, in which we split the participants into two different control groups (A and B). where Group A would have the opportunity to interact with the vehicle and Group B would not have any such opportunity. In this Final Scenario, we fabricated an accident on purpose in order to observe how participants deal with it and to compare how various control groups with and without interaction perceived their level of danger.

Figure 6.17 depicts the stacked histogram of the proportion of test cases with test group with versus without interaction. We observed that the participants' perception of safety was slightly higher in people from group A where they could interact with vehicles compare to test group B where they couldn't. As Expected participants' felt safer when they had some kind of interaction. This provides more concrete evidence on RQ$_4$ (*"what is the human perception of SDC's test failures/safety when humans can interact with the car?"*) as we observed in Section 5. Observing from the comments, *"If i had the control then we could have avoided the crash. "*(P12),*"It was safe until the accident. I'm pretty sure that I could have avoided the accident if I still have control over the car. "*(P30),*"Interaction would be nice, but driving into the lamp post was very unexpected, so maybe I couldn't prevent it"*(P34) Those in group B without interaction strongly believed they could avoid accidents, demonstrating the significance of

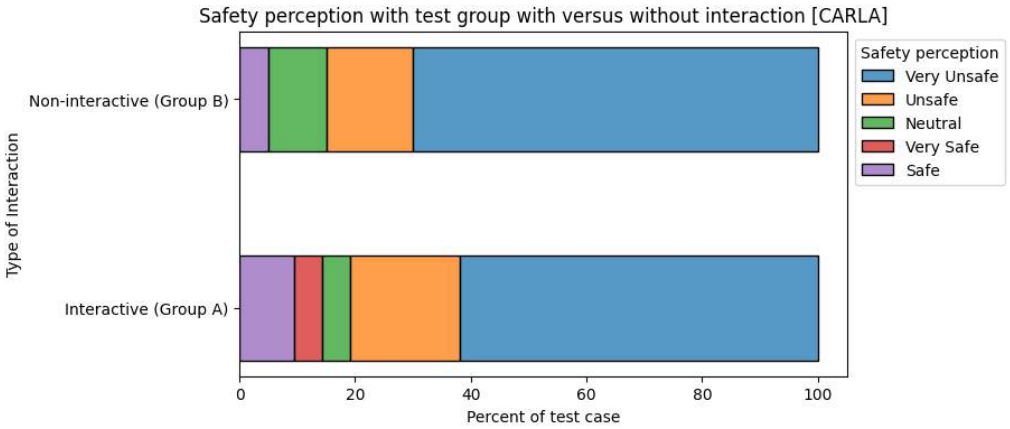


Figure 6.17: It is clear from the graphs that participants in test Group B (which did not involve any interaction) experienced a greater feeling of vulnerability on safety compared to participants in test Group A. (with interaction).

the slight interaction in enhancing safety. These were distinct and inexplicable results, as none of the automated testing research considers any human interaction, but these results clearly show how much this impairs overall perceived safety.

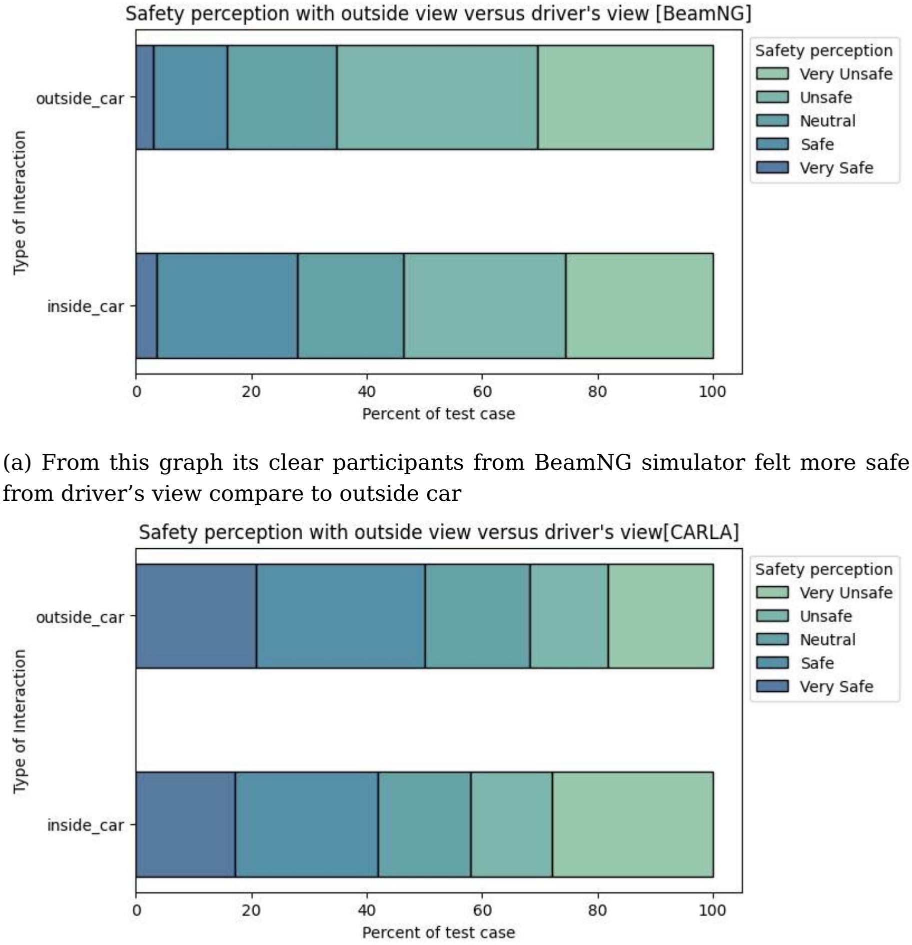
Table 6.7: Statistics for the test scenario based on test groups.

Variable	**Interaction**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety	Yes No	0.0 0.0	0.76 0.5	4.0 4.0	0.01e-3 (non-gausian) 0.06e-4 (non-gausian)	0.60	-

Further, we analyze the distribution of the perceived levels of safety among the participants with different test groups, we analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.7. The p-value threshold in this approach was set to 0.05 (as a rule of thumb) which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.7 the "p-value" parameter has a value of 0.60, which reveals that there is no statistical evidence that there is a difference between the distributions.

Finding 12. According to the results of the control experiment, the test group without interaction strongly believed that with interaction, they could prevent an intentional accident. This demonstrates the importance of interaction in testing a self-driving car, as compared to the test group without interaction. These results were not statistically significant.

6.3 Does the safety perception change from a different view (outside view versus driver's view)?



(a) From this graph its clear participants from BeamNG simulator felt more safe from driver's view compare to outside car

(b) From this graph its clear participants from CARLA simulator felt more unsafe from driver's view compare to outside car

Figure 6.18: Visualization of Safety perception from the outside view versus the driver's view

As Figure 6.18a depicts the stacked histogram of the proportion of test cases with different views (outside view versus driver's view) on BeamNG simulator, we observed that the participants' perception of safety was higher in the driver's view compared to an outside view of the car. It is also depicted in the histogram seen in Figure 6.18a, a greater proportion of participants felt safe or very safe in the same scenario from the driver's view than from the car's outside view.

Judging from the comment, *"The view angle was not the best"* (P14) (driver's view), *"Again, I perceived the bumps and the fact that the car is "behaving" badly, but due to my restricted point of view, I felt a little bit safer than previously."*(P28) (driver's view),*"I put 'neutral' because i felt it safer than from the outside view. However, the car still went a bit off-road, especially at the end.c."* (P4) (driver's view), we can see that they felt uncomfortable in the angle of view in the BeamNG simulator from inside the car compared to outside the car. We hypothesize that, because BeamNG VR is not a Full VR experience, participants may feel safer in the outside view.

Additionally, we analyze the distribution of the perceived levels of safety among the participants with different views (outside view versus driver's view), we analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.8. The p-value threshold in this approach was set to 0.05 (as a rule of thumb), which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.7 the "p-value" parameter has a value of 1, which reveals that there is no statistical evidence that there is a difference between the distributions.

Table 6.8: Statistics for the test scenario based on different views (outside view versus driver's view)

Variable	**view**	**Min**	**Mean**	**Max**	**Distribution**	**p-value**	$\hat{A}_{12}$
Level of safety BeamNG	outside driver's	0.0 0.0	1.23 1.52	4.0 4.0	0.06e-9 (non-gausian) 0.06e-9 (non-gausian)	1.0	-
Level of safety CARLA	outside driver's	0.0 0.0	1.71 1.78	4.0 4.0	0.09e-13 (non-gausian) 0.09e-13 (non-gausian)	1.0	-

Further, we analyze the distribution of the perceived levels of safety among the different views (outside view versus driver's view) on CARLA simulator, as we see in the Figure 6.18b depicts the stacked histogram of the proportion of test cases with different views (outside view versus driver's view) on CARLA simulator, we observed that the participants' perception of safety was higher in the outside view compared to a simulator view of the car. It is also depicted in the histogram seen in Figure 6.18b, a greater proportion of participants felt safe or very safe in the same scenario from the outside view than from the car's driver's view. The results are the opposite of what we observed with the BeamNG simulator, demonstrating that the simulator and view angle on each simulator plays a significant role in how participants perceive safety. We didn't find any qualitative comments on these results. Further, we analyze the distribution of the perceived levels of safety among the participants with different views (outside view versus driver's view) on CARLA simulator, and we analyze the difference in the distribution of the two categories by performing the Shapiro-Wilk test on the data samples obtained, which can be seen in Table 6.8. The p-value threshold in this approach was set to 0.05 (as a rule of thumb), which indicates that if the p-values obtained during the test were less than 0.05, then there is a statistically significant difference between the scores. As you can see in Table 6.7 the "p-value" parameter has a value of 1, which reveals that there is no statistical evidence that there is a difference between the distributions.

Finding 13. The results show the view has a drastic impact on safety perception. As seen BeamNG simulator, participants felt safer from the outside view of the car, and in CARLA participants felt safer from the driver's view. This also provides evidence that results in various bases depending on the simulator and angle of view. This is an important finding because SBST and SDC researchers only evaluate cars from the outside, and the results show that human safety perceptions differ from the outside view and the driver's perspective.

6.4 Summarizing qualitative aspects from participants on the simulators

In the final step of the experiment, we have the participants rate the overall quality of the simulators (BeamNG and CARLA) as well as the driving performance of the AI-agent.

BeamNG simulator

- **Environment:** Participants felt the landscape and road design were unrealistic in the real world. Also, there were no other cars or vehicles, which made the simulation in the surroundings, with which most of the participants had less resemblance with the real world. As previous research on SBST [22] and [8] only focused on the ego vehicle for testing SDCs using BeamNG simulator, future research testing SDCs should focus on test scenarios involving multiple vehicles.
- **Graphics:** Some of the participants think that the BeamNG simulator is comparable to cutting-edge computer games that centre on driving a car; based on the feedback, it seemed like the graphics were low, and the car was more fictional and rigid. Surprisingly, BeamNG has more realistic physics in practice. Some factors that contributed to this low level of realism are trivial scenarios with no traffic (other cars, pedestrians), static objects, etc.
- **Obstacles:** The obstacles were not placed realistically, and their placement was also unrealistic. To make the game more realistic, it was suggested that more dynamic obstacles, traffic, and pedestrian be added.
- **Driving of AI:** Another primary factor that contributes to an unsafe test case in the BeamNG simulator is a vehicle that veers dangerously close to the roadway's edge. The vast majority of participants suggested slowing down and improving artificial intelligence's driving performance in curved sections of the road. Previous research on SBST [22] and [8] on testing SDCs does not focus on using advanced AI-agent which is a limitation of BeamNG simulator. In future research testing SDCs, it is better to focus on the integration of advanced AI for testing SDCs.

6.4.1 CARLA simulator

- **Environment:** The majority of people who took part responded positively to the city map, the traffic rules, the city speed limits, and the dynamic weather conditions, all of

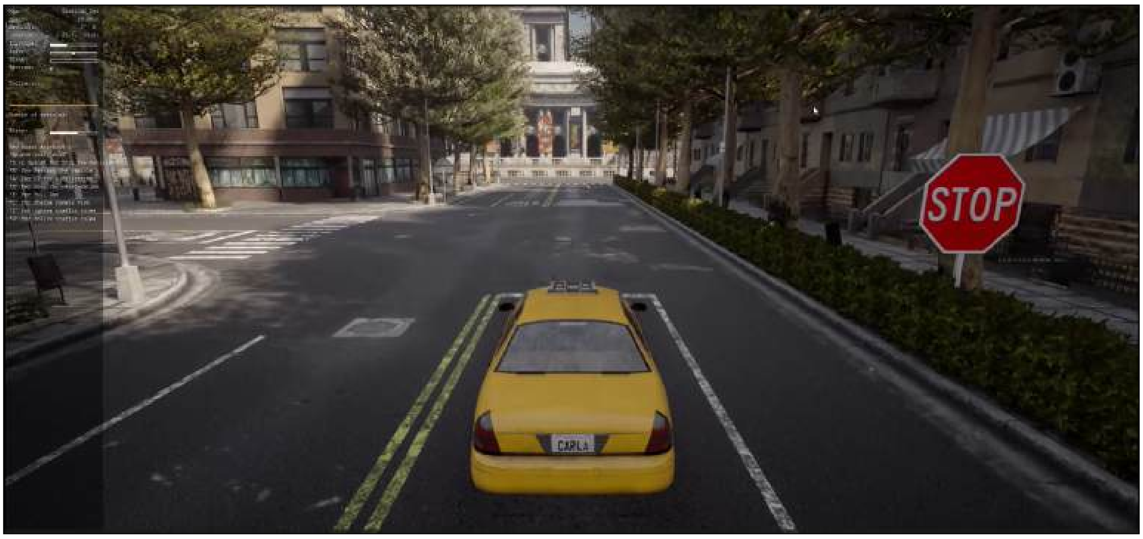


Figure 6.19: The figure of CARLA simulator where the car didn't stop for stop sign

which contributed to a greater sense of immersion in the scenario on CARLA. However, some of them thought the physics engine of the car was not as realistic as it should be, and one of the most noteworthy pieces of feedback/bug we found in CARLA VR with the dynamic weather was that raindrops came inside the car, which was not as realistic as it should have been.

- **Graphics:** CARLA simulator was a full 3D VR experience with a 360° view, which was a huge indication of trust in the simulator. However, there were delays in the animation, and the latency was not significantly reduced in VR in comparison to CARLA. special VR with an outside view, but the quality was very low, which affected how the participants perceived the environment.
- **Obstacles:** Participants liked CARLA dynamic obstacles like other cars and pedestrians, but one of the main feedbacks from the participants was that traffic was haltingly slow, which made the overall experience frustrating.
- **Driving of AI:** In summary, the AI was much better at driving, but switching lanes was not very safe and could have caused accidents. The artificial intelligence in the CARLA simulator didn't seem to track the stop signs, and the turn signal on the car wasn't working when making turns as we can see in the Figures 6.19 and 5.5a. This was one of the major flaws that we discovered CARLA. The artificial intelligence was unstable when turning, and in one of the test cases, it went above the roundabout. This occurred for the majority of the participants.

6.4.2 Participant' feedback on the experiment

Comments collected from the survey participants mentioned their experiences and suggestions to improve are summaries Table 6.9:

Table 6.9: What makes the safety perception low for CARLA and BeamNG

Category	Description	Comments	Participants
Limitations of the AI-agent	Enhance the level of artificial intelligence that controls the vehicle.	*"I think there is a good improvement, things can be tested in simulation. From the other side, there are margin to improve the AI driving the car. I would be happy to take part again on the study."*(P1)	P1
Driving Confidence in SDC	There is no longer any confidence in the driverless car.	*"I do not feel comfortable to use a real SDC once. There are too many misbehaviors and the VR was not close enough to reality."*(P2) *" I already thought before hand self driving cars are unsafe however now I feel even more that way. I think there needs to be done much more research"*(P3) *"I can say that self-driving cars still has a lot of improvements to make to become really safe. "*(P20) *"Self-driving cars are unsafe "*(P30)	P2, P3, P20
Raise awareness of the Safety of SDCs	Bring more attention to the issue of autonomous vehicles and their safety.	*"I think it made me more aware of all the different scenarios that a car can find in real life.."*(P5) *"I am not yet convienced that SDCs are at a level where I would sit back and relax"*(P8) *"It is very difficult for self driving car to tackle with real world drivers as they can change their decision. "*(P10) *"I understand that it is difficult for AI to make steady decisions especially when the pedestrains are unpredictable. "*(P15) *"Made me more skeptical towards it. But the good thing is that with the failed simulators, the car can learn lots of different things which would eventually make it safer. Exciting times ahead. "*(P19) *"I have more respect of it now, I think its not thta easy to give control over own life a AI, but still it stayed on the road and followed red lights. "*(P21) *" But it was also seen, that the self driving cars are not safe yet."*(P30) *" Made me understand how many scenarios have to be faced by the car. And how difficult it is to design one."*(P32) *" I am a strong evangelist for self-driving cars, becaus I use since 3 years a Tesla. The car drives on the highway smoother and more secure than myself. Unfortulately FSD is not yet allowed in Europe, but I am looking forward to have it."*(P39)	P5, P10, P15, P19, P21, P30, P32, P39
Interaction with the car		*"Could have a higher level of control over the interactions "*(P12) *"The interactive experiments made me see that a combination of self-driving AI with human control could be a way to solve the initial phases of self-driving cars, until cars are properly trained. "*(P5) *"find the usage of interaction and interesting idea. "*(P18) *"I really appreciated the possibility to interact with the simulator "*(P27)	P12, P5, P18, P27
Raise the level of complexity	Add more traffic and obstacles to the game to make the scenarios more difficult.	*"I would suggest to test it with a lot of traffic jams and using differenct scearios like understanding if it can recognize ambulance siren and give the space."*(P1) *"I would suggest to make longer scenarios "*(P22)	P1, P22, P31

		"The simulations are sometimes too short to get a feeling for the capabilities of the AI that is driving the car."(P31)	
Experiment setup	well planned and professional experiment setup	*"The setup done was very professional, all details were briefed up and all my queries were answered during the experiment. "*(P17) *"It is a nice way to learn self driving and experience driving without being on the road. "*(P34) *"The experiment was cool."*(P36) *"Great experience and VR usecase is great."*(P39) *"very good testing setup to start withd, full interaction would be nice to somewhat overrun the AI driver in critical scenarios."*(P41)	P17, P34, P39, P41

6.4.3 Lessons Learned

As indicated in section 5, incorporating humans into the testing process helps to improve both the level of confidence in SDCs as well as the quality of the test case. One of the most important findings was that the participants' confidence levels increased as a direct result of the interaction that took place in the vehicle. Which answer the RQ_4 will clear evidence as shown in Table 5.10 with a p-value > 0.001. When we did the further investigation on how the participant's background influences their perceived safety in Section 6.1 we saw that age plays an important role. in how realistic the simulators are. Participants in the age group 18–30 felt simulators. were more realistic than those who were older than 30 years. Other factors also affect the safety perception of SDCs scenarios for the participants who, as IT versus non-IT experts, had already used VR, and their level of experience didn't influence their perception of safety. Gender, IT versus non-IT expertise, VR usage, gamers, and level of experience have a significant impact on the realism of the test scenario.

Using VR in SDC revealed a variety of additional difficulties. During the course of our experiment, we had some participants who experienced motion sickness, as well as first-time VR users who reported feeling lightheaded and having the requirement to take extended breaks between simulations. which would make it useful for future researchers to consider this aspect, improve the graphics with simulators, set up the experiment with longer breaks, and give more warm-up scenarios so that the user can get used to the VR before real-life scenarios.

In order to obtain objective results from the SDCs experiment, it is crucial to diversify the test group. When we ask participants **Question**: *"Did this experiment change the way you thought about the Self-driving Cars safety?"* 39% (Yes), 41% (No), and 19% (Maybe), you can find further qualitative feedback on Table 6.9 . In summary, the majority of participants believe that SDC-Alabaster is closely associated with human perceptions of SDC test failures and safety and that the SDC test case is somewhat relevant to the real world.

Chapter 7

Threat to Validity

Internal Validity. Threats to internal validity may concern, as for previous work [8, 21], the cause-and-effect relationships between the technologies used to generate the scenarios and their components and the corresponding outcomes, which are dependent on the realism of our scenarios. Indeed, in the BeamNG simulator we did not recreate every element, especially dynamic objects, found on actual roads (traffic, pedestrians, etc.). To increase our internal validity, we used the CARLA simulator, which has dynamic objects such as weather, traffic, pedestrians, etc., to simulate a realistic environment.

Another threat is a potential bias of the result, as some of the participants were using VR for the first time, and a few of them have little experience of evaluating critical driving scenarios; hence, the participant's perception of safety could be biased. To reduce this bias, we recruited participants from a variety of fields and experiences, as shown in Table 4.2, and provided them with warm-up test scenarios so they could become accustomed to the environment; the results of the warm-up round were not included in our analysis.

Although BeamNG and CARLA have different roads, we use the same automatically generated tests as we see in the section in 3.2, which were developed for the BeamNG simulator and not for CARLA, so there may be differences in the roads' features and attributes between the two simulators, which might lead to different test case outcomes. For future work, we intend to create a road and test case that can be used in both simulators and is therefore generalized.

We also use different AI agents for BeamNG and CARLA, which may have vastly different driving and traffic management characteristics. For future work, we plan to change SDC-Alabaster and generalize the AI agent so that future researchers and practitioners can integrate any AI agent and utilize SDC-Alabaster for automated testing of their SDCs.

External Validity. Finally, threats to external validity might be the low number of participants, as we had 41 participants in our experiment, and there could be gender bias because of the unequal distribution of participants (men: 31, female: 10), as we saw in the discussion section 6.1.2. Furthermore, most of the participants have experience in computer science or related fields, as we can see in Table 9.2, so the population was not diverse and there could be uncontrolled confounding variables biasing the results. We will overcome this threat by planning to conduct more experiments with people of diverse backgrounds.

In scenarios, we use a flat 2D plane. We do not take into account Z-coordinates, which would be more realistic and produce more generalizable results.

Our findings cannot be generalized to the entire universe of open-source CPS simulation environments in other domains. Consequently, additional replications and studies incorporating additional data and other CPS domains are desirable.

Chapter 8

Conclusion and Future Work

Automated Testing Self-driving cars are challenging due to the lack of human feedback and loss of confidence. To overcome this, we introduced human-in-the-loop simulation-based testing for self-driving cars, called SDC-Alabaster, which generated test scenarios with static and dynamic objects and executed the test in the BeamNG and CARLA SDC simulation environments. The SDC-Alabaster also supports the visualization of simulated test scenarios in a VR headset, allowing humans to feel more immersed in the SDCs.

In order to answer our technical research question, we implemented static objects like road bumps, trees, and cylinders in BeamNG simulator and dynamic traffic (other cars, pedestrians), as well as dynamic weather in CARLA simulator(RQ$_1$). We integrated both BeamNG and CARLA simulators with a VR headset. BeamNG does not support VR, so we used an external driver to convert the experience into a VR format(RQ$_2$).

To address RQ$_3$ and RQ$_4$, we conducted a controlled experiment with 41 participants from diverse backgrounds and experiences. The experiment lasted a maximum of one hour and included various tests in BeamNG and CARLA simulators with a different view (outside car and driving view), with and without a VR headset, as well as test scenarios with vehicle interaction.

Our findings from the experiment show that SDC-Alabaster case classification (pass/fail) closely resembles the human perception of SDCs' test failures/safety (RQ$_3$), and perceptions of safety and realism vary with the simulating environment (*i.e.*, with or without VR). In addition, we discovered that the failure cases that are most important to tests are perceived as less realistic (RQ3). Our results show the perception of realism and safety among users is significantly dependent on the presence of obstacles in the given scenario. With a p-value>0.64e-8, the distribution is 56%($\hat{A}_{12}$) statically different for the perception of realism, and with a p-value>0.2e-37, the distribution is 100%($\hat{A}_{12}$) statically different for the perception of safety. Our results also show Carla was more realistic and safer than BeamNG because participants found CARLA's scenarios more realistic with a p-value> 0.01e-16, the distribution is 85%($\hat{A}_{12}$) statically different, and for safety with a p-value> 0.05e-10, the distribution is 68%($\hat{A}_{12}$) statically different(RQ$_3$). We also found that interactions with cars make humans feel safer compared to when there is no interaction. We also found that interactions with vehicles make humans safer compared to when there is no interaction; with statistical evidence of a p-value > 0.001, the distribution is 36%($\hat{A}_{12}$) statistically different (RQ$_4$). In addition, we discovered that the age, gender, field of expertise, previous use of virtual reality, computer gaming experience, and the number of years of testing experience of the participants all play a significant role in the level of safety and perception of realism.

8.1 Future Work

Future work will be directed in various directions. Here are some future work suggestions, some of which are based on the experiment's results:

- (i) **Voice feedback in VR:** To enhance human interaction with a car using a VR headset by incorporating voice-command feedback and the user's ability to provide feedback and control the vehicle without removing the VR headset.
- (ii) **Training model with a generated dataset**: We will generate a large number of datasets that could be used to train the AI model and enhance the AI-agents of SDCs based on the feedback and interaction with the test scenario.
- (iii) **Generate dynamic roads and maps in CARLA**: We have seen that the BeamNG simulator allows us to dynamically create roads; accordingly, we would like the CARLA simulator to also allow us to generate and integrate dynamic maps and roads.
- (iv) **Improve the AI-agent traffic cars**: The post-survey responses of the participants provided us with feedback to improve the CARLA traffic simulator's AI-agent for other vehicles. One of the objectives of future research is to implement AI-adaptive traffic vehicles.
- (v) **Integrate participants' driving**: To integrate human driving to simulate and replicate human-based experimentation and to use test scenarios to compare the results of participant driving and AI-agent driving. In addition, the test scenario should use actual roads.
- (v) **Recognize "STOP" sign**: In our experiment, we found the CARLA AI-agent doesn't recognize "STOP" signs, which was a major finding and drawback of CARLA simulator. For future work to fix the AI-agent to recognize signs and other traffic signs.

Also, we aim to replicate the study by involving additional participants from diverse backgrounds.

Chapter 9

Appendices

9.1 Additional data on the participant

In this section, we may observe all additional data about participants that will not be considered for results analysis.

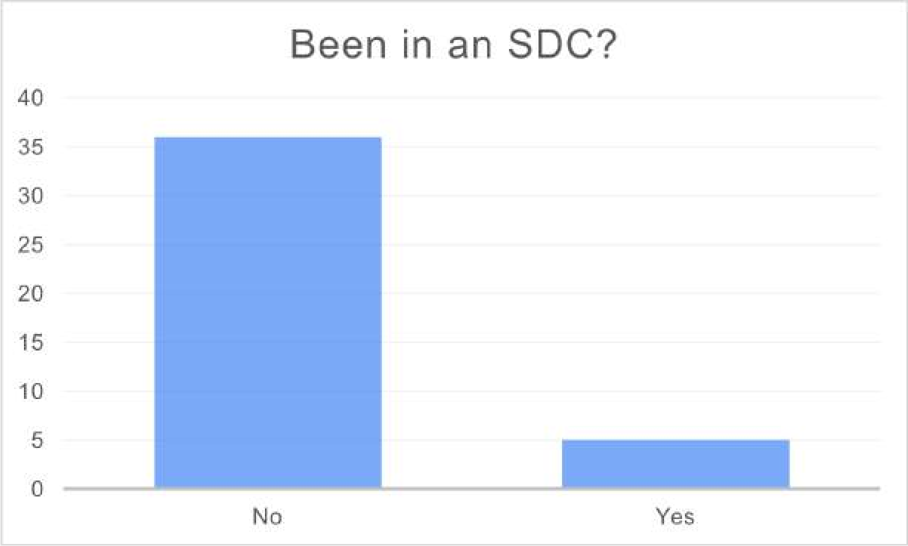


Figure 9.1: This graph demonstrates that the majority of participants were trying SDC for the first time.

Table 9.1: Summarizes participants with which type of vehicle drive and how many years of driving experience they have

Vehicle driven	**Years of driving experience**					**Total**
	less than one year	*1-2 years*	*3-6 years*	*6-10 years*	*>10 years*	
2 wheelers (Motorbikes)	2	-	3	-	-	5
2 wheelers (Motorbikes), 4 wheelers (light-duty vehicles like cars and vans)	-	2	4	3	3	12
4 wheelers (light-duty vehicles like cars and vans)	-	2	9	3	7	21
4 wheelers (light-duty vehicles like cars and vans), 4> Wheelers(heavy-duty vehicles like trucks, buses, and coaches)	-	-	1	-	1	10
None	1	-	-	-	-	1
Total	3	4	17	6	11	**41**

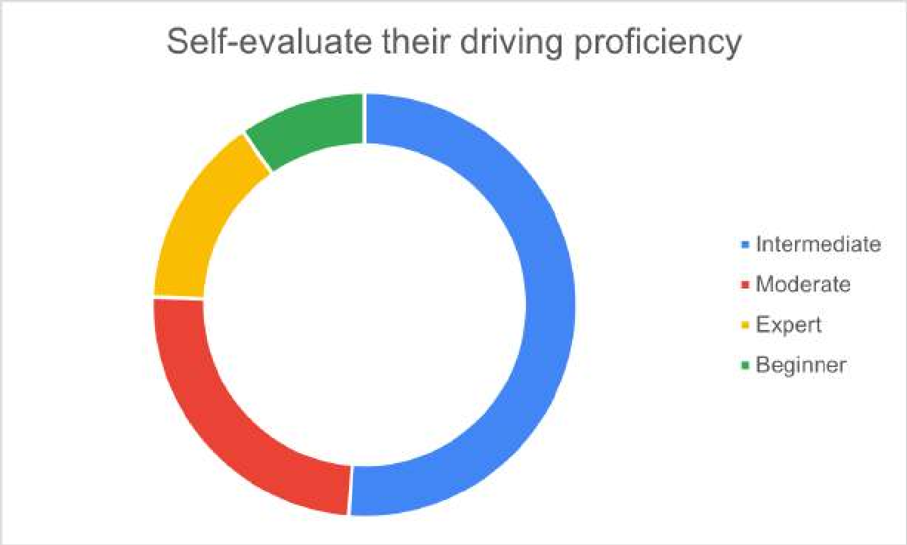


Figure 9.2: This graph clearly shows most of the participants have some kind of experience driving.

Bibliography

[1] R. B. Abdessalem, A. Panichella, S. Nejati, L. C. Briand, and T. Stifter. Testing autonomous cars for feature interaction failures using many-objective search. In M. Huchard, C. Kästner, and G. Fraser, editors, *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE 2018, Montpellier, France, September 3-7, 2018*, pages 143–154. IEEE, ACM, 2018.

[2] A. Afzal, D. S. Katz, C. L. Goues, and C. S. Timperley. A study on the challenges of using robotics simulators for testing. *arXiv preprint arXiv:2004.07368*, 2020.

[3] M. Althoff, M. Koschi, and S. Manzinger. Commonroad: Composable benchmarks for motion planning on roads. In *IEEE Intelligent Vehicles Symposium, IV 2017, Los Angeles, CA, USA, June 11-14, 2017*, pages 719–726. IEEE, 2017.

[4] BeamNG.research. BeamNGpy. `https://github.com/BeamNG/BeamNGpy`. Accessed: 2020-08-10.

[5] BeamNG.tech. Beamng.research. `https://documentation.beamng.com/beamng_tech/`. Accessed: 2022-07-31.

[6] L. P. Berg and J. M. Vance. Industry use of virtual reality in product design and manufacturing: a survey. *Virtual Real.*, 21(1):1–17, 2017.

[7] C. Birchler, N. Ganz, S. Khatiri, A. Gambi, and S. Panichella. Cost-effective simulation-based test selection in self-driving cars software. *arXiv preprint arXiv:2211.11409*, 2022.

[8] C. Birchler, N. Ganz, S. Khatiri, A. Gambi, and S. Panichella. Cost-effective simulation-based test selection in self-driving cars software with sdc-scissor. In *29th IEEE International Conference on Software Analysis, Evolution, and Reengineering, Honolulu, USA (online), 15-18 March 2022*. ZHAW Zürcher Hochschule für Angewandte Wissenschaften, 2022.

[9] C. Birchler, S. Khatiri, P. Derakhshanfar, S. Panichella, and A. Panichella. Automated test cases prioritization for self-driving cars in virtual environments. *arXiv preprint arXiv:2107.09614*, 2021.

[10] C. Birchler, S. Khatiri, P. Derakhshanfar, S. Panichella, and A. Panichella. Single and multi-objective test cases prioritization for self-driving cars in virtual environments. *ACM Trans. Softw. Eng. Methodol.*, apr 2022. Just Accepted.

[11] D. Bogdoll, S. Orf, L. Töttel, and J. M. Zöllner. Taxonomy and survey on remote human input systems for driving automation systems. In K. Arai, editor, *Advances in Information and Communication*, pages 94–108, Cham, 2022. Springer International Publishing.

[12] Carla. carla. `https://github.com/carla-simulator/carla/tree/master/PythonAPI`. Accessed: 2022-11-07.

[13] E. Castellano, A. Cetinkaya, C. H. Thanh, S. Klikovits, X. Zhang, and P. Arcaini. Frenetic at the SBST 2021 tool competition. In *14th IEEE/ACM International Workshop on Search-Based Software Testing, SBST 2021, Madrid, Spain, May 31, 2021*, pages 36–37. IEEE, 2021.

[14] E. Castellano, S. Klikovits, A. Cetinkaya, and P. Arcaini. Freneticv at the sbst 2022 tool competition. In *2022 IEEE/ACM 15th International Workshop on Search-Based Software Testing (SBST)*, pages 47–48, 2022.

[15] M. S. Corporation. Carsim adas: Moving objects and sensors. `https://www.carsim.com/products/supporting/objects_sensors/index.php`. Accessed: 2022-08-20.

[16] P. Daignault and P. Delhomme. Attitudes des jeunes automobilistes à l'égard des principales actions contre l'insécurité routière en france. *Pratiques Psychologiques*, 17(4):373–389, 2011.

[17] M. Dalboni and A. Soldati. Soft-body modeling: A scalable and efficient formulation for control-oriented simulation of electric vehicles. In *IEEE Transportation Electrification Conference and Expo (ITEC)*, pages 1–6, 2019.

[18] M. R. Desselle, R. A. Brown, A. R. James, M. J. Midwinter, S. K. Powell, and M. A. Woodruff. Augmented and virtual reality in surgery. *Computing in Science Engineering*, 22(3):18–26, 2020.

[19] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[20] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. López, and V. Koltun. CARLA: an open urban driving simulator. In *1st Annual Conference on Robot Learning, CoRL 2017*, volume 78 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 2017.

[21] A. Gambi, T. Huynh, and G. Fraser. Generating effective test cases for self-driving cars from police reports. In M. Dumas, D. Pfahl, S. Apel, and A. Russo, editors, *Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2019, Tallinn, Estonia, August 26-30, 2019*, pages 257–267. ACM, 2019.

[22] A. Gambi, G. Jahangirova, V. Riccio, and F. Zampetti. SBST tool competition 2022. In *15th IEEE/ACM International Workshop on Search-Based Software Testing, SBST@ICSE 2022, Pittsburgh, PA, USA, May 9, 2022*, pages 25–32. IEEE, 2022.

[23] A. Gambi, M. Mueller, and G. Fraser. Asfault: Testing self-driving car software using search-based procedural content generation. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pages 27–30, 2019.

[24] A. Gambi, M. Mueller, and G. Fraser. *Automatically Testing Self-Driving Cars with Search-Based Procedural Content Generation*, page 318–328. Association for Computing Machinery, New York, NY, USA, 2019.

[25] A. Gambi, M. Mueller, and G. Fraser. Automatically testing self-driving cars with search-based procedural content generation. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2019, page 318–328, New York, NY, USA, 2019. Association for Computing Machinery.

[26] C. A. González, M. Varmazyar, S. Nejati, L. C. Briand, and Y. Isasi. Enabling model testing of cyber-physical systems. In *Proceedings of the 21th ACM/IEEE International Conference on Model Driven Engineering Languages and Systems*, MODELS '18, page 176–186, New York, NY, USA, 2018. Association for Computing Machinery.

[27] G. Grano, A. Ciurumelea, S. Panichella, F. Palomba, and H. C. Gall. Exploring the integration of user feedback in automated testing of android applications. In *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 72–83, 2018.

[28] R. Gutiérrez-Moreno, R. Barea, E. L. Guillén, J. Araluce, and L. M. Bergasa. Reinforcement learning-based autonomous driving at intersections in CARLA simulator. *Sensors*, 22(21):8373, 2022.

[29] HTC. Htc vive pro 2. `https://vive.com/us/product/vive-pro2/overview/`. Accessed: 2022-11-22.

[30] C. G. Huélamo, J. del Egido, L. M. Bergasa, R. Barea, E. L. Guillén, J. F. Arango, J. Araluce, and J. López. Train here, drive there: ROS based end-to-end autonomous-driving pipeline validation in CARLA simulator using the NHTSA typology. *Multim. Tools Appl.*, 81(3):4213–4240, 2022.

[31] D. Humeniuk, G. Antoniol, and F. Khomh. Ambiegen tool at the sbst 2022 tool competition. In *2022 IEEE/ACM 15th International Workshop on Search-Based Software Testing (SBST)*, pages 43–46, 2022.

[32] S. Int. Org. Standardization Geneva. Road vehicles - safety and cybersecurity for automated driving systems - design, verification and validation. *ISO*, 2020.

[33] N. Kalavas. Human feedback could help in the a.i. of self-driving cars. `https://www.y-mobility.co.uk/human-feedback-could-help-in-the-artificial-intelligence-ai-of-self-`
Accessed: 2022-05-08.

[34] P. Kaur, S. Taghavi, Z. Tian, and W. Shi. A survey on simulators for testing self-driving cars. In *2021 Fourth International Conference on Connected and Autonomous Driving (MetroCAD)*, pages 62–70, 2021.

[35] S. Khatiri, C. Birchler, B. Bosshard, A. Gambi, and S. Panichella. Machine learning-based test selection for simulation-based testing of self-driving cars software. *arXiv preprint arXiv:2111.04666*, 2021.

[36] N. Koenig and A. Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, volume 3, pages 2149–2154 vol.3, 2004.

[37] P. Koopman and W. Michael. Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety*, 2016.

[38] C. Law. Driverless car accidents – who's at fault? https://www.natlawreview.com/article/driverless-car-accidents-who-s-fault. Accessed: 2022-04-23.

[39] E. A. Lee. Cyber physical systems: Design challenges. In *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)*, pages 363–369, 2008.

[40] L. Li, W. Huang, Y. Liu, N. Zheng, and F. Wang. Intelligence testing for autonomous vehicles: A new approach. *IEEE Transactions on Intelligent Vehicles*, 1(2):158–166, 2016.

[41] A. Loquercio, E. Kaufmann, R. Ranftl, A. Dosovitskiy, V. Koltun, and D. Scaramuzza. Deep drone racing: From simulation to reality with domain randomization. *IEEE Transactions on Robotics*, 36(1):1–14, 2020.

[42] J.-L. Martin, S. Lafont, M. Chiron, B. Gadegbeku, and B. Laumon. Différences entre les hommes et les femmes face au risque routier. *Revue d'Épidémiologie et de Santé Publique*, 52(4):357–367, 2004. "Genres et Santé".

[43] J. Mesit and R. K. Guha. A general model for soft body simulation in motion. In *Proceedings of the 2011 Winter Simulation Conference (WSC)*, pages 2685–2697, 2011.

[44] O. Michel. Cyberbotics ltd. webots™: Professional mobile robot simulation. *International Journal of Advanced Robotic Systems*, 1(1):5, 2004.

[45] T. Mikkonen, K. Kemell, P. Kettunen, and P. Abrahamsson. Exploring virtual reality as an integrated development environment for cyber-physical systems. In M. Staron, R. Capilla, and A. Skavhaug, editors, *45th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2019, Kallithea-Chalkidiki, Greece, August 28-30, 2019*, pages 121–125. IEEE, 2019.

[46] T. Mikkonen, K.-K. Kemell, P. Kettunen, and P. Abrahamsson. Exploring virtual reality as an integrated development environment for cyber-physical systems. In *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 121–125, 2019.

[47] S. Nair, S. Shafaei, D. Auge, and A. C. Knoll. An evaluation of "crash prediction networks" (CPN) for autonomous driving scenarios in CARLA simulator. In H. Espinoza, J. A. McDermid, X. Huang, M. Castillo-Effen, X. C. Chen, J. Hernández-Orallo, S. Ó. hÉigeartaigh, and R. Mallah, editors, *Proceedings of the Workshop on Artificial Intelligence Safety 2021 (SafeAI 2021) co-located with the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021), Virtual, February 8, 2021*, volume 2808 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[48] M. Online. Autopilot fail! moment driverless tesla being summoned by owner across washington air field crashes into 2m private jet and keeps going even after collision. https://www.dailymail.co.uk/news/article-10745095/Moment-driverless-Tesla-summoned-owner-Washington-air-field-crashes-2m-jet.html. Accessed: 2022-04-23.

[49] S. Panichella, A. Gambi, F. Zampetti, and V. Riccio. Sbst tool competition 2021. In *2021 IEEE/ACM 14th International Workshop on Search-Based Software Testing (SBST)*, pages 20–27, 2021.

[50] N. Rajabli, F. Flammini, R. Nardone, and V. Vittorini. Software verification and validation of safe autonomous cars: A systematic literature review. *IEEE Access*, 9:4797–4819, 2021.

[51] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic. Cyber-physical systems: The next computing revolution. In *Design Automation Conference*, pages 731–736, 2010.

[52] V. Riccio and P. Tonella. Model-based exploration of the frontier of behaviours for deep learning system testing. In *Proceedings of the ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE '20, page 13 pages. Association for Computing Machinery, 2020.

[53] E. Rohmer, S. P. N. Singh, and M. Freese. V-rep: A versatile and scalable robot simulation framework. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1321–1326, 2013.

[54] G. Rong, B. H. Shin, H. Tabatabaee, Q. Lu, S. Lemke, M. Možeiko, E. Boise, G. Uhm, M. Gerow, S. Mehta, E. Agafonov, T. H. Kim, E. Sterner, K. Ushiroda, M. Reyes, D. Zelenkovsky, and S. Kim. Lgsvl simulator: A high fidelity simulator for autonomous driving. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6, 2020.

[55] R. Satava. Medical applications of virtual reality. *J Med Syst 19*, 1995.

[56] E. Seedhouse. Presence within the virtual reality environment of the international space station. *Virtual Reality 2022*, 2022.

[57] S. Sontges and M. Althoff. Computing the drivable area of autonomous road vehicles in dynamic road scenes. *IEEE Trans. Intell. Transp. Syst.*, 19(6):1855–1866, 2018.

[58] T. Stolte, S. Ackermann, R. Graubohm, I. Jatzkowski, B. Klamann, H. Winner, and M. Maurer. Taxonomy to unify fault tolerance regimes for automotive systems: Defining fail-operational, fail-degraded, and fail-safe. *IEEE Transactions on Intelligent Vehicles*, 7(2):251–262, 2022.

[59] T. N. Y. Times. 2 killed in driverless tesla car crash, officials say. `https://www.nytimes.com/2021/04/18/business/tesla-fatal-crash-texas.html`. Accessed: 2022-04-23.

[60] P. Tonella. Evolutionary testing of classes. In G. S. Avrunin and G. Rothermel, editors, *Proceedings of the ACM/SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2004, Boston, Massachusetts, USA, July 11-14, 2004*, pages 119–128. ACM, 2004.

[61] Unreal. Unreal engine technologies. `https://www.unrealengine.com/`. Accessed: 2022-08-20.

[62] A. S. Vempati, H. Khurana, V. Kabelka, S. Flueckiger, R. Siegwart, and P. Beardsley. A virtual reality interface for an autonomous spray painting uav. *IEEE Robotics and Automation Letters*, 4(3):2870–2877, 2019.

[63] T. washington post. Tesla driver faces felony charges in fatal crash involving autopilot. `https://www.washingtonpost.com/technology/2022/01/20/tesla-autopilot-charges/`. Accessed: 2022-04-23.

[64] Waymo. Waymo safety report. `https://storage.googleapis.com/waymo-uploads/files/documents/safety/2021-12-waymo-safety-report.pdf`. Accessed: 2022-10-24.

[65] J. Wu, Z. Huang, C. Huang, Z. Hu, P. Hang, Y. Xing, and C. Lv. Human-in-the-loop deep reinforcement learning with application to autonomous driving. *CoRR*, abs/2104.07246, 2021.

[66] R. V. Yampolskiy. Unpredictability of ai, 2019.

[67] D. Yeo, G. Kim, and S. Kim. Toward immersive self-driving simulations: Reports from a user study across six platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[68] O. Yildirim, C. Pidel, and M. West. Future mobility solutions: A use case for understanding how vr influences user perception. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 184–187, 2020.

[69] A. Yoganandhan, S. Subhash, J. Hebinson Jothi, and V. Mohanavel. Fundamentals and development of self-driving cars. *Materials Today: Proceedings*, 33:3303–3310, 2020. International Conference on Nanotechnology: Ideas, Innovation and Industries.

[70] H. Yoshitake, K. Futawatari, and M. Shino. A vr-based simulator using motion feedback of a real powered wheelchair for evaluation of autonomous navigation systems. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '21 Adjunct, page 26–29, New York, NY, USA, 2021. Association for Computing Machinery.

[71] E. Zapridou, E. Bartocci, and P. Katsaros. Runtime verification of autonomous driving systems in carla. In J. Deshmukh and D. Ničković, editors, *Runtime Verification*, pages 172–183, Cham, 2020. Springer International Publishing.

[72] W. Zhang, S. Fu, Z. Cao, Z. Jiang, S. Zhang, and S. Xu. An sdr-in-the-loop carla simulator for c-v2x-based autonomous driving. In *39th IEEE Conference on Computer Communications, INFOCOM Workshops 2020, Toronto, ON, Canada, July 6-9, 2020*, pages 1270–1271. IEEE, 2020.