Verification and Validation of Automated Systems' Safety and Security

Evaluation report including the evaluation of the improved V&V processes as well as framework limitations

Document Type	Report
Document Number	D5.6
Primary Author(s)	Alper Kanak (ERARGE)
Document Date	2023-05-26
Document Version	1.8 Final
Dissemination Level	Public (PU)
Reference DoA	2022-12-14
Project Coordinator	Behrooz Sangchoolie, behrooz.sangchoolie@ri.se, RISE Research Institutes of Sweden
Project Homepage	www.valu3s.eu
JU Grant Agreement	876852

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876852. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Austria, Czech Republic, Germany, Ireland, Italy, Portugal, Spain, Sweden, Turkey.

Disclaimer

The views expressed in this document are the sole responsibility of the authors and do not necessarily reflect the views or position of the European Commission. The authors, the VALU3S Consortium, and the ECSEL JU are not responsible for the use which might be made of the information contained in here.

Project Overview

Manufacturers of automated systems and the manufacturers of the components used in these systems have been allocating an enormous amount of time and effort in the past years developing and conducting research on automated systems. The effort spent has resulted in the availability of prototypes demonstrating new capabilities as well as the introduction of such systems to the market within different domains. Manufacturers of these systems need to make sure that the systems function in the intended way and according to specifications which is not a trivial task as system complexity rises dramatically the more integrated and interconnected these systems become with the addition of automated functionality and features to them.

With rising complexity, unknown emerging properties of the system may come to the surface making it necessary to conduct thorough verification and validation (V&V) of these systems. Through the V&V of automated systems, the manufacturers of these systems are able to ensure safe, secure and reliable systems for society to use since failures in highly automated systems can be catastrophic.

The high complexity of automated systems incurs an overhead on the V&V process making it time-consuming and costly. VALU3S aims to design, implement and evaluate state-of-the-art V&V methods and tools in order to reduce the time and cost needed to verify and validate automated systems with respect to safety, cybersecurity and privacy (SCP) requirements. This will ensure that European manufacturers of automated systems remain competitive and that they remain world leaders. To this end, a multi-domain framework is designed and evaluated with the aim to create a clear structure around the components and elements needed to conduct V&V process through identification and classification of evaluation methods, tools, environments and concepts that are needed to verify and validate automated systems with respect to SCP requirements.

In VALU3S, 13 use cases with specific safety, security and privacy requirements will be studied in detail. Several state-of-the-art V&V methods will be investigated and further enhanced in addition to implementing new methods aiming for reducing the time and cost needed to conduct V&V of automated systems. The V&V methods investigated are then used to design improved process workflows for V&V of automated systems. Several tools will be implemented supporting the improved processes which are evaluated by qualification and quantification of safety, security and privacy as well as other evaluation criteria using demonstrators. VALU3S will also influence the development of safety, security and privacy standards through an active participation in related standardisation groups. VALU3S will provide guidelines to the testing community including engineers and researchers on how the V&V of automated systems could be improved considering the cost, time and effort of conducting the tests.

VALU3S brings together a consortium with partners from 10 different countries, with a mix of *industrial partners* (25 partners) from automotive, agriculture, railway, healthcare, aerospace and industrial automation and robotics domains as well as leading *research institutes* (6 partners) and *universities* (10 partners) to reach the project goal.

Consortium

RISE RESEARCH INSTITUTES OF SWEDEN AB	RISE	Sweden
STAM SRL	STAM	Italy
FONDAZIONE BRUNO KESSLER	FBK	Italy
KNOWLEDGE CENTRIC SOLUTIONS SL - THE REUSE COMPANY	TRC	Spain
UNIVERSITA DEGLI STUDI DELL'AQUILA	UNIVAQ	Italy
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO	ISEP	Portugal
UNIVERSITA DEGLI STUDI DI GENOVA	UNIGE	Italy
CAMEA, spol. s r.o.	CAMEA	Czech
IKERLAN S. COOP	IKER	Spain
CAF Signalling S.L	CAF	Spain
R G B MEDICAL DEVICES SA	RGB	Spain
UNIVERSIDADE DE COIMBRA	COIMBRA	Portugal
VYSOKE UCENI TECHNICKE V BRNE - BRNO UNIVERSITY OF TECHNOLOGY	BUT	Czech
ROBOAUTO S.R.O.	ROBO	Czech
ESKISEHIR OSMANGAZI UNIVERSITESI	ESOGU	Turkey
KUNGLIGA TEKNISKA HOEGSKOLAN	KTH	Sweden
STATENS VAG- OCH TRANSPORTFORSKNINGSINSTITUT	VTI	Sweden
UNIVERSIDAD DE CASTILLA - LA MANCHA	UCLM	Spain
FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V.	FRAUNHOFER	Germany
SIEMENS AKTIENGESELLSCHAFT OESTERREICH	SIEMENS	Austria
RULEX INNOVATION LABS SRL	RULEX	Italy
NXP SEMICONDUCTORS GERMANY GMBH	NXP-DE	Germany
PUMACY TECHNOLOGIES AG	PUMACY	Germany
UNITED TECHNOLOGIES RESEARCH CENTRE IRELAND, LIMITED	UTRCI	Ireland
NATIONAL UNIVERSITY OF IRELAND MAYNOOTH	NUIM	Ireland
INOVASYON MUHENDISLIK TEKNOLOJI GELISTIRME DANISMANLIK SANAYI VE TICARET LIMITED SIRKETI	IMTGD	Turkey
ERGUNLER INSAAT PETROL URUNLERI OTOMOTIV TEKSTIL MADENCILIK SU URUNLER SANAYI VE TICARET LIMITED STI.	ERARGE	Turkey
OTOKAR OTOMOTIV VE SAVUNMA SANAYI AS - OTOKAR AS	OTOKAR	Turkey
TECHY BILISIM TEKNOLOJILERI DANISMANLIK SANAYI VE TICARET LIMITED SIRKETI - TECHY INFORMATION TECHNOLOGIESAND CONSULTANCY LIMITED COMPANY	TECHY	Turkey
ELECTROTECNICA ALAVESA SL	ALDAKIN	Spain
INTECS SOLUTIONS SPA	INTECS	Italy
LIEBERLIEBER SOFTWARE GMBH	LLSG	Austria
AIT AUSTRIAN INSTITUTE OF TECHNOLOGY GMBH	AIT	Austria
E.S.T.E. SRL	ESTE	Italy
NXP SEMICONDUCTORS FRANCE SAS	NXP-FR	France
BOMBARDIER TRANSPORTATION SWEDEN AB	BT	Sweden
QRTECH AKTIEBOLAG	QRTECH	Sweden
MONDRAGON GOI ESKOLA POLITEKNIKOA JOSE MARIA ARIZMENDIARRIETA S COOP	MGEP	Spain
INFOTIV AB	INFOTIV	Sweden
BERGE CONSULTING AB	BERGE	Sweden
CARDIOID TECHNOLOGIES LDA	CARDIOID	Portugal

Executive Summary

This deliverable describes the joint work on Task 5.2 and Task 5.3 of the VALU3S project. This document presents the evaluation of the improved Verification and Validation (V&V) processes and follows the qualitative and quantitative evaluation of the demonstrations with respect to the previously submitted planning and implementation (D5.2 [8] and D5.3 [9]). D5.6 presents potential limitations of the Framework and a final comparison of the performance baselines of the verification and validation (V&V) workflows and the achieved results by the end of the project.

The deliverable lists the demonstrators (planned in both Task 5.1 and Task 5.2) and their various aspects in use cases (UCs) that can be used to show progress at the final stage of the project. This is briefly introduced in Chapter 1. Individual demonstrators (corresponding to all use cases (UC)[1] introduced in D1.1 [1]) are then described in detail in Chapter 3. Chapter 2 presents the overview of the evaluation methodology whereas Chapter 4 gives an overview of the impact by considering the PESTLEE (Political, Economic, Social, Technological, Legal, Ethical, and Environmental) criteria. Finally, Chapter 5 concludes the report.

The deliverable focuses on both quantitative and qualitative assessment of the project results observed in demonstrators and use cases. For each use case, the scope is revisited and also, individual V&V challenges (based e.g., on evaluation scenarios) are re-elicited with respect to the demonstrators. Quantitative evaluation is one of the main contributions of this deliverable where measurable and objective results are presented as a result of mathematical and statistical modelling, experimental analysis and in terms of a numerical value. On the other hand, qualitative evaluation results, based on subjective judgment to analyse the value or prospects based on non-quantifiable information, are presented for each use case and the demonstrator. For each partner and the evaluation scenario, an overview of contributions to the selected evaluation scenarios per UC have been updated (by relying on the previous WP5 deliverables). There exist repeated tools and methods planned to be used for the evaluation which is crucial to reduce the workload and improve the collaboration among partners. Moreover, individual evaluation criteria are listed with updates (as compared to baseline) where applicable and/or possible.

In the VALU3S project, 28 evaluation criteria have been defined for the measurement of V&V process improvement and for the measurement of quality attributes of developed demonstrators, in particular their safety, cybersecurity, and privacy (SCP). Finally, potential demonstration setups are described per use case.

Part of the content of this deliverable has been taken from the reports already submitted earlier:

- D1.1 – Description of use cases as well as scenarios [1]

[1] Except for UC12 since the UC provider terminated its participation in the project and UC14 was added by the lead partner CARDIOID.

- D1.2 – SCP requirements as well as identified test cases [2]
- D3.5 – Interim description of methods designed to improve the V&V process [5]
- D3.6 – Final description of methods designed to improve the V&V process [6]
- D4.4 – Initial Detailed Description of Improved Process Workflows [3]
- D4.5 – Initial Implementation of V&V Tools Suitable for the Improved Process Workflows [4]
- D5.1 – Initial Demonstration Plan and a List of Evaluation Criteria [7]
- D5.2 – Final demonstration plan and a list of evaluation criteria [8]
- D5.3 – Initial Demonstrator Implementation Status Report [9]
- D5.4 – Demonstrator prototypes [10]
- D5.5 – Final demonstrator implementation status report [11]

Outcomes of this deliverable are used in other deliverables of Task 5.3 at the end of the project.

Contributors

Lukáš Maršík	CAMEA	Niyazi Uğur	ERARGE
Hamid Ebadi	INFOTIV	Çağrı Terzibaş	ERARGE
Thanh Bui	RISE	İbrahim Arif	ERARGE
Jack Jensen	BERGE	Deborah Hugon	STAM
Bernhard Fischer	SIEMENS	Aleš Smrčka	BUT
Jose Luis de la Vara	UCLM	Martin Karsberg	INFOTIV
Arturo García	UCLM	Joakim Rosell	RISE
Giovanni Giachetti	UCLM	Marie Farrell	NUIM
Sina Borrami	ALSTOM	Matt Luckcuck	NUIM
Håkan Palm	ALSTOM	Rosemary Monahan	NUIM
Emanuele Mingozzi	ESTE	Oisin Sheridan	NUIM
Katia Di Blasio	INTECS	Gürol Çokünlü	OTOKAR
Stefano Tonetta	FBK	Ömer Şahabaş	OTOKAR
Massimo Nazaria	FBK	Muhammet Saral	OTOKAR
Alberto Tacchella	FBK	Beata Davidova	ROBO
Metin Ozkan	ESOGU	Ugur Yayan	IMTGD
Ahmet Yazıcı	ESOGU	Alim Kerem Erdogmus	IMTGD
Elif Değirmenci	ESOGU	Cem Baglum	IMTGD
Yunus Sabri Kırca	ESOGU	Lourenço Rodrigues	CARDIO
Fabio Patrone	UNIGE	Walter Tiberti	UNIVAQ
Giovanni Gaggero	UNIGE	Luigi Pomante	UNIVAQ
Georgios Giantamidis	UTRCI (Collins)	Francesco Smarra	UNIVAQ
Stylianos Basagiannis	UTRCI (Collins)	Robert Sicher	LLSG
Thomas Bauer	FRAUNHOFER IESE	Alessandro D’Innocenzo	UNIVAQ
Xabier Mendialdua	IKER	Davide Ottonello	STAM
Íñigo Elguea	ALDAKIN	Maytheewat Aramrattana	VTI
Krasen Parvanov	QRTECH	Mikel Aldalur	IKER
Juan Manuel Morote	UCLM	Stylianos Basagiannis	UTRCI
José Proença	ISEP	Nestor Arana	MGEP
Matt Luckcuck	NUIM	Peter Folkesson	RISE
Marie Farrell	NUIM	Bernd Bredehorst	PUMACY
Rosemary Monahan	NUIM	Zain Shawar	PUMACY
Oisín Sheridan	NUIM	Christoph Schmittner	AIT
Alper Kanak	ERARGE	Mateen Malik	RISE
Sercan Tanrıseven	ERARGE	Davide Ottonello	STAM
Salih Ergün	ERARGE	Luis Alonso	TRC
Peter Folkesson	RISE		

Reviewers

Ashfaq Farooqui	RISE	2023-05-09
Behrooz Sangchoolie	RISE	2023-05-12, 2023-05-23, 2023-05-26
Aleš Smrčka	BUT	2023-05-03
Lukáš Maršík	CAMEA	2023-05-03
Manuel Schmidt	NXP-DE	2023-05-03

Revision History

Version	Date	Author (Affiliation)	Comment
0.1	2023-01-18	Alper Kanak (ERARGE)	Initial deliverable structure
0.2	2023-04-01	Salih Ergün, S. Halit Ergün, Alper Kanak (ERARGE)	PESTLEE analysis
1.0	2023-04-25	UC contributors	Use Case and demonstrator contributions
1.1	2023-04-28	İbrahim Arif (ERARGE)	Qualitative Assessment Results
1.2	2023-05-02	Alper Kanak (ERARGE)	Ready for the first round of reviews
1.3	2023-05-10	Alper Kanak (ERARGE)	Sent to the coordinator for the final check
1.4	2023-05-12	Behrooz Sangchoolie (RISE)	Review of the first final draft, while making formatting changes and leaving additional comments to be addressed.
1.5	2023-05-16	Alper Kanak (ERARGE)	Finalisation of all review and corrections. Submitted to the coordinator.
1.6	2023-05-18	Alper Kanak (ERARGE)	Reviews and updates received from all partners, processed and readied for submission.
1.7	2023-05-23	Behrooz Sangchoolie (RISE)	Review of the second final draft, while making formatting changes.
1.8	2023-05-26	Behrooz Sangchoolie (RISE)	Final version of the report to be submitted.

Table of Contents

List of Figures

List of Tables

Acronyms

ADAS	Advanced Driver-Assistance System
AHRS	Attitude and Heading Reference System
AI	Artificial Intelligence
AMS	Analog Mixed Signal
CAD	Computer-Aided Design
CCF	Common Cause of Failure
CI	Continuous Integration
CIS	Computer Interlocking System
CV	Computer Vision
DDMA	Doppler Division Multiplexing Access
DSL	Domain Specific Language
ECC	Error Correction Code
ECG	Electrocardiogram
ECU	Electronic Control Unit
EER	Equal Error Rate
FAR	False Acceptance Rate
FDD	Failure Detection Diagnosis
FLA	Failure Logic Analysis
FMEA	Fault Mode and Effects Analysis
FMECA	Failure Mode, Effects and Criticality Analysis
FN	False Negative
FP	False Positive
FPGA	Field-Programmable Gate Array
FRR	False Rejection Rate
FTA	Fault Tree Analysis
FTTI	Fault Tolerant Time Interval
HiL	Hardware-In-the-Loop
HMI	Human Machine Interface
HRI	Human-Robot-Interaction
HRV	Heart Rate Variability
HSM	Hardware Security Module
IC	Integrated Circuit
IMU	Inertial Measurement Unit
IoU	Intersection over Union

IP	Intellectual Property
IPC	Industrial PC
KSS	Karolinska Sleepiness Scale
mAP	mean Average Precision
ML	Machine Learning
MMC	Matthews Correlation Coefficient
MQTT	Message Queuing Telemetry Transport
OPC-UA	Open Platform Communications Unified Architecture
PESTLEE	Political, Economic, Social, Technological, Legal, Environmental and Ethical
PIR	Passive InfraRed
PLC	Product Life-Cycle
QEMU	Quick Emulator
RAMS	Reliability, Availability, Maintainability and Safety
SCP	Safety, Cybersecurity, and Privacy
SIL	Safety Integrity Level
STB	System Test Box
STRIDE	Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege
SoC	System on a Chip
TAR	True Acceptance Rate
TCU	Traction Control Unit
THR	Tolerable Hazard Rate
TP	True Positive
TRL	Technology Readiness Level
TN	True Negative
UC	Use Case
UDP	User Datagram Protocol
UWB	Ultra-wideband
V&V	Verification and Validation
VR	Virtual Reality
WP	Work Package

Chapter 1 Introduction

The main objective of the VALU3S project is to lower the effort and cost of engineering processes by focusing on one (or more) of the most resource-consuming steps of the product life cycle – verification and validation (V&V). V&V is not just a single engineering phase, but a complex process integrated into different engineering phases, applied in different levels of details of development, begins before even a single line of code is produced and does not end after a product is deployed to the market. The VALU3S project aims at V&V of automated systems, which require special approaches in providing confirmation of services and warranties which are different from the traditional techniques. Within the project, a V&V framework has been developed, which will integrate newly proposed and/or improved versions of already existing V&V methods and tools supporting these methods. The framework has been applied in the development phase of products in different domains (agriculture, aerospace, automotive, healthcare, industrial robotics, and railway) to show the improvements gained by the framework. The V&V process is being improved not just by reducing the effort and cost but also by increasing the quality of products while reducing the time needed for V&V. This main result has been demonstrated within intermediate demonstrations and finalised at the end of the project by providing an evaluation report for all use cases and by demonstrating the utilisation of newly developed methods and tools in these use cases.

The purpose of this document is to present the qualitative and quantitative evaluation results associated with the selected use cases and demonstrations. The demonstration consists of several so-called *demonstrators,* which are prepared by individual UC providers. For the final demonstration, crucial parts have been selected from all the use cases to cover the majority of the domains, the dimensions, and the layers of the V&V framework, and the entire set of developed or improved methods and tools. Since the VALU3S consortium consists of experts in different fields of V&V, demonstrators are in the form of collaborative work of all partners led by use case providers. The demonstrations are supported by the evaluation reports which present the level of quality of the developed systems that have been reached and how much time and cost are required for V&V processes that have been reduced. To provide a credible evaluation, several metrics are used, focusing on measuring both SCP features and cost, effort, and quality of the V&V processes used in engineering processes in different use cases. For the qualitative assessment, experts' opinions have been collected by online questionnaires to observe the subjective factors like perceived usefulness, perceived ease of use, perceived trust, attitude toward using, behavioural intention to use, motivation, compatibility, return of investment expectancy, performance expectancy, perceived risk, and social influence.

1.1 Structure of the Deliverable

This section contains a summary of the information specific to the points that will be discussed in the following chapters, following the "Introduction" in Chapter 1.

Chapter 2 presents an overview of the evaluation methodology followed for quantitative and qualitative assessment in a nutshell.

Chapter 3 iterates over the individual demonstrators, including their brief introduction, description of the evaluation criteria, baseline evaluation, quantitative and qualitative evaluation results and observed limitations, lessons learnt and best practices. measurements of SCP and V&V criteria.

Chapter 4 presents an overall impact analysis based on Political, Economic, Social, Technological, Legal, Environmental and Ethical (PESTLEE) criteria.

Chapter 5 presents the conclusion of this deliverable, focusing on summarising the main achievements and potential future activities.

Appendix A and Appendix B present the questionnaire details and snapshots from the online questionnaire forms.

Chapter 2 Evaluation Methodology

The evaluation of the presented tools and toolchains is implemented mainly in two dimensions: quantitative assessment and qualitative assessment. According to NIST definitions [12], quantitative assessment is based on the use of a set of methods, principles, or rules for assessing performance based on the use of numbers where the meanings and proportionality of values are maintained inside and outside the context of the assessment. On the other hand, qualitative assessment is based on the use of a set of methods, principles, or rules for assessing risk based on nonnumerical categories or levels, subjective measures such as willingness, perceived usefulness, technology acceptance, etc. The following subsections give an overview of the quantitative and qualitative assessment methodology applied in use cases. Additionally, Section 2.1 presents the mapping studies to identify the commonalities among VALU3S use cases.

2.1 Quantitative Assessment Methodology and Mapping Strategy and Results

For quantitative assessment, VALU3S partners have defined different quality metrics which describe what is to be measured specifically for presenting useful indicators and recommendations for decision-makers. The quantitative assessment is based on objective criteria that present numbers-based, countable, or measurable identifiers. Quantitative assessment is use-case-specific and tells measurable observations about developed products and processes behind the development. These metrics have already been documented and adopted by different standards but, in most cases, they do not take into account the applicability of the selected metrics. VALU3S partners have analysed 13 different real-world use cases based on the sample cyber-physical systems that are actually being used in six domains of application: automotive, railway, aerospace, agriculture, healthcare, and industrial robotics. These cyber-physical systems are automated and partly include AI-based decision-making systems, which makes safety and cybersecurity their primary concern.

The quantitative assessment methodology is two-fold: i) V&V Evaluation Criteria (Eval_VV series presented in Table 2-1); ii) SCP Evaluation Criteria (Eval_SCP series presented in Table 2-2). The methodology is applied by the use case and demonstrator leaders as they determine how quantitative evaluation is implemented (e.g., by experiments, analyses, tests, etc.) to measure the quality of development and V&V activities. Partners also aligned these metrics with the literature and existing standards and searched for commonalities among the domains. The refinement process of the metrics resulted in the selection of criteria which provide means for practical measurement of the development of cyber-physical systems. The quantitative assessment is presented for each use case and underlying demonstrator in Chapter 3.

Table 2-1. V&V Evaluation Criteria

V&V Evaluation Criteria Identifier according to D5.2 [8]	V&V Evaluation Criteria Name	Criteria's Description
Eval_VV_1	Time of test execution	This criterion will show if and how a new test set will be optimized w.r.t. used methods, improved tools, and available resources. The criterion measures and compares test execution time.
Eval_VV_2	Coverage of test set	Measuring how much of software/hardware test coverage items (e.g., lines of code, branches, faults, and attacks depending on selected test design technique) has been covered by a test set (set of test cases, also known as test suite). Increased coverage means increased trust in the analysed system.
Eval_VV_3	Number of test cases	Quantify a test set proving that a reduced number of test cases can ensure desired quality (coverage), e.g., in combination with erroneous outputs a.k.a. silent data corruption (SDC), number of safety/security requirement violations, and number of malicious attacks and faults detected.
Eval_VV_4	Effort for test creation	Estimation of effort for deriving/maintaining test suites, e.g., for fault injection and runtime verification campaigns (manual design vs. model-based generation).
Eval_VV_5	Joint management of SCP requirements	To minimize risks and costs, the potential impact of SCP requirements on the design must be analysed early with management flow for joint SCP requirements analysis. Examples of approaches considering safety and security are SAHARA and FMVEA, as well as an approach for co-engineering with interaction points that considers safety, cybersecurity, and performance.
Eval_VV_6	Cost of finding and fixing a coding bug	The cost of someone writing a test and the cost of someone finding and fixing the bug which the test exhibits.
Eval_VV_7	Development quality statistics	Preferably the statistics should be evaluated after every software change or at scheduled times, e.g., periodically every night. Evaluation might be connected with running regression tests and using the results as part of the statistics.
Eval_VV_8	Effort needed for test	This evaluation criterion is used to measure the effort (e.g., person-hours) required to perform a test on a system. This measure is especially useful to compare the effort spent doing manual work versus automated work.
Eval_VV_9	Service actions needed	The number and complexity of service actions needed after deployment of the system to the field. This can include the first installation of the system, on-site debugging and tuning of the configuration during the first weeks of operation, and regular service checks or demanded actions.
Eval_VV_10	Reduced cost and time for work on the certification process and functional safety	Successful certification of industrial robots, robotic systems, and control systems demands compliance with all applicable technical guidelines and standards in the specific application scenario or environment. Therefore, a system audit must cover and pass several costly and time-consuming aspects before certification can take place.
Eval_VV_11	Randomness and security assessment process performance	Assessing the randomness and cryptographic strength (Eval_SCP_9) should be time- and effort-efficient as the cyber-physical systems to be validated and verified are complex systems and needs to be restarted as soon as possible for their actual work. Moreover, the employment of less personnel effort is also crucial to improve labour efficiency as the proposed method will enable the verification of the crypto-key generation scheme by-design that will lead the overall design to a more resilient system.

V&V Evaluation Criteria Identifier according to D5.2 [8]	V&V Evaluation Criteria Name	Criteria's Description
Eval_VV_12	Effort required to the user for prepare and run the tool	This evaluation criterion is focused on the tools which require effort from the user. In fact, each type of tool can require a different amount of effort from the user to allow the proper set up and running of the application.
Eval_VV_13	Reliability measures of decisions	Verification and validation automatic checkers are considered decision-making systems which classify the designs, implementation, or behaviour of a system under test (SUT) if it fulfils the specified requirements. The criterion measures true/false positives/negatives ratios as these directly influence the time & cost of debugging the SUT.

Table 2-2. SCP Evaluation Criteria

SCP Evaluation Criteria Identifier according to D5.2 [8]	**SCP Evaluation Criteria Name**	**Requirement Type**	**Criteria's Description**
Eval_SCP_1	Error Coverage	Safety & Cybersecurity	Error coverage is defined as the conditional probability that a system recovers, given the occurrence of a fault. Similar to other metrics such as program vulnerability factor (PVF), error coverage does not distinguish between different failure modes. However, in practice, silent data corruptions (SDCs) are considered the most severe failure mode, because users will trust the system output in the absence of an error indication. This is because the erroneous outputs are generated with no indication of failure, making them very difficult to detect. Therefore, instead of error coverage, some researchers have used error resiliency as the dependability metric.
Eval_SCP_2	Number of Safety /Security Requirement Violations	Safety & Cybersecurity & Privacy	Measuring the number of violated SCP attributes/requirements/properties that have been checked by runtime monitors, software testing, and/or formal verification is useful for comparing the effect of changes to requirements engineering, development, and verification processes. Safety/security requirement violations may indicate inconsistencies amongst requirements if any do exist.
Eval_SCP_3	Number of Malicious Attacks and Faults Detected	Safety & Cybersecurity	This evaluation criterion measures the number of malicious attacks and faults detected in the system-under-test (SUT). The number of detected attacks and faults shall then be compared with the actual number of malicious attacks and faults that have occurred or injected into the SUT, in order to reflect on the safety and security aspects of the SUT. The SUT would be considered safe if all the faults are correctly detected. Similarly, the SUT would be considered secure if all the attacks are correctly detected.
Eval_SCP_4	Metrics to Evaluate AI/ML Algorithms	Safety	While data preparation and training a machine learning model is a key step in the machine learning pipeline, it's equally important to measure the performance of the trained model. It is significant to use multiple evaluation

SCP Evaluation Criteria Identifier according to D5.2 [8]	SCP Evaluation Criteria Name	Requirement Type	Criteria's Description
			metrics to evaluate the model; the metrics include classification metrics, regression metrics, ranking metrics, and computer vision metrics.
Eval_SCP_5	Potential Impact of Incidents and Attacks	Safety & Cybersecurity	Unpleasant consequences of performed attacks and induced incidents can be various and range from service disruptions to damage to the attacked systems and even harm to surrounding objects and people. Defining different levels allows classifying the impact of incidents and attacks depending on their malicious effects on the attacked system and on the actions that are required to make the system operative again.
Eval_SCP_6	Metrics to Evaluate Cybersecurity	Cybersecurity	The development of secure systems needs to be validated using relevant metrics to judge the quality of cybersecurity, so that over several development iterations, the number of, for example, identified threats, becomes fewer. The metrics include, among others, the number of threats and the number of attack paths.
Eval_SCP_7	Number of prevented accidents	Safety	This criterion aims at situations which could lead to an accident related to the safety of usage of a system. There are two approaches to measuring the number of prevented accidents. One approach is to analyse a model of a system and provide a detailed report of conditions of possible accidents. The other approach is to experimentally evaluate the system and report each of the accidents.
Eval_SCP_8	Authentication Accuracy and Time Applied to Human Users and Components	Cybersecurity & Privacy	This criterion focuses on two main subcomponents of cybersecurity: (i) Active authentication of system components at certain time intervals to verify that each component is not under attack, (ii) Role-based access module for user authentication against unprivileged uses.
Eval_SCP_9	Randomness and cryptographic algorithm strength	Cybersecurity	Ensuring the entire security covering end nodes and central mechanisms through highly secure cryptographic backends.
Eval_SCP_10	Software fault tolerance robustness	Safety	An error is detected if its presence is indicated by an error message or error signal; errors that are present but not detected are latent errors. A fault is the adjudged or hypothesized cause of an error. Consequently, the software fault tolerance robustness is demonstrated by different testing methods, including fault injection to ensure error-tolerant use case activity.
Eval_SCP_11	Simulation-level System Robustness	Safety	Testing on a simulation-level of the system under test with a fault injection plug-in for system robustness assessment. Similar to the Software fault tolerance robustness criterion, the assessment is not restricted to software and instead can potentially include, e.g., model-in-the-loop or hardware-in-the-loop components.
Eval_SCP_12	Number of Attack/ Incident	Safety	This metric expresses how many attack/incident typologies the method/tool is capable to deal with, giving a useful indication of its functionalities and, indirectly, its level of detail.

SCP Evaluation Criteria Identifier according to D5.2 [8]	SCP Evaluation Criteria Name	Requirement Type	Criteria's Description
	Typologies Examined		
Eval_SCP_13	Accuracy of Simulated Sensor Output	Safety	The accuracy of simulated sensor output generated by a simulation environment can be compared with real sensor data from a controlled and virtually replicated environment to verify the simulator output. This relates to the safety concepts, avoiding live testing with real personal and vehicles.
Eval_SCP_14	Simulator Environment Quality	Safety	Metrics to track simulator visual quality, and model resolution. Photo-realistic 3D rendering measures such as polygon count, ray-traced distance, virtual sensor resolution, etc.
Eval_SCP_15	Simulator Environment Functionality	Safety	Metrics on features and functionality supported by the simulator environment. A mature simulator gives flexibility to the operators to configure aspects of the environment and extract data from the environment using a variety of sensors. The expected increased functionality from the simulator will gain more opportunities to simulate complex and possible harmful events while completely eliminating the risk of injuries.

2.2 Qualitative Assessment Methodology and Technology Acceptance Model

For the qualitative evaluation of VALU3S outcomes, Technology Acceptance Model (TAM) [13] is extended to orient the organisational investments for better planning, decision-making, design and implementation of the automated cyber-physical system solutions in six VALU3S domains. The proposed model, namely the Qualitative Assessment Model (QAM) is inspired by TAM that effectively utilises the subjective measures of TAM to evaluate the behavioural intention of end-users (see Figure 2-1). QAM postulates that the behavioural intention of an individual towards using a cyber-physical system is influenced by practical or psychological factors like user interfaces, enrolment or verification procedures, data security policies, devices, and other auxiliary tools. The related factors are measured by questionnaires or usage statistics automatically collected by system components. The applied methodology is based on the quantitative analysis results for better convincing the subjects.

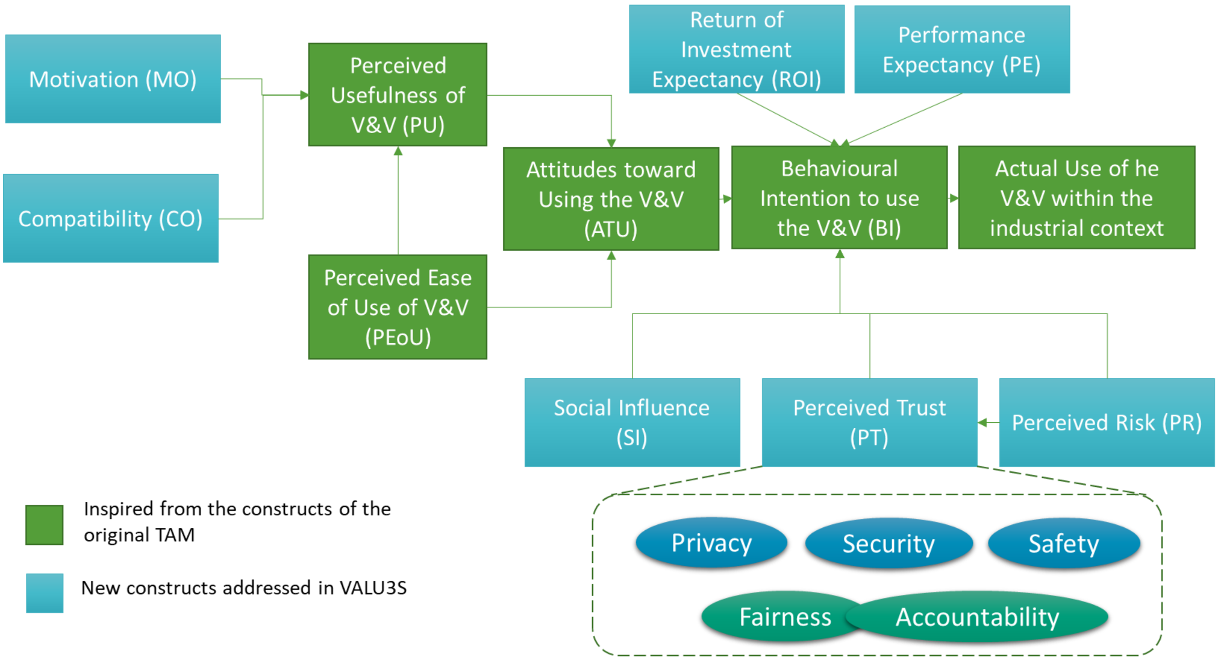

Motivation (MO)
Compatibility (CO)
Perceived Usefulness of V&V (PU)
Perceived Ease of Use of V&V (PEoU)
Attitudes toward Using the V&V (ATU)
Return of Investment Expectancy (ROI)
Performance Expectancy (PE)
Behavioural Intention to use the V&V (BI)
Actual Use of he V&V within the industrial context
Social Influence (SI)
Perceived Trust (PT)
Perceived Risk (PR)
Privacy
Security
Safety
Fairness
Accountability
Inspired from the constructs of the original TAM
New constructs addressed in VALU3S

Figure 2-1 QAM constructs

QAM is composed of a set of constructs aiming to cover the crucial aspects of a typical digital cyber-physical system designed for the VALU3S domain applications. QAM constructs are defined first by inspiration from the widely-accepted technology acceptance models like TAM or UTAUT (Unified Theory of Acceptance and Use of Technology – UTAUT) [14] and second by experience from the required engineering activities and later during demonstrations in VALU3S (e.g., expert opinions, discussions, literature surveys, etc.). QAM constructs are listed below:

- **Perceived usefulness (PU)**: PU is defined as the prospective user's subjective probability that using a VALU3S solution where the operations facing users are designed so that a user believes that related tasks facilitate the processes.
- **Perceived Ease of Use (PEOU)**: PEOU is defined as the degree to which a person believes that using VALU3S solutions would be free from effort.
- **Attitude Toward Using (ATU)**: Attitude toward system use is postulated to partially mediate the effect of perceived ease of use and perceived usefulness on behavioural intention. The attitude can be at a personal level, where the user herself/himself may have a personal view toward using the system. If the organisation has an organisational culture that shapes this attitude, then the attitude can be observed at an organisational level.
- **Behavioural Intention to use (BI) and Actual use (A)**: Positing particular beliefs, PEOU, PU and ATU, reflect the primary relevance for technology acceptance behaviours and determines the behavioural intention of people (BI) to use the system (Actual Use).
- **Motivation (MO)**: This construct is defined to assess the motivation of users for a better understanding of whether the proposed solutions excite and appreciate them to realise in their organisations.
- **Compatibility (CO)**: Compatibility is a technical requirement that may influence the decision of buying a V&V solution and integrate it into the existing system. This construct is used to

measure the willingness of paying some effort to integrate the offered innovation into the mainstream systems, e.g., IT networks or production lines.

- **Return Of Investment Expectancy (ROI)**: This construct is designed to measure the financial impact of deploying the V&V tools in existing settings. ROI can be measured in many ways but here a subjective estimate is expected to understand if the deployment will reduce the costs and create an advantage in the market.
- **Performance Expectancy (PE)**: Performance can also be measured in many ways, as the quantitative analyses are expected to present a detailed assessment. PE is designed to observe the subjects' opinions about how the offered innovations improve the quality of products, services or outputs, shorten the time and reduce faults, or increase efficiency.
- **Perceived Trust (PT)**: PT indicates whether a person perceives that a particular technological solution offered in VALU3S is secure, safe, accountable, privacy-preserving, fair and trustworthy or not.
- **Perceived Risk (PR)**: PR is about the opinions related to the level of risk that deals with any potentially complicated problem, investment plans, compliance with standards and regulations, and security, privacy and safety concerns.
- **Social Influence (SI)**: SI is defined as the degree to which an individual perceives the importance of others' beliefs, which he/she should use to accept the VALU3S solution stack.

QAM is applied by a questionnaire given in Appendix A. A set of questions are asked to subjects and they are expected to interpret the questions in two dimensions: i) personal opinions regardless of any organisational culture related to their affiliated institutions; ii) organisational opinions that reflect the overall glance of their affiliated organisations, i.e., organisational culture, strategic vision and mission and existing or future provisions and practices.

The questionnaire is applied to three different types of profiles selected from volunteered experts between the ages of 18 and 65 having no disabilities. Three different profiles of the subjects are as follows:

- Internal subjects who have taken a role in the design, implementation and testing of the tools and/or toolchains used in the selected use case(s) of VALU3S.
- External subjects who have not taken a direct role in developing the tools and preparation of the selected demonstrators, but are employed in one of the VALU3S participant organisations
- External subjects who are not involved either in VALU3S project activities or employed in one of the VALU3S participant organisations.

The questionnaire is applied over an online survey (see Appendix B for snapshots) where subjects can find all relevant information about the use cases and demonstrators. Especially for external subjects, VALU3S partners employ at least one person to accompany the subjects and help them throughout the questionnaire. The questionnaire is applied in the following methodology:

- Step 1: A short description of the VALU3S, project scope, main objectives and targeted area(s) of implementation are presented.
- Step 2: A short description of the main objective of the questionnaire is presented.

- Step3: Videos, presentations, posters or any descriptive material about the use cases and demonstrators are presented to each subject.
- Step 4: A summary of the quantitative analysis results within the context of the targeted area of implementation is presented (to strengthen the participants' attraction and make them more convinced).
- Step 5: To prevent any GDPR-related issues a disclaimer is prepared to inform the subjects about the process. Information about the disclaimer is shared and a question & answering session is applied if subjects have any concerns about the process. If a group of participants is available, the first 4 steps are applied to the entire group by letting them speak freely. This open-speech approach helps resolve any problem or question mark in participants' minds.
- Step 6: Validation of the subject's consent is assured.
- Step 7: Application of the questionnaire per subject (not at group level to prevent any positive or negative influence on each other)
- Step 8: The desktop studies are applied by the WP5 leaders in an anonymised way and presented in this deliverable.
- Step 9: All participating subjects are informed about the results by sharing the D5.6 with each of them.

2.2.1 Statistical Evaluation and Presentation of Results

The subjects' responses are statistically analysed in the following way:

- Correlation Analysis: Correlations among the QAM constructs are presented in tabular format. The correlation coefficient (a value between -1 and +1) tells how strongly two variables are related to each other. A correlation coefficient of +1 indicates a perfect positive correlation. As variable X increases, variable Y increases. As variable X decreases, variable Y decreases. Similarly, a correlation coefficient of -1 indicates a perfect negative correlation. As variable X increases, variable Z decreases. As variable X decreases, variable Z increases.
- Reliability analysis: To measure the reliability of the questionnaire, Cronbach Alpha values are computed for each QAM construct. Cronbach's Alpha ranges between 0 and 1, with 0 being the lowest possible value and 1 being the highest. A value of 0 indicates no internal consistency or reliability in the questionnaire, while a value of 1 indicates perfect internal consistency. A score close to 1 indicates that the items in the questionnaire are highly correlated with each other and provide a consistent measure of the underlying construct being assessed. In contrast, a score close to 0 indicates that the items are not well-correlated and do not provide a consistent measure of the construct. For an easy interpretation, Cronbach alpha values greater than 0.7 mean the internal consistency is acceptable whereas values less than 0.5 mean unacceptable as these constructs are not reliable in their context.
- Regression Analysis: In statistical modelling, regression analysis is used to estimate the relationships between two or more variables: i) a dependent variable (a.k.a. criterion variable) is the main factor you are trying to understand and predict, and ii) independent variables (a.k.a. explanatory variables, or predictors) are the factors that might influence the dependent variable. Regression analysis helps analysts understand how the dependent variable changes when one

of the independent variables varies and allows them to mathematically determine which of those variables has an impact. The results are given in the following ways:

- Multiple R: It is the Correlation Coefficient that measures the strength of a linear relationship between two variables. The correlation coefficient can be any value between -1 and 1, and its absolute value indicates the relationship strength. The larger the absolute value, the stronger the relationship: 1 means a strong positive relationship; -1 means a strong negative relationship; 0 means no relationship at all.
- R Square: It is the Coefficient of Determination, which is used as an indicator of the goodness of fit. It shows how many points fall on the regression line. The R2 value is calculated from the total sum of squares, more precisely, it is the sum of the squared deviations of the original data from the mean. Generally, R Square of 95% or more is considered a good fit.
- Standard Error: It is another goodness-of-fit measure that shows the precision of your regression analysis - the smaller the number, the more certain you can be about your regression equation. While R2 represents the percentage of the dependent variables' variance that is explained by the model, Standard Error is an absolute measure that shows the average distance that the data points fall from the regression line.
- P-value: The P-value is used to determine the probability of the results of hypothetical tests. One can analyse the results based on 2 hypotheses: the Null hypothesis and the Alternative hypothesis. If the P-value is >0.10, then the data is not significant. If the P-value is <=0.10, then the data is marginally effective. If the P-value is <=0.05, then the data is significant, and if the P-value is <0.05, then the data is highly important.
- t-Stat: The t-value measures the size of the difference relative to the variation in your sample data. Put another way, T is simply the calculated difference represented in units of standard error. The greater the magnitude of T, the greater the evidence against the null hypothesis.

The results are also presented in the VALU3S online web repository for each use case and the selected demonstrators. One can find brief information about the participants' profile, mean and standard deviations of subject responses to questions, and the results of the hypotheses testing as presented in Figure 2-2.

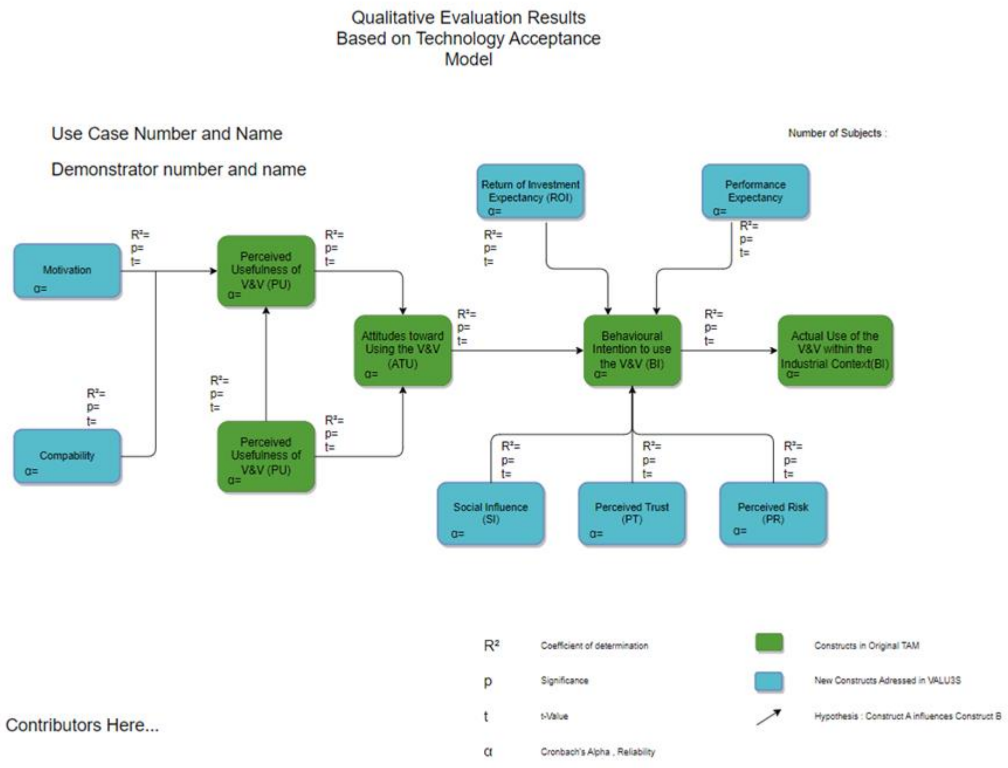

Qualitative Evaluation Results
Based on Technology Acceptance
Model
Use Case Number and Name
Demonstrator number and name
Number of Subjects :
Motivation
α=
Compability
α=
Perceived Usefulness of V&V (PU)
α=
Perceived Usefulness of V&V (PU)
α=
Attitudes toward Using the V&V (ATU)
α=
Return of Investment Expectancy (ROI)
α=
Performance Expectancy
α=
Behavioural Intention to use the V&V (BI)
α=
Actual Use of the V&V within the Industrial Context(BI)
α=
Social Influence (SI)
α=
Perceived Trust (PT)
α=
Perceived Risk (PR)
α=
R²=
p=
t=
R² Coefficient of determination
p Significance
t t-Value
α Cronbach's Alpha , Reliability
🟩 Constructs in Original TAM
🟦 New Constructs Adressed in VALU3S
↗ Hypothesis : Construct A influences Construct B
Contributors Here...

Figure 2-2 Mock-up design of the presentation of QAM results over the VALU3S web repository.

Chapter 3 Evaluation Results

Demonstrating the usefulness of the framework developed within the VALU3S project relies on (i) demonstrators and (ii) the evaluation report. Demonstrators are key subjects that relate the VALU3S objectives to demands on real engineering processes. The evaluation report will provide plausible justification for the improvement of the V&V processes and the quality of developed products. The improvements are simply presented by comparing quantifiable (e.g., effort, cost & time) and other quality attributes of V&V performed in all the use cases with and without the V&V framework developed in VALU3S. In this chapter, we present individual demonstrators, including a baseline evaluation dealing with the quantitative and qualitative assessment and specific conclusions considering the observed limitations, lessons learnt and best practices. Note that some of the demonstrations are then classified as **Lead** demos to be presented during the project's final event, and the rest we call **complementary** (those support the project as a whole and demonstrated not in the final event but over other channels via online presentations, videos, etc.).

A properly-described status of each V&V process at the beginning of the project is provided to compare it with the status at the end of the project, and thus, demonstrate the development of V&V within the project. The initial status of the V&V process, i.e., the description of the used techniques and tools and available resources supporting V&V, is called the V&V *baseline*. Individual demonstrators that have been selected for final demonstrations and specification of evaluation of their baseline have the key role in the demonstrator description, which consists of the following elements:

1. A brief recapitulation of the corresponding UC[2],
2. V&V challenges,
3. List of contributors,
4. Contributors' roles and evaluation scenarios
5. Demonstration
6. Quantitative Results
7. Quantitative Results
8. Observed Limitations, Lessons Learnt and Best Practices

The initial definition of the demonstrators, the specification of their baselines and the evaluation criteria are covered in the first deliverable of WP5, i.e., D5.1 [7]. The demonstrator baseline is covered by the D5.2 [8] including the details about the evaluation criteria. As the initial state of all the demonstrations has been identified and described, we will continue by implementing the demonstrators. This process is partly documented in D5.3 [9]. An interim demonstration is supported by this report. The final demonstrations are documented, evaluated, and presented in D5.5 - *Final demonstrator implementation status report [11]*, D5.7 - *Updated web-based repository, linking V&V evaluation results to the framework*, and D5.8 - *Final demonstration* (see Table 3-1 and Figure 3-1 for a detailed plan).

[2] UC intro is briefed in order not to repeat the information already present in D5.3.

Table 3-1 Demonstration plan and reporting.

Step of Demonstration Plan	Reporting in Deliverables	Delivery Project Month	Note
1	D5.1 [7]	8	Definition of demonstration and a part of the specification of baseline.
2	D5.1 [7], D5.2 [8]	8, 18	Definition of evaluation criteria in D5.1, evaluation of baseline in D5.2.
3	D5.3 [9], D5.4 [10]	24, 30	Initial demonstration status report in D5.3, demonstrator prototypes in D5.4.
4	D5.5 [11], D5.6 (current report),	36	Demonstrator report, evaluation, and update of web-based framework linking the evaluation results.
5	D5.7, D5.8 (these deliverables will be submitted before the end of May 2023)	38	Final demonstration as well as inclusion of demonstrators in the web-based repository.

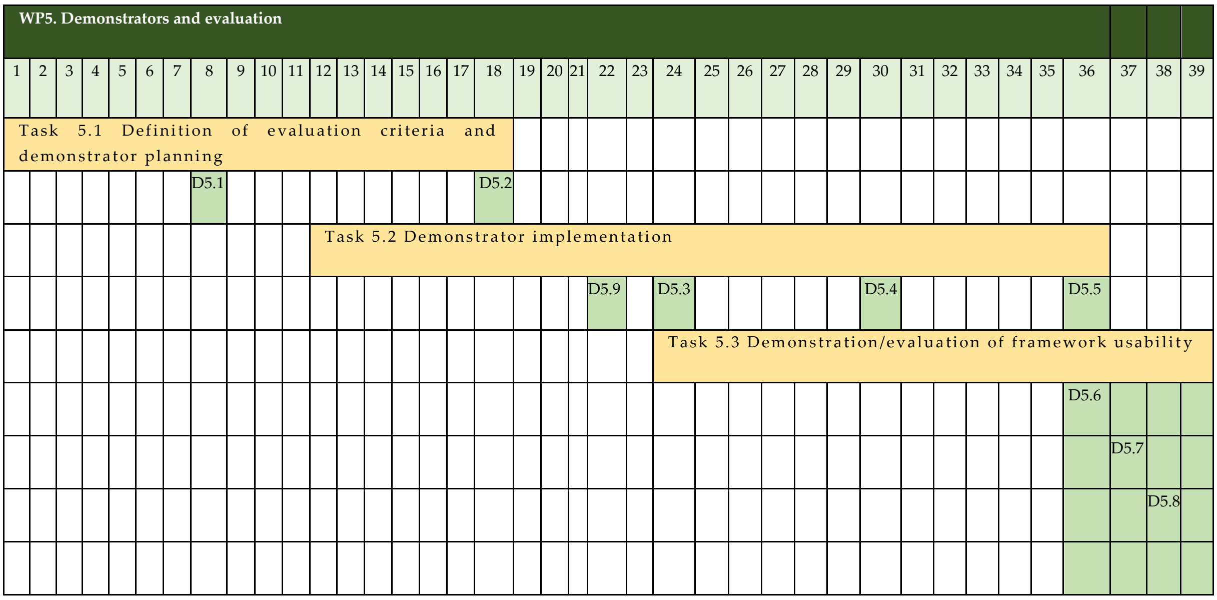


Figure 3-1 Timing of tasks in Work Package 5 and delivery timeline for different deliverables.

Each demonstrator has adopted a different set of evaluation criteria and selected metrics that could be used for a successful demonstration. These metrics need to be measured (or supported by qualified estimation), and corresponding quantified and qualified measurements are reported. To demonstrate the progress, we need to obtain the baseline measurable indicators (again measurements or estimations) which is a crucial part of this deliverable. Comparison with updated measurements (based on prototypes) has then been addressed further in D5.5 [11]. The rest of this chapter describes the individual demonstrators.

3.1 Use Case 1 - Intelligent Traffic Surveillance (UC1)

UC1 deals with the testing and verification of reliability and security of smart and mostly wireless sensors (cameras, radars, etc.) used in intelligent traffic surveillance. It works with traffic monitoring systems similar to the CAMEA's Unicam platform (used e.g., for Spot Speed Enforcement, Section Speed Enforcement, Travel Time, Red Light Enforcement, or Weigh-in-Motion system). An example of a complex UNICAM system can be seen in Figure 3-2.

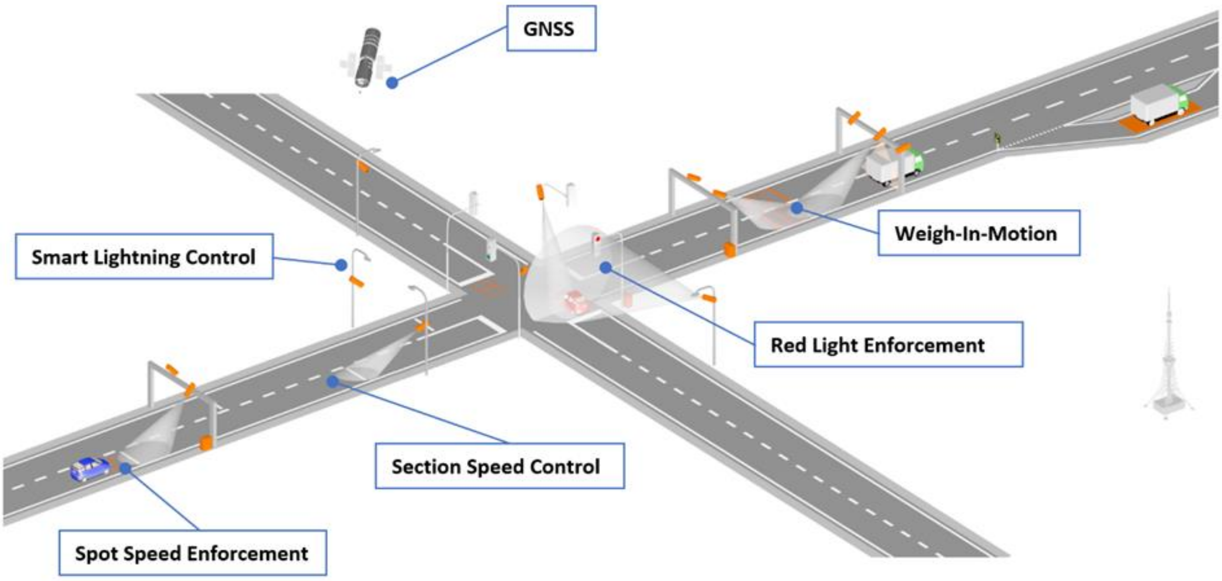


Figure 3-2 CAMEA Unicam intelligent traffic surveillance system

3.1.1 V&V challenges

Within this use case, some V&V challenges arise based on defined evaluation scenarios and the following corresponding baseline:

1. The Radar/camera-advanced detection and tracking system uses Machine Learning component(s), which is data-driven and has opaque nature.
2. Data for V&V of the system's performance and robustness are not feasible to capture only from real-world settings:
   - Privacy issues: The data includes camera images of vehicles with license plates (LPs) which are protected by GPDR.
   - Coverage and corner cases: Under representative data for corner cases such as vehicles going in wrong directions, or other misbehaviours. Recreation of such situations in real settings may danger human life.
   - Inflexibility in improving dataset distribution, e.g., if V&V strategy may result in a need to focus on similar cases of some problematic situations of interest.
3. Node connection to cloud – The system is operating in a cloud infrastructure with distributed processing. The communication between nodes is relying on TCP/UDP with proprietary communication and encryption components. CAMEA must ensure that there are no blind spots

in the information security processes, i.e., to have a visibility of the threat landscape, option caused by siloed processes and potentially inconsistent procedures.

3.1.2 Contributors

Partners contributing to the UC: CAMEA, BUT, RISE, BERGE, INFOTIV, QRTECH, AIT.

3.1.3 Contributors' Roles & Evaluation Scenarios

Evaluation scenarios defined for this use case are listed below; the assignments of the contribution of all UC partners can be found in Table 3-2:

1. VALU3S_WP1_Automotive_1 - Radar/camera advanced detection and tracking (covering potentially problematic situations, testing accuracy and reliability of detector algorithms, using simulation-based synthetic data)
2. VALU3S_WP1_Automotive_2 - Radar + camera cooperation (testing/validating of triggering camera based on the radar detection input, time criticality + detection reliability and accuracy, testing communication between sensors based on simulation-based synthetic data)
3. VALU3S_WP1_Automotive_3 - Node connection to the cloud (verifying reporting mechanism including connection to the server, detection result buffered for server query, testing proper functionality in the lab scenario under various simulated conditions)

Table 3-2 Overview of contribution to evaluation scenarios by UC1 partners

Evaluation Scenario	CAMEA	BUT	RISE	INFOTIV	BERGE	QRTECH	AIT
VALU3S_WP1_Automotive_1	X		X	X	X		
VALU3S_WP1_Automotive_2	X		X	X	X		
VALU3S_WP1_Automotive_3	X	X	X	X		X	X

3.1.4 Demonstrations

Demonstration for UC1 was planned by individual partners. Therefore, multiple demonstrable items are covering defined challenges and scenarios and partially cover VALU3S dimensions (see Table 3-3).

Table 3-3 Overview of demonstration prepared by UC1 partners.

Item #	Demonstration name	Description/Purpose	Format	Responsible
1	Traffic simulator-based vehicle detection	Show overall V&V processes of Intelligent Traffic Surveillance system by CAMEA (Machine learning based camera & radar detector) using BERGE simulator. Show the system functionalities (LP detection API) and settings of environment/scenario parameters to support V&V requirements.	Lead, presentation, video	RISE, BUT, AIT, QRTECH

Item #	Demonstration name	Description/Purpose	Format	Responsible
2	Testing network communication using NetLoiter	Show how to set up and run tests of the system performing under different network conditions. The remote radar processing application will be tested with NetLoiter tool.	Lead, presentation	BUT, CAMEA
3	Model-based threat analysis for CAMEA system architecture	Demo model-based threat analysis for CAMEA system architecture using ThreatGet. This shows the envisioned usage architecture of the Intelligent Traffic Surveillance system and identifies potential attack scenarios and the effects of different security measures on the overall threat landscape for the system. In addition, it shows how reports and documentation for V&V and compliance can be generated.	Complementary	AIT, QRTECH

3.1.5 Quantitative Results

List of evaluation criteria of the V&V process:

1. Eval_VV_2 – Coverage of Test Set (VALU3S_WP1_Automotive_1) – with the BERGE simulator, toolchain, and method, it is possible to verify the LP detection system using the synthetic data corresponding to various traffic scenes (3 scenes), illuminations, weather conditions, traffic participants, sensor specifications and mounting positions. This enables the coverage of rare but potentially critical traffic situations including tests of the accuracy of processing and detection algorithms, etc. (CAMEA).
    - Within the baseline: There had been a very limited set for testing traffic monitoring systems. Testing has been mostly done in real-time after deployment with almost no possibility of going back to problematic situations. We need to wait until it happens again in real-time and try to catch it by the whole system (or go with our vehicles and imitate it).
    - Improvement: Developed a framework and toolchain that support generating test scenarios with multiple input parameters including the scenes, the traffic scenarios, traffic participant trajectories (including speed), and environment conditions. This supports the simulation of rare and safety-critical traffic scenarios. The scenarios and test cases have been implemented for demonstration purposes only, the actual number and details of test cases for the real operation can be defined after the project.
2. Eval_VV_8 – Effort Needed for Test (VALU3S_WP1_Automotive_1) - the effort needed is reduced for tests of traffic monitoring systems based on cameras and radar that are connected to the cloud. This includes extended coverage of rare but potentially critical traffic situations, tests of the accuracy of processing, and detection algorithms and others (CAMEA).
    - Baseline: testing was done mostly after deployment in real-time execution of the system. As there is usually a short time from installation to operation for customers,

more people are needed (developers, support, service, etc.…). Every system installation can differ (more or less) and, therefore, there is always some effort spent (and much more for newly designed systems). Effort can then be estimated in person months on average (depending on the complexity of the system). Estimated baseline: 3-4 person-months for repeatedly installed systems, up to 20 person-months for novel systems. The estimated baseline for testing a single scenario is up to 1 person-day.

- Improvement:
    i. The testing of camera/radar-based detection and tracking system (LP detection system) has been improved using the approach and the provided toolchain (BERGE simulator and ScenarioGenerator). The average effort spent for the setup of one new traffic scenario in the simulator is 0.5 person-day, and the test execution of one scenario is within the order of minutes (i.e., improvement factor ~2x).
    ii. The testing of the radar connected to the cloud has been improved using a simulated network environment. The effort spent on testing using NetLoiter as Hardware-in-the-loop is about 1 person-hour per test scenario (i.e., improvement factor ~8x).

3. Eval_VV_9 – Service Actions Needed (VALU3S_WP1_Automotive_2) - this evaluation criterion is applied to the development process of traffic monitoring systems based on cameras and radar that are connected to the cloud. As the project duration is too short for evaluation of the whole life cycle of the project, only an estimation of the metric is carried out (CAMEA).
    - Baseline: Service actions are often connected with bugs and issues coming from weak testing procedures. Most bugs are in newly developed and deployed systems. Estimated baseline: 1-2 person-months for repeatedly installed (but with unique installation parameters in most cases), 4-5 person-month for novel systems.
    - Improvement: Once a system is tested properly, the number of service actions is drastically reduced. An example can be an issue with the radar being upgraded with new firmware. A bug in a system has been found during the VALU3S project which causes the radar to be "bricked". For such a case, complicated service action is required with on-site maintenance as well as the repair of the device. Since the problem has been found using advanced network testing, the risk of the service action has lowered. The bug has not been fixed yet, so only the estimation is provided: effort spent on maintenance is reduced by 1 person-day per upgrade, which can vary from 1 to 4 upgrades per month, the distance to the locality (2-3 hours to 3-4 hours) and depends on the quality of used network connection (for some cases, LTE connection with the unstable signal is used). The improvement factor is then ~1.1x (90 person-hours to 82 person-hours).

List of evaluation criteria for SCP:

1. Eval_SCP_1 – Error Coverage (VALU3S_WP1_Automotive_3) - the metric is used for evaluating penetration testing in this use case (QRTECH).

- Number of planned penetration tests by type: 5 (Man-In-The-Middle, Denial of Service, Encryption/Certificates Evaluation, Reconnaissance scanning, Tampering (Physical) attack on Camera).
- Number of components/attack surface covered in the testbed: 3 (Camera, Locality, Network infrastructure) of 5.
- Number of evaluated security requirements: 1.
- Number of test cases involved from use case: 2.

2. Eval_SCP_2 – Number of Safety/Security Requirement Violations (VALU3S_WP1_Automotive_3) - this metric will be used for evaluating the results from testing that is planned to be performed in the use case (BUT).
    - Baseline: 3 out of 3 requirements within the evaluation scenario “Node connection to the cloud” have been violated.
    - Improvement: 1 of 3 requirements violated (UC1_R_15), 1 validated (UC1_R_16) and 1 validated partially (UC1_R_17). Improvement factor ~3x.
3. Eval_SCP_4 – Metrics to Evaluate AI/ML Algorithms (VALU3S_WP1_Automotive_1) - This applies to the detection and tracking of road users by radar and/or camera sensors, in different environmental conditions (harsh weather, occlusion). System performance as a function of weather and other environmental conditions and traffic situations (RISE, INFOTIV).
    - Baseline: 02 – Accuracy, confidence.
    - Improvement: 08 metrics have been implemented including mAP, IoU, Confidence, Levenshtein distance, TP, TN, FP, and FN. Additional metrics can easily be implemented upon the operational requirements after the project (e.g., the number of consecutive frames that have FP/FT over a limited time window).
4. Eval_SCP_6 – Metrics to Evaluate Cybersecurity (VALU3S_WP1_Automotive_3) - these metrics will be used for evaluating the results from threat analysis and penetration testing that is planned to be performed in UC1 (QRTECH).
    - The number of identified threats: 67 - Threat Model created with data flow diagram based on the system architecture. Using the model and tool (Microsoft Threat Modelling Tool) as well as the STRIDE framework a total of 67 threats were identified. The STRIDE framework stands for Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege, and it is a threat modelling methodology used to identify potential security threats and risks in a system. The categorisation of the threats using the mentioned STRIDE framework and further priority categorization was conducted was done by priority levels (Low(L), Medium(M), High(H)). The result of the categorization and prioritization of the 67 threats can be seen in the list below:
        i. Spoofing: 15 - 5H, 7M and 3L
        ii. Tampering: 5 - 3M, 2L
        iii. Repudiation: 8 - 6H and 2M
        iv. Information Disclosure: 10 - 1H, 6M and 3L
        v. Denial of Service: 12 - 11H and 1M
        vi. Elevation of privilege: 15 - 9M and 6L
        vii. Custom: 2 - 1H and 1M

5. Eval_SCP_13 – Accuracy of Simulated Sensor Output (VALU3S_WP1_Automotive_1) - a metric of accuracy, depending on the output format of the physical system, e.g., number or ratio of points accurately representing target scene in case of radar sensor (BERGE, INFOTIV).
   - Fréchet inception distance (FID) within the baseline: To have an even more realistic image to use it is possible to add FID to the simulator output. In FID an AI is trained with real image data captured by CAMEA. When the AI is ready it can create a new, adapted, image with the style of real captured images (Figure 3-3).

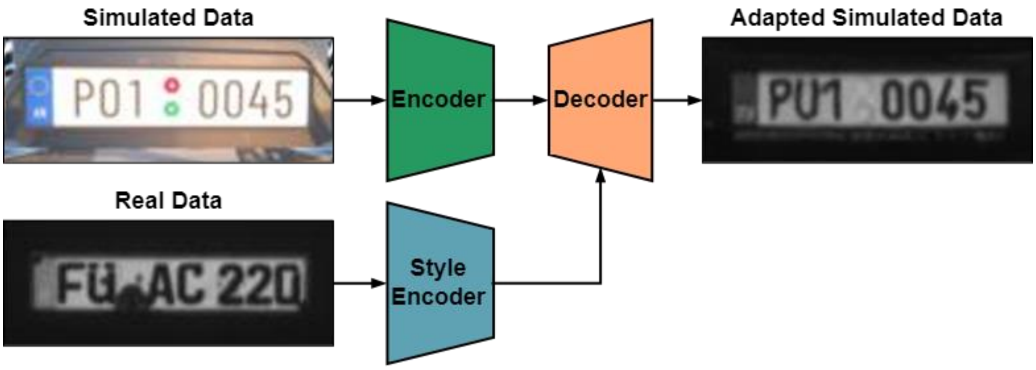


Figure 3-3 Real, simulated and adapted simulated capture data.

   - Accuracy of the virtual radar sensor within the baseline: The accuracy of the radar is directly connected to the set resolution of the radar. The resolution has two settings that both can be adjusted/set to the value the user wants. First is the angle of the cone the radar has. The higher the value, the wider the cone. The second values are the number of traces within the cone, the higher the value the more line traces and therefore higher the accuracy of the radar.

6. Eval_SCP_14 – Simulator Environment Quality (VALU3S_WP1_Automotive_1) - this metric will be used to track the visual quality increase (BERGE, INFOTIV).
   - Baseline: Not available
   - Improvement: The BERGE simulator has been developed using the Unreal engine, and supports the following items:
     i. Polygon count: There are 3 scenes in the simulator that all have different poly counts:
        1. City scene: Total of 20,2 million polygons.
        2. Country road scene: Total of 8,8 million polygons.
        3. Highway scene: Total of 28,5 million polygons.

   - Raytracing distance within the baseline was set to 200 meters to capture the reflections within a sphere. The ray tracing bounces are set to one, not to affect the frames per second too much (because more than one bounce require more computational resources and would affect the performance).

7. Eval_SCP_15 – Simulator Environment Functionality (VALU3S_WP1_Automotive_1) - this metric is used for the measurement of different environmental conditions used as an input for the CAMEA traffic surveillance system (BERGE, INFOTIV**).**

- Baseline: The Unreal engine does not provide the required functionality; those functionalities have been developed within the project into BERGE Simulator.
- Improvement: Several functionalities to support user setting-specific scenarios. There is also possible to pre-set a scenario through a JSON file so the user can generate synthetic datasets without having to use the graphical user interfaces. The input parameters that can be defined include:
    i. Weather conditions (sun, cloud, rain, snow)
    ii. Time of day (illumination condition)
    iii. Date
    iv. Scene (city, country road and highway)
    v. Vehicles (number of, speed, colour, license plate)
    vi. Placement and direction of vehicles
    vii. Pedestrians
    viii. Animals
    ix. Place lights in the scene.
    x. Camera position
    xi. Camera settings
    xii. Type of output (image, distorted, depth map, segmented image, segmented license plate, radar data/point cloud)
    xiii. File info of output (resolution, file format, bitmap)

Overall improvement addressing the main project objective (i.e., to reduce time and cost spent on V&V) can be summarised with the following chart depicted in Figure 3-4. The chart includes only evaluation criteria which directly or indirectly express the effort connected to the V&V activities as well as the quality of the product.

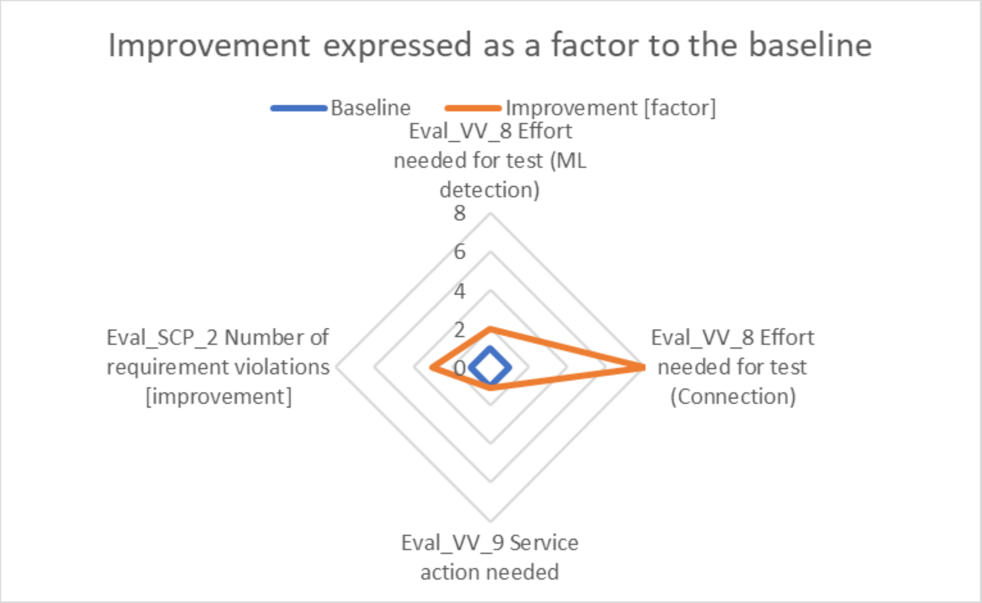


Figure 3-4. UC1 improvements as a factor to the baseline

3.1.6 Qualitative Results

Demonstrator 1: Network Intrusion Detection of a Node to Cloud Data

Participants Profile: QAM is applied to 10 subjects (9 males, 1 female) aged in the range of 24-34. The education level is relatively high as the subject pool is composed of 1 Post-Doc or higher degree and 1 PhD researcher and 8 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as researchers, scientists, managers, R&D engineers, software/hardware engineer, mechatronics/design engineers etc. having experience in the fields of "testing of IoT devices, HIL testing, NIDS, embedded systems, computer vision, AI/ML, data analytics, semantic web, IoT, digital twin etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-4. According to the results, the majority of the construct couples seem correlated with each other except for the PR-SI couple.

Table 3-4. UC1 – Demonstrator 1 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.629	1									
MO	0.585	0.539	1								
CO	0.589	0.800	0.380	1							
ROI	0.469	0.352	0.689	0.523	1						
PE	0.581	0.456	0.584	0.648	0.831	1					
PT	0.444	0.342	0.637	0.484	0.887	0.799	1				
PR	0.150	0.263	0.391	0.291	0.507	0.467	0.655	1			
SI	0.457	0.019	0.162	0.278	0.579	0.435	0.460	-0.240	1		
ATU	0.569	0.777	0.472	0.972	0.607	0.701	0.642	0.456	0.251	1	
BI	0.683	0.507	0.864	0.345	0.618	0.673	0.602	0.573	0.060	0.429	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-5, the majority of the questions asked to subjects are sufficiently reliable. However, answers related to questions considering PE and BI seem not reliable enough. The main reason can be the concentration level of subjects might not be high.

Regression Analysis: Finally, regression analysis is applied to estimating the relationships among QAM constructs, as seen in Table 3-6. For this demonstrator, there exists an inversely proportional relation between BI and {ATU, ROI, PE, SI, PT, PR} whereas other pairs influence each other positively. The significance of the test seems sufficient and statistically meaningful.

Table 3-5. UC1 - Demonstrator 1 Reliability Analysis

Cronbach-Alpha values	
PU	0.599
PEOU	0.462
MO	0.543
CO	0.525
ROI	0.360
PE	-3.252
PT	0.310
PR	0.768
SI	0.111
ATU	0.602
BI	-1.600

Table 3-6. UC1 - Demonstrator 1 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.304xPU + 4.184	0.342	0.076	2.041
H2	CO-PU	Right	CO = 0.27xPU + 4.44	0.347	0.073	2.061
H3	PEOU-PU	Right	PEOU = 0.57xPU + 2.7	0.396	0.051	2.289
H4	PU-ATU	Right	PU = 0.3xATU + 4.1	0.324	0.086	1.959
H5	PEOU-ATU	Right	PEOU = 0.47xATU + 2.8	0.603	0.008	3.486
H6	ATU-BI	Inverse	ATU = -0.33BI + 4.06	0.184	0.216	1.343
H7	ROI-BI	Inverse	ROI = -0.5xBI + 3.32	0.382	0.057	2.222
H8	PE-BI	Inverse	PE = -0.83xBI + 1.19	0.453	0.033	2.574
H9	SI-BI	Inverse	SI = -0.05xBI + 5.64	0.004	0.869	0.171
H10	PT-BI	Inverse	PT = -0.53xBI + 3.12	0.363	0.065	2.133
H11	PR-BI	Inverse	PR = -0.44xBI + 3.65	0.328	0.084	1.976

Demonstrator 2: V&V of ML-based system using simulators

Profile: QAM is applied to 11 subjects (10 Males, 1 female) aged in the range of 24-54. The education level is relatively high as the subject pool is composed of 5 Post-Doc or higher-degree and 6 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as “researchers, academician, managers, global consultants, deans, R&D engineers, software/hardware engineers etc.” having experience in the fields of “Quality & Analysis, NIDS, embedded systems, computer vision, automotive, AI/ML, data analytics, semantic web, IoT, digital twin etc.”.

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-7. According to the results, the majority of the construct couples seem fully correlated with each other.

Table 3-7. UC1 – Demonstrator 2 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.526	1									
MO	0.712	0.910	1								
CO	0.600	0.878	0.878	1							
ROI	0.706	0.907	0.968	0.846	1						
PE	0.479	0.600	0.684	0.607	0.750	1					
PT	0.577	0.840	0.907	0.726	0.931	0.576	1				
PR	0.270	0.455	0.546	0.459	0.563	0.450	0.600	1			
SI	0.261	0.567	0.495	0.225	0.528	0.339	0.675	0.468	1		
ATU	0.317	0.512	0.615	0.466	0.537	0.335	0.530	0.333	0.198	1	
BI	0.710	0.792	0.903	0.675	0.850	0.662	0.755	0.301	0.477	0.694	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-8, the majority of the questions asked to subjects are sufficiently reliable. However, answers related to questions considering PR and BI seem not reliable enough. The main reason can be the concentration level of subjects might not be high.

Table 3-8. UC1 - Demonstrator 2 Reliability Analysis

Cronbach-alpha values	
PU	0.381
PEOU	0.623
MO	0.793
CO	0.518
ROI	0.472
PE	0.291
PT	0.403
PR	-0.035
SI	0.611
ATU	0.026
BI	-0.165

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-9. For this demonstrator, there exists an inversely proportional relation between BI and {ATU, ROI, PE, SI, PT, PR} whereas other pairs influence each other positively. The significance of the test seems sufficient and statistically meaningful.

Table 3-9. UC1 - Demonstrator 2 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.40xPU + 3.36	0.507	0.014	3.041
H2	CO-PU	Inverse	CO = 0.37xPU + 4.01	0.360	0.051	2.250
H3	PEOU-PU	Inverse	PEOU = 0.52xPU + 3.29	0.277	0.096	1.857
H4	PU-ATU	Inverse	PU = 0.51xATU + 1.96	0.100	0.343	1.001
H5	PEOU-ATU	Inverse	PEOU = 0.81xATU + 1.06	0.262	0.107	1.788
H6	ATU-BI	Inverse	ATU = 0.68xBI + 1.86	0.482	0.018	2.895
H7	ROI-BI	Inverse	ROI = -0.97xBI + 0.8	0.722	0.001	4.838
H8	PE-BI	Inverse	PE = -1.08xBI + 0.16	0.438	0.027	2.648
H9	SI-BI	Inverse	SI = -0.58xBI + 2.66	0.228	0.138	1.629
H10	PT-BI	Inverse	PT = -0.88xBI + 1.13	0.570	0.007	3.452
H11	PR-BI	Inverse	PR = -0.304xBI + 3.78	0.091	0.369	0.946

3.1.7 Observed Limitations, Lessons Learnt and Best Practices

Frequent interaction and close collaboration of all involved partners contributed to three topics. The findings, limitations, and lessons learnt are the following:

ML-based detection: V&V of ML-based components is still an open research question; the proposed method and toolchain provide an example of a best practice assurance process. The tools developed during the VALU3S project use state-of-the-art technology and apply the feedback from the use case provider enabling the tools to be used in the real world or real-world-based situations (e.g., generation of license plates for a simulated scenario based on a real set of license plates).

Testing of network applications: There are several tools aimed at testing network applications. The tools are either too generic and focus on lower-level network links, or too specific, focusing on tailored networking issues. To the best of our knowledge, the tool chain developed within VALU3S aims both at easy usage and a simple definition of purpose-specific needs. The limitations of the developed workflow are the performance (e.g., testing on high throughput network links) and easy and stable remotely controlled fault-injector (e.g., using a predefined, well-defined and easily maintainable fault configuration, or management via cables instead of wireless connection as it might and has already been jammed during testing).

Cybersecurity assessment: Understanding the infrastructure, including, devices, services, and protocols is essential for creating a comprehensive threat model and conducting a threat analysis as part of the cybersecurity assessment V&V. The STRIDE framework provides a structured approach to identify potential security threats in systems and it covers the most common types of security threats that can occur. Penetration testing workflow following the PTES framework provides comprehensive coverage in a methodical systematic approach that can provide detailed coverage.

According to the qualitative analysis results, the mean and standard deviation of expert responses to UC1 demonstrations are given in Table 3-10 and Table 3-11. The expert responses to first demonstrator seem more positive than the second demonstrator. Responses to all QAM constructs for Demonstrator-1 are between 5,15 and 5.90 (out of 7) meaning that there is a strong intention to use network intrusion detection solutions in traffic surveillance systems where the users are aware of potential cyber security threats against node-to-cloud communication infrastructures.

Table 3-10 Mean and standard deviation of experts' responses to UC1 - Demonstration 1

UC1/ Demonstrator-1	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,84	5,39	5,44	5,22	5,22	5,61	5,24	5,15	5,46	5,59	5,90
Std Dev	0,62	0,68	1,19	1,36	1,09	0,70	0,98	1,14	1,10	1,13	0,87

For the second demonstration experts still have a positive attitude however, they feel more sceptic about the V&V of ML-based systems using simulators. Since the PU and MO are quite high there is a perceptional motivation to use Demonstrator-1 outputs. However, other factors like PT, PR, ROI, CO, PE, SI ANTU and PEOU are comparatively low. If any potential problem related to the users' responses was ignored, e.g., statistical significance or unreliability of responses, the main reason behind the relatively low acceptance might be the barriers related to the use of AI-based techniques, user reluctance or suspicion of AI-based solutions and their potential safety consequences.

Table 3-11 Mean and standard deviation of experts' responses to UC1 - Demonstration 2

UC1/ Demonstrator-2	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,73	4,73	5,16	4,68	4,52	4,64	4,61	4,53	4,33	4,89	5,16
Std Dev	0,80	0,81	1,40	1,30	1,10	0,77	1,08	1,24	1,03	1,29	1,25

Nevertheless, when considered together with the quantitative results, the technology providers can increase the level of acceptance by a more proactive dissemination strategy to increase consciousness and awareness. The advancement in the technology and promising results obtained throughout the project may help stakeholders to benefit more from the UC1 results.

3.2 Use Case 2 – Car Teleoperation (UC2)

UC2 from Roboauto (ROBO) focuses on the cybersecurity of the transmission line and the routers in car teleoperation. Any vehicle equipped with sufficient sensors and electronic actuators can be controlled remotely. Every teleoperated system consists of a remotely-controlled vehicle(s) (car, towing tractor etc.) and a remote station operated by a human driver. These parts are connected by the heart of the system which is called the Gateway. All vehicles and remote stations connect to the Gateway where the vehicle or the remote station is authenticated. Once the authentication is complete, the vehicle–remote station status is reported to the Gateway. The Gateway sends the data to the fleet management where it is visualized. An operator selects a proper vehicle and a remote station to connect via fleet management. Once connected, the driver can see the image streamed from the vehicle as well as its speed and other (telemetry) data, see Figure 3-5. If the remote operation is selected by the operator/driver, the vehicle can be driven to a location of choice. Upon reaching the destination, the driver/operator can "leave" the vehicle and switch to another one.

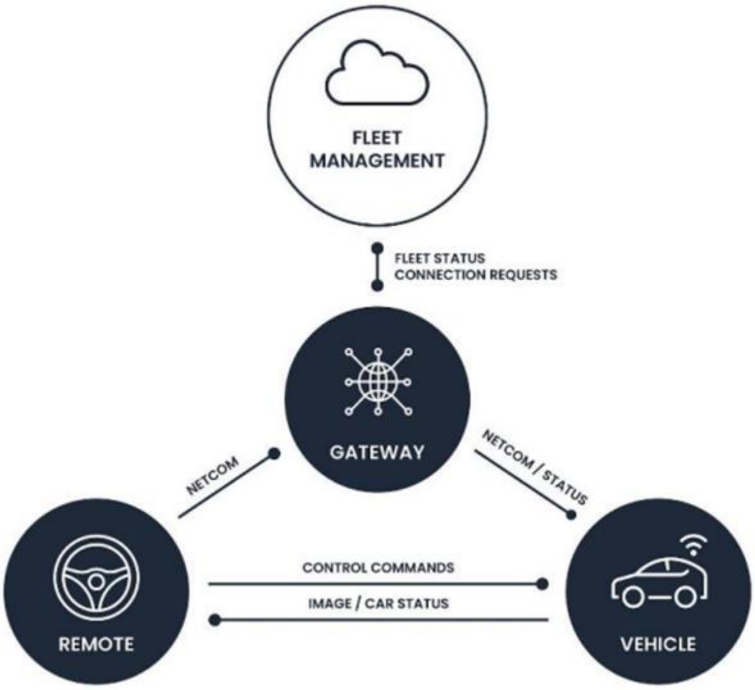


Figure 3-5 Diagram of streamed data flow

3.2.1 V&V challenges

Within this use case, we are facing some V&V challenges that arise from various factors that affect the evaluation scenarios and the corresponding baseline. Some of the key challenges include:

- **Complexity of the system** - The system under test is very complex with multiple elements communicating asynchronously with each other. It is difficult to test only a part of the system as everything communicates with each other.
- **Network communication** - The communication between the elements takes place over the network and needs to be simulated for the project. Simulating network communication is not easy, and it adds to the complexity of the system.
- **Customer requirements –** ROBO's customers require fast interactions and changes, which puts high pressure on us to reduce the time spent on validation and verification. On the other hand, the complexity and quality expectations are constantly increasing, making it challenging to meet these requirements.
- **Hardware in the loop (HIL)** - We do not have access to a real car to perform HIL testing. This makes it difficult to test the system under realistic conditions.

- **Software in the loop (SIL)** - We lack people who can test the system, and there is no unified solution for testing changes in the system. This makes it challenging to validate and verify the system in a timely and efficient manner.
- **Real-time -** Challenges arise in this use case due to the real-time nature of the system, which requires tools that can operate in real-time and carefully configured test cases that can effectively test the system's real-time behaviour.

3.2.2 Contributors

Partners contributing to the UC: ROBO, BUT, RISE, VTI, QRTECH.

3.2.3 Contributors' Roles & Evaluation Scenarios

Evaluation scenarios defined for this use case are listed below; the assignments of the contribution of all UC partners can be found in Table 3-12:

1. VALU3S_WP1_Automotive_4 - Transmission line under different performance conditions - Evaluation of transmission line under big latency, jitter (varying latency), and heavy line usage on the transmission line between teleoperated vehicle and the remote station.
2. VALU3S_WP1_Automotive_6 - Transmission line switching - Evaluate the safety of the teleoperated vehicle during Wi-Fi AP switching.
3. VALU3S_WP1_Automotive_14 - Transmission line cybersecurity – Evaluate cases of man-in-the-middle attacks tampering with data.

Table 3-12 Overview of UC2 partners' contributions to evaluation scenarios

Evaluation Scenario	ROBO	BUT	RISE	VTI	QRTECH
VALU3S_WP1_Automotive_4	x	x	x	x	
VALU3S_WP1_Automotive_6	x				
VALU3S_WP1_Automotive_14	x		x	x	x

Individual UC partners are contributing to evaluation scenarios as follows:

- ROBO, as the UC provider, supports all UC partners and provides and updates a mock-up of the real teleoperated system so all the partners can easily test their methods and tools from anywhere.
- BUT contributes mostly to the VALU3S_WP1_Automotive_4 evaluation scenario. The test cases related to checking if the system under test runs correctly under different network conditions use the tool NetLoiter (a new tool developed by BUT) which systematically simulates fault injection in network traffic. The tool is coupled in a toolchain with Plogchecker (a new tool currently developed) which is used as a test oracle whether the system monitored by the tool runs correctly. Plogchecker checks during run-time the log produced by the system and reports violations of expected sequences of log events. Netloiter implements the Software Implemented Fault Injection method and Plogchecker applies Runtime Verification based on Formal

Specification. The contribution of BUT focuses on formalising the description of correct runs of SUT and on development and improving the tools to meet the needs of these use cases. RISE contributes by evaluating 'transmission line' disturbances by injecting faults and attacks on the transmission line. The "transmission link" between vehicles and remote stations. Commands are sent from the remote station to the teleoperated vehicles. VALU3S_WP1_Automotive_4 and VALU3S_WP1_Automotive_14 are relevant to work packages in the UC2 for evaluation.

- RISE/VTI contribute to the development of the ComFASE tool and the simulation environment which supports V&V of ROBO's car teleoperation. ComFASE is a "Communication-based Fault and Attack Simulation Engine" which enables fault and attack injection on the vehicle communication system to evaluate the safety implications on the interconnected vehicle system. This interconnection is either vehicle-to-vehicle (V2V) or vehicle-to-everything (V2X) such as teleoperation. In the UC2 perspective, ComFASE is used to evaluate the transmission link disturbances by injecting communication faults and attacks. This work is a collaboration of RISE and VTI. The development of ComFASE is performed by RISE. VTI, on the other hand, integrated the Roboauto SIL (software in the loop also called the 'mock-up') with the existing wireless communication network simulation framework, namely Veins_INET which is based on OMNeT++ network simulator. As ComFASE can be integrated and used for fault and attack injection testing with any simulation framework that is based on OMNeT++ so the Veins_INET extension makes it possible to perform fault and attack injection in the wireless communication network to evaluate the Roboauto's SIL system. The focus of RISE and VTI is on the VALU3S_WP1_Automotive_4 evaluation scenario. Therefore, this allows V&V to be performed on an actual software component (i.e., the mock-up).
- QRTECH contributes to VALU3S_WP1_Automotive_14 by identifying threats and vulnerabilities as well as exploiting vulnerabilities in the penetration testing phase, to improve the system's cybersecurity posture and reduce threats.

This use case deals with testing and verification of reliability and security of car teleoperation. Figure 3-6 illustrates the V&V tools that are being developed and are demonstrated in this use case as well as the V&V methods associated with the tools.

Baseline of the V&V process:

For continual functionality checking a set of regression tests is executed after each change. These tests include not only the unit testing of the modules but also the mock-up setup of a whole system and scripted testing scenarios. Automatic testing pipelines are deployed as part of a CI in GitLab.

Every month a new version of the software is released, and additional manual integration tests are performed in a laboratory environment with a small robotic vehicle. A set of integration tests covering potential problems with deploying and running a car teleoperation system is continuously updated and extended. Some features are also field-tested on the prototypes; however, this is not regular.

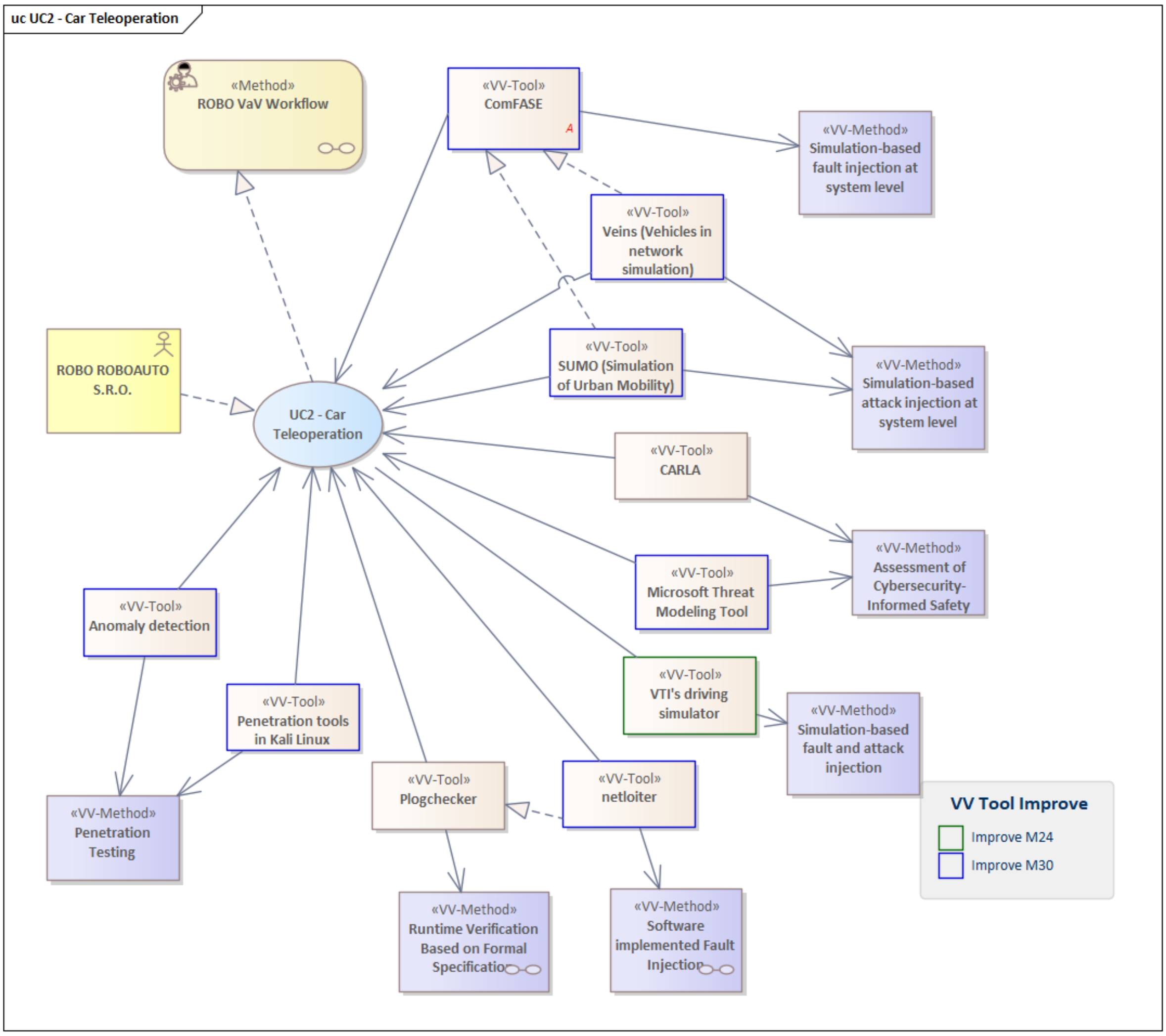

uc UC2 - Car Teleoperation
«Method» ROBO VaV Workflow
«VV-Tool» ComFASE
«VV-Method» Simulation-based fault injection at system level
«VV-Tool» Veins (Vehicles in network simulation)
ROBO ROBOAUTO S.R.O.
«VV-Tool» SUMO (Simulation of Urban Mobility)
«VV-Method» Simulation-based attack injection at system level
UC2 - Car Teleoperation
«VV-Tool» CARLA
«VV-Method» Assessment of Cybersecurity-Informed Safety
«VV-Tool» Microsoft Threat Modeling Tool
«VV-Tool» Anomaly detection
«VV-Tool» VTI's driving simulator
«VV-Method» Simulation-based fault and attack injection
«VV-Tool» Penetration tools in Kali Linux
«VV-Tool» Plogchecker
«VV-Tool» netloiter
«VV-Method» Penetration Testing
«VV-Method» Runtime Verification Based on Formal Specification
«VV-Method» Software implemented Fault Injection
VV Tool Improve
Improve M24
Improve M30

Figure 3-6 Tools and Methods for UC2 – Car Teleoperation

List of evaluation criteria of the V&V process:

1. Eval_VV_1 – Time of Test Execution - The time of test execution of the faulty network connection is not available as there is currently no such test.
    - Based on other testing scenarios it is estimated that manual network connection faults testing could take up to one man-day.
    - It is expected that automated tests will improve the time spent on testing. The method closely related to this criterion in UC2 is the Software-Implemented Fault-Injection.
2. Eval_VV_2 – Coverage of Test Set - The test set, based on fault and attack injection, will be delivered by our partner RISE/VTI.
    - The simulation environment is up and ready for test execution and evaluation. A list of tests is created together with the UC2 providers to evaluate the teleoperation system's functionality.
3. Eval_VV_3 – Number of Test Cases - Test set reduction is dependent on the test set which should be delivered by our partners: RISE and VTI. The test set will be based on fault and attack injection.

- Several tests are executed to verify the teleoperation functionality according to the test requirements.

List of evaluation criteria for SCP:

1. Eval_SCP_2 – Number of Safety/Security Requirement Violations - There are currently 12 SCP requirements. The test set, based on penetration testing and software fault injection, will be delivered by our partners and BUT.
    - No estimation of violated requirements made so far.
2. Eval_SCP_7 – Number of Prevented Accidents - A list of accidents that could occur is currently being prepared.
    - For now, no accidents could be prevented by testing as the list is not finished yet.
3. Eval_SCP_11 – Simulation-level System Robustness - list of injected faults and attacks into simulated wireless communication:
    - Delay attack
    - Denial of service (DoS) attack
    - Disconnection while the system is running.

3.2.4 Demonstration

To demonstrate the progress and help partners of UC2 in VALU3S, a complex mock-up setup of the real teleoperation system was created by Roboauto. For this project, the environment has been enhanced with automated inputs and a simulated vehicle, all of which are connected to the teleoperation system that remains unchanged.

This mock-up setup allows all partners to easily access the teleoperation technology via their own devices and test their methods without any barriers before the demonstration in the real operation. Outputs of the partners' testing show how these methods can help in future development, what errors have been found and how the product can be improved, and the errors reduced. This process demonstrates how relevant changes can be implemented, and therefore, makes the system safer and more robust.

After this phase of testing in the mock-up setup, the same tests are performed in a real operation environment with a real vehicle to compare possible discrepancies. Based on the results of the testing, successful methods might be integrated into the real testing process to automatically generate the results of the evaluation criteria. The demonstrated items are as follows (see Table 3-13):

Table 3-13 Overview of demonstration prepared by UC2 partners.

Item #	Demonstration name	Description/Purpose	Type	Responsible
1	V&V of Car Teleoperation application under Faults and Attack in Wireless Communication Channel	ComFASE is a communication-based fault and attack injection tool that is developed to inject faults and attacks in the teleoperation communication system for the purpose of verification and validation of the safety features implemented in the teleoperated vehicle system. For this purpose, the teleoperated modules provided by the UC provider are integrated into the WiFi-based communication simulation framework to test the functionality. The complete simulation environment includes the UC related modules, WiFi-based communication framework and ComFASE.	Lead, demonstration	RISE/ VTI/ROBO
2	Testing network communication using NetLoiter	Show how to setup and run tests of system performing under different network conditions. Demonstration of automated validation of requirements on network link reliability.	Lead, demonstration	BUT
3	Integration of threat modelling and penetration testing	Integration of threat modelling and penetration testing for improving the use case workflow.	Complementary , demonstration	QRTECH

3.2.5 Quantitative Results

List of evaluation criteria of the V&V process:

1. Eval_VV_6 – Cost of finding and fixing a coding bug - Simple (and automated) tests relatively reduce the time needed for bug fixing as several bugs could be manifested in a single execution of the test suite (ROBO/BUT).
   - Baseline: When testing a system for remote control of a vehicle, the estimated time for manual preparation of 6 simple tests is 6-12 hours. Manual test execution time takes 30 minutes for each test. Finding and fixing a bug (considering one bug per 6 tests) takes an average of 1 hour (20 minutes to 1 hour, occasionally several days) in such scenarios where a real vehicle is replaced by a simulation model. The cost of a bug found using manual testing using 6 tests is 5,5 hours.
   - Improvement: Automation reduces the time for test preparation (10 minutes per test) and test execution (1 minute per test). Moreover, it can uncover more bugs in the future development of the system. The cost of a bug is approximately 3 hours (approximately 2x improvement).

2. Eval_VV_2 – Test coverage – Measuring how much software test coverage items have been covered by a test set (set of test cases, also known as test suite). Measurement provided by ROBO/BUT.
    - Baseline: Measuring software artefacts covered by tests is crucial feedback on how well the system has been tested. The module which manages a safe and secure link between the vehicle and the remote station is one of the critical parts of the system for a remotely operated car. Six automated tests of simulated driving verify behaviour in different situations and cover up to 16% of source-code statements (up to 56% of the code of a module for network communication).
    - Improvement: Incorporating fault injection of a network link during testing increases the statement coverage up to 76% of the code of a module for network communication (improvement factor ~1.36x).

List of evaluation criteria of the SCP:

1. Eval_SCP_1 – Error Coverage - An error coverage criterion is dependent on the fault and attack injection test set and penetration testing test set which are delivered by QRTECH and RISE/VTI. According to these tests, the following improvements have been recorded as compared to the baseline criteria that was set to 0 at the beginning of the project:
    - Number of penetration tests by type of attack: 6 (Arp poisoning, Man-In-The-Middle, Denial of Service, Brute Force attack, Encryption/Certificates Evaluation, Reconnaissance scanning)
    - Number of components involved in the tests: 3 of 3
    - Number of evaluated security requirements: 1
    - Number of test cases involved from use case: 2
2. Eval_SCP_2 – Number of Safety/Security Requirement Violations (VALU3S_WP1_Automotive_4) – this metric has been used for evaluating the teleoperation system from testing the network application that has been performed in the use case by BUT.
    - Baseline: 7 non-validated requirements within evaluation scenario “Transmission line under different performance conditions”.
    - Improvement: only 1 requirement is not validated (UC2_R_5), the rest 6 requirements have been confirmed (see D5.5 [11] for more details about the network settings). The untested requirement depends on testing with human-in-the-loop which was out of the scope of test automation (as the main contribution of BUT). Improvement factor ~3x.
3. Eval_SCP_3 – Number of Malicious Attacks and Faults Detected - Malicious attacks and faults are defined by QRTECH. According to these tests, the following improvements have been recorded as compared to the baseline criteria that was set to 0 at the beginning of the project:
    - Number of possible detected attacks and faults by type: 6
        - Denial of service and DDoS: 4 (Jamming Wi-Fi, UDP flooding, SYN flooding, ICMP flood)
        - Social engineering and phishing/spear phishing
        - Network intrusion
        - Brute force attack/Weak passwords
        - Malware infection
        - Physical attack/tampering on vehicle

4. Eval_SCP_6 – Metrics to Evaluate Cybersecurity - The number of threats will be determined by threat analysis by our partner QRTECH. The time of availability of the system under test is measured by RISE/VTI.
    - The number of identified threats: 65 - Threat Model created with data flow diagram based on the system architecture. Using the model and tool (Microsoft Threat Modelling Tool) as well as the STRIDE framework a total of 65 threats were identified. The STRIDE framework stands for Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege, and it is a threat modelling methodology used to identify potential security threats and risks in a system. The categorisation of the threats using the mentioned STRIDE framework and further priority categorization was conducted was done by priority levels (Low(L), Medium(M), High(H)). The result of the categorization and prioritization of the 65 threats can be seen in the list below:
        - Spoofing: 4 - 3M and 1L
        - Tampering: 5 – 1H, 4M
        - Repudiation: 6 – 6L
        - Information Disclosure: 13 - 9M and 4L
        - Denial of Service: 30 - 11H, 18M and 1L
        - Elevation of privilege: 7 – 5H and 2M

Summarised improvement incorporating V&V technology developed by UC2 contributors uses evaluation criteria measuring effort, test coverage, and a number of violated requirements. All of these are connected to the evaluation scenario VALU3S_WP1_Automotive_4 – Transmission line under different performance conditions. The chart in Figure 3-7 graphically represents the improvement using factors to the baseline (factors have been used to harmonise the results). Improvement connected with cybersecurity assessment is hard to define due to the missing baseline; there has not been any cybersecurity analysis to be compared with the final state.

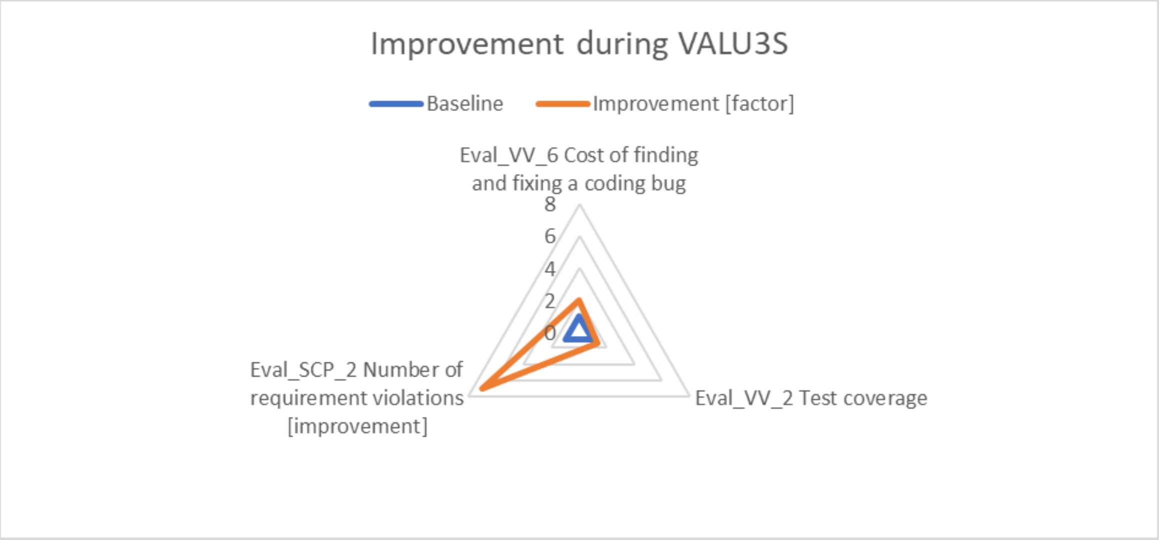


Figure 3-7 UC2 improvements as a factor to the baseline

Demonstrator 1: V&V of Car Teleoperation application under Faults and Attack in Wireless Communication Channel

In this use case, the evaluation of demonstrator 1 is performed by RISE and VTI. In fact, they:

- Developed ComFASE tool that injects faults and attacks in the realistic V2V or V2X communication system model.
- Developed a simulation environment. The core parts of the simulation environment are Roboauto's mock-up, WiFi communication network and ComFASE fault and attack injection tool. The complete simulation environment (simulation environment architecture and real-test setup) is presented in Figure 3-8 and Figure 3-9.
- Modelled and injected three types of jamming attacks to evaluate and validate the fall-back mechanism of the teleoperation system. The attack models, their implementation and test requirements needed for validation are presented in Table 3-14.

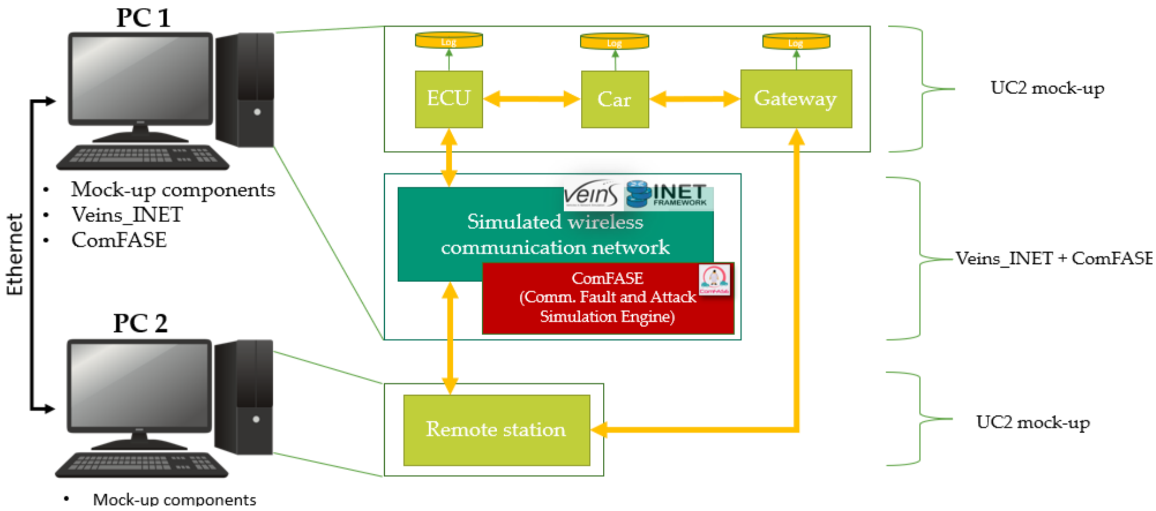


Figure 3-8 ComFASE simulation environment

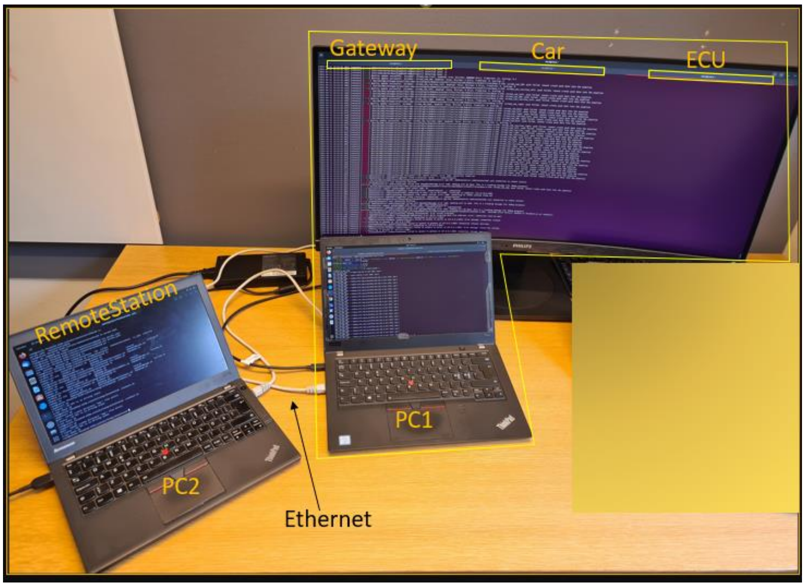


Figure 3-9 Real-test setup for UC2

Table 3-14 Attack models, implementation, and test requirements

Attack Model	Attack Implementation	Test requirement (TR)
Disconnection	Both control commands and video stream between the car/ecu and remote station are stopped.	**TR1**-Message received within 1.5 seconds » normal operation. **TR2**-No message within 1.5 seconds » Disconnect and trigger vehicle safe stop.
Delay Attack	Additional delay is added to data packets in the transport layer (UDP). The added delay is removed before the simulation ends.	**TR3**-Message delay < 150 ms » normal operation. **TR4**-Message delay >= 150 ms » trigger vehicle safe stop
DoS Attack	Additional delay is added to data packets in the transport layer (UDP). The added delay remains until the end of the simulation.	**TR3**-Message delay < 150 ms » normal operation. **TR4**-Message delay >= 150 ms » trigger vehicle safe stop

Attack Injection Results

Baseline testing: Prior to any attack injection it is imperative to validate the results from the simulation environment (mock-up + Veins_INET) w.r.t. baseline that is a standalone mock-up without any additional communication network layers. The results of the baseline testing are given in Table 3-15 and Figure 3-10.

Table 3-15 Baseline testing results

Setup	Connection lost	Latency (ms)		
		Min.	**Max.**	**Mean.**
Baseline	15.082	0	5.3	1.83
Veins_INET	16.575	1.67	83.67	9.68

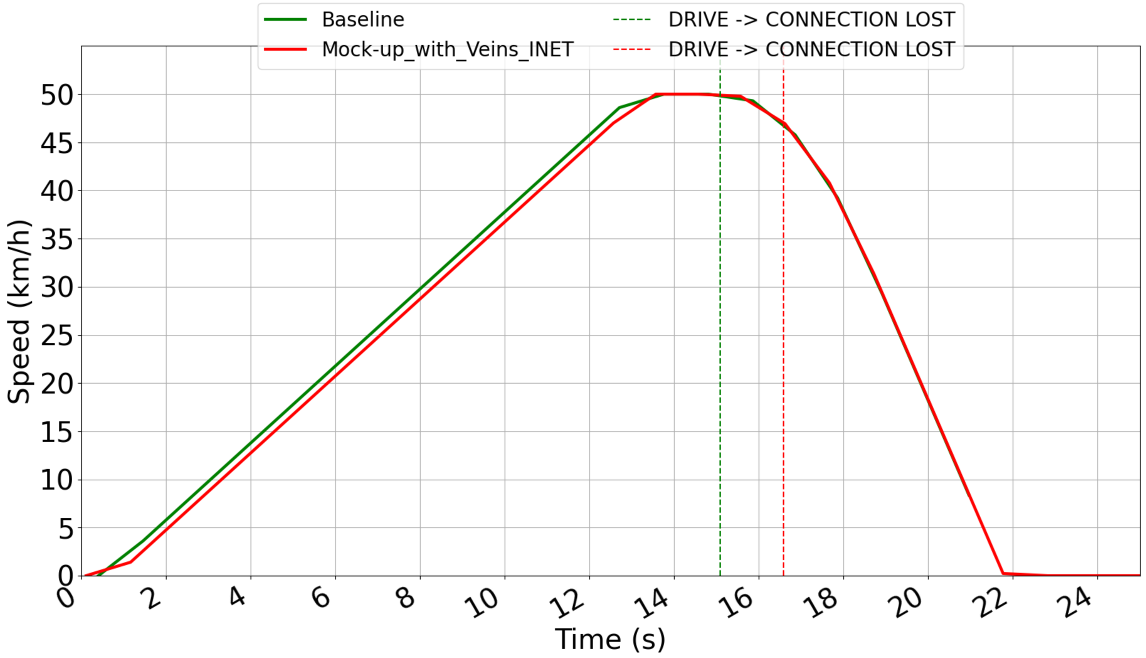


Figure 3-10 Baseline testing results as a function of time and speed

A few tests are executed for each of the implemented attack models. Below we provide some results of disconnection and delay attacks.

Disconnection: In this attack, the complete communication loss (i.e., both control command channel and video streams are stopped) occurs between the 'car' and the 'remotestation' modules. The top speed test is executed where the vehicle accelerates to achieve the top speed (i.e., 50 km/h in this case). After achieving the top speed, the vehicle speed goes to zero when the test script is finished or if the communication is lost.

The disconnection attack is activated at different times to verify the test requirement (TR1, TR2). The results of the disconnection testing are given in Table 3-16 and Figure 3-11.

Table 3-16 Disconnection attack testing results

Attack Activation Time	Connection Lost Recorded Time
4.5 s	7.496
5 s	7.998
7 s	9.974
10 s	13.002

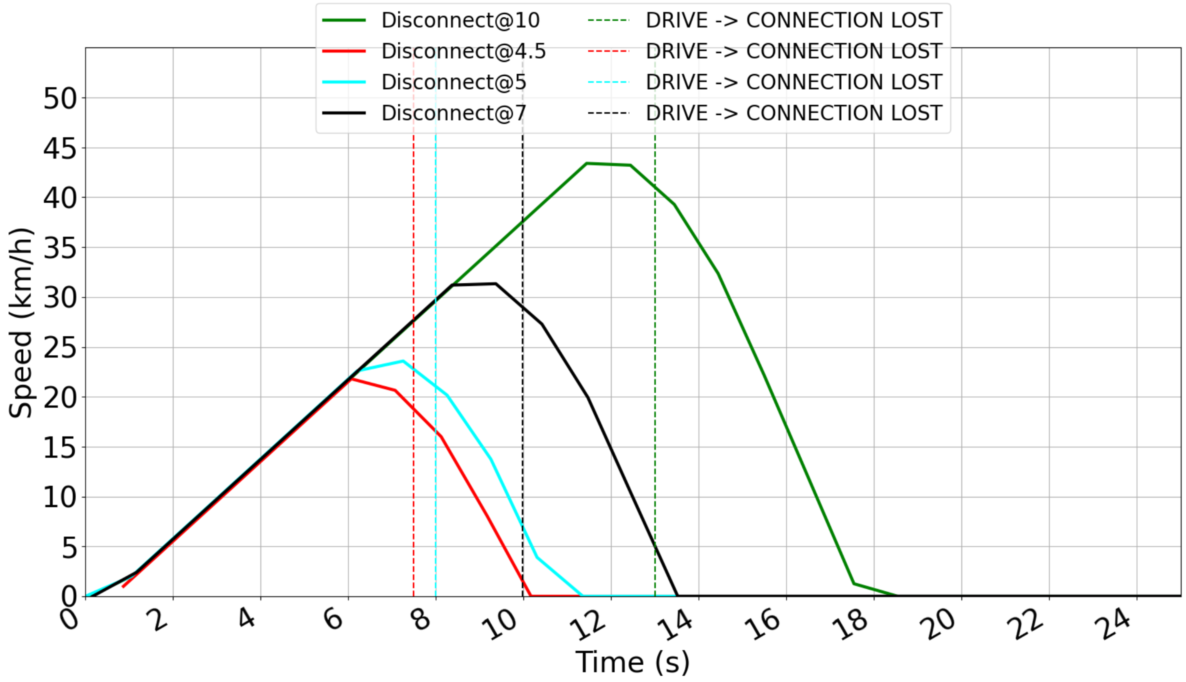


Figure 3-11 Testing of disconnection attack as a function of time and speed

Delay Attack: In this attack scenario, an additional delay is added to packets in the transport layer (UDP). Additional delay is the delay on top of the existing simulation delay (i.e., 3 ms approximately). The delay is added as an interval in real-time (real-time is slower than simulation time) where 1 second simulation time is approximately 0.03 – 0.04 seconds in real-time. The results of the delay attack are given in Table 3-17 and Figure 3-12.

Table 3-17 Delay attack testing results

Setup (DelayAmount @T=startTime-endTime)	**Latency (ms)**		
	Min.	**Max.**	**Mean.**
300ms@T=5-8s	1.33	72.67	8.31
4s@T=5-8s	2	137	56.66
4s@T=5-10s	2	307	118.33
6s@T=5-10s	1	338	94.32
10s@T=5-10s	1	381	32.41

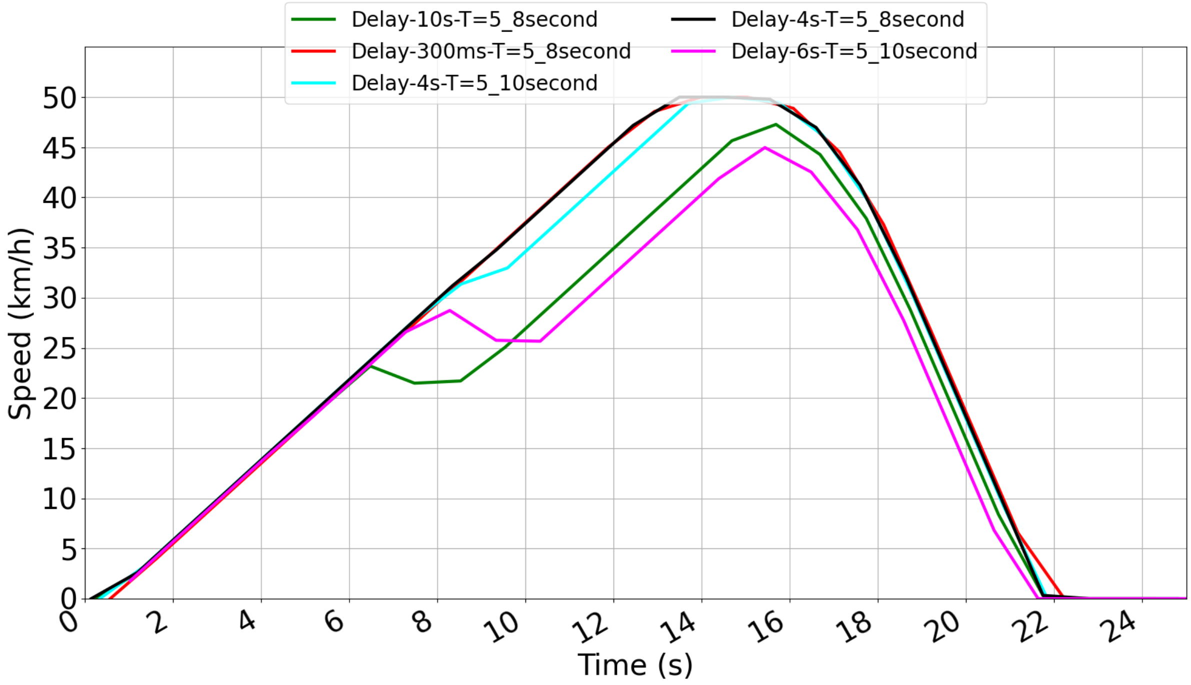


Figure 3-12 Testing of delay attack as a function of time and speed

From the above graphs, the total delay is less than 1.5 seconds in all experiments. When the communication delay is in the range of 150 ms and 1.5 seconds, the vehicle safe stop is triggered but the communication between the car and 'remotestation' is not disconnected by the system. The communication delay of less than 150 ms does not affect the normal operation of the teleoperation system. This validates TR3 and TR4.

In the DoS attack, the delay exceeded the 1.5-second threshold, so the results are the same as the disconnection test. Therefore, the graphs are not shown for DoS results.

3.2.6 Qualitative Results

Roboauto's main focus in the VALU3S project within the car teleoperation use case will be on the safety of the transmission line. In cooperation with our partners, we derived evaluation scenarios from these two areas, and plan to verify them on both mock-up setup (simulation using the real teleoperation system backend, automated inputs, and outputs), and real teleoperated vehicle (car or laboratory robot). The qualitative assessment is implemented by mainly considering the experts' feedback on the Lead demonstrator (demonstrator-1).

Participants Profile: QAM is applied to 15 subjects (14 Males, 1 female) aged in the range of 24-54. The education level is relatively high as the subject pool is composed of 2 Post-Doc or higher-degree and 3 PHD researchers and 10 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "academicians, CEOs, project managers, data scientists, R&D engineers, ML engineers, software/hardware engineers, etc." having experience in the fields of "formal verification, automotive verification, automotive, remote control of vehicles, cyber-security, system integration, real-time physiological computing, AI/ML, data analytics, semantic web, IoT, digital twin etc.".

Correlation Analysis: Correlations among the TAM constructs are presented in Table 3-18. The results show that MO is not correlated with other constructs except ROI.

Table 3-18. UC2 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.599	1									
MO	0.442	0.128	1								
CO	0.644	0.795	-0.036	1							
ROI	0.408	0.895	0.020	0.680	1						
PE	0.282	0.861	-0.164	0.670	0.905	1					
PT	0.424	0.927	-0.091	0.743	0.951	0.967	1				
PR	0.178	0.589	-0.112	0.401	0.688	0.626	0.687	1			
SI	0.287	0.700	-0.362	0.779	0.682	0.867	0.835	0.516	1		
ATU	0.110	0.485	-0.491	0.382	0.637	0.707	0.658	0.414	0.631	1	
BI	0.105	0.687	-0.121	0.520	0.858	0.877	0.854	0.662	0.715	0.588	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-19, the majority of the questions asked to subjects are sufficiently reliable. However, answers related to questions considering ATU and BI seem not reliable enough. The main reason can be the concentration level of subjects might not be high.

Table 3-19. UC2 Reliability Analysis

Cronbach-alpha values	
PU	0.361
PEOU	0.313
MO	0.509
CO	0.152
ROI	0.235
PE	0.706
PT	0.753
PR	0.415
SI	0.653
ATU	-0.282
BI	-1.071

Regression Analysis: Finally, regression analysis is applied to estimating the relationships among QAM constructs, as seen in Table 3-20. For this demonstrator, there exists an inversely proportional relation between BI and {ATU, ROI, PE, SI, PT, PR} as well as PU-ATU whereas other pairs influence each other positively. The significance of the test seems sufficient and statistically meaningful.

Table 3-20. UC2 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	p-Value	t-Value
H1	MO-PU	Right	MO = 0.41xPU + 3.431	0.195	0.099	1.777
H2	CO-PU	Right	CO = 38xPU + 4.002	0.414	0.010	3.033
H3	PEoU-PU	Right	PEoU = 0.33xPU + 4.107	0.359	0.018	2.697
H4	PU-ATU	Inverse	PU = -0.21xATU + 3.497	0.012	0.696	0.400
H5	PEoU-ATU	Right	PEoU = 0.49xATU + 2.32	0.235	0.067	2.001
H6	ATU-BI	Inverse	ATU = -0.57xBI + 2.68	0.346	0.021	2.623
H7	ROI-BI	Inverse	ROI = -0.93xBI + 0.88	0.737	0.000	6.030
H8	PE-BI	Inverse	PE = -0.93xBI + 0.23	0.769	0.000	6.585
H9	SI-BI	Inverse	SI = -0.78xBI + 1.67	0.511	0.003	3.689
H10	PT-BI	Inverse	PT = -0.98xBI + 0.58	0.730	0.000	5.925
H11	PR-BI	Inverse	PR = -0.67xBI + 2.30	0.438	0.007	3.182

3.2.7 Observed Limitations, Lessons Learnt and Best Practices

The VALU3S project provided a valuable learning opportunity to the UC providers and the technology developers in the realm of validating and verifying complex autonomous systems. Use Case 2, with its intricate combination of several interconnected modules which cannot be tested standalone, posed a challenging task for system development and function testers. Successful execution of this endeavour necessitated close collaboration among engineers with expertise in these diverse fields.

An integrated simulation environment was designed to test and validation of the system under test. Test requirements, test cases, and evaluation scenarios are established for the test and verification process. The simulation-based testing of the teleoperation application not only streamlined the verification and validation process but also resulted in significant time and cost savings.

Testing of network applications: Many tools are aimed at testing network applications, but they hardly fit the specific needs of testing the teleoperation system. The toolchains developed within VALU3S aim both at easy usage and a simple definition of purpose-specific needs (ComFASE focus mainly on the simulation of the wireless networks, and NetLoiter aims at testing the real network connections). The developed toolchains enabled the UC2 provider to verify the teleoperation system in-the-lab in otherwise hard-to-reach conditions.

Cybersecurity assessment: Close collaboration between use case partners provided the opportunity to conduct a security assessment of the system and components. System simulation and mock-up components can be used to address some of the limitations of penetration testing such as the need of real hardware and completed systems before starting to perform testing. Thus, cybersecurity testing can be conducted earlier in the development process. The combination of penetration testing and threat modelling supported with open-source tools provide a cost-efficient and effective way to conduct a security assessment. STRIDE threat modelling framework provides a structured approach to identify potential security threats in systems and it covers the most common types of security threats that can occur. Penetration testing workflow following the PTES framework provides comprehensive coverage in a methodical systematic approach that can provide detailed coverage. The use of free open-source tools for performing threat modelling and penetration testing (Kali Linux) allows for effective, time and cost-reductive workflow. In addition to performing the test case verifying the security requirements, the methodology increases security awareness on the system and component level for stakeholders by exposing threats and security risks and highlighting security best practices and standards that can increase the overall security posture of the system.

The applied demonstrations have yielded the following findings:

- As many modules need to work together to realize a working setup. This required a lot of effort to build the simulation environment that can be used for testing and evaluating the system under test according to the test requirements. To ensure seamless interoperability, it is crucial to conduct comprehensive pre-checks and alignment of interfaces of all the submodules of the simulation environment with each other.
- Initially, the evaluation scenarios were quite high-level, and the test requirements were ambiguous.
- Through close collaboration between UC2 providers and technology providers, a detailed understanding of the system under test was acquired.
- Moreover, the ambiguities in the test requirements and test cases were clarified. This iterative process has resulted in a set of precise test requirements that accurately reflect the test need. These test requirements are now used to verify and validate the system functionality.
- A high-performance computer is required to produce more accurate results, which could be a limitation in terms of the hardware resources available. This was a lesson learned for us during the development, and verification of the simulation environment.
- With the testing environment now prepared, a multitude of testing activities can be conducted in simulations, including extreme test cases. This is particularly advantageous for the UC2 provider, as it enables testing to be carried out earlier in the system development process. Additionally, this approach offers significant time and cost savings, making it an efficient and cost-effective solution.

As presented in Table 3-21, the expert's opinions about car teleoperation indicate that the proposed technology stack is generally accepted. The mean values of PU, MO, PE and BI are over 5 out of 7. PT, ROI, PEOU, SI and ATU are fair. The responses related to CO and PR are quite lower than expected (4.31/7.00). If the potential problems related to the statistical reliability and significance are ignored, the qualitative assessment results show that the experts are hopeful about the novel car teleoperation solution demonstrated in UC2. It is noteworthy that if the users see real-life experiments and actual use of the technology in real settings, their attitude toward using the system will get higher. Additionally, when the developers share more quantified results with the end users and certify their solutions by showing that the teleoperated vehicles are compliant with recent standards and regulations, like UNECE155/156 or ISO 26262.

Table 3-21 Mean and standard deviation of experts' responses to UC2

UC2	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,66	4,77	5,40	4,31	4,80	5,52	4,88	4,53	4,68	4,66	5,33
Std Dev	0,98	1,80	1,05	1,64	1,63	1,67	1,54	1,73	1,61	1,82	1,76

3.3 Use Case 3 - Radar System for ADAS (UC3)

This VALU3S use case "Radar system for ADAS" looks at V&V challenges from the perspective of verification and validation needs of the ADAS IC manufacturers. Due to the faster development cycles and higher complexity of these modern ADAS systems traditional V&V methods have reached their limits.

One of the challenges to be addressed is the necessity to include the verification and validation at the system level from the design to the production of IC components. Future V&V must include the interaction between the sensor IC and all the peripherals at an early stage to save time in the process and to grant the safety and reliability of a productive system. The demonstrator of UC3 focuses on demonstrating a first approach to implementing such a system testing in the V&V workflow (see Figure 3-13). Especially, the testing of radar systems in simulated driving scenarios can play a vital role in overcoming these challenges and thus, will be investigated in the use case.

Figure 3-13 Illustration of an automated car in a simulated urban scenario

3.3.1 V&V challenges

The following V&V challenges are addressed by the demonstrator:

- Customers want to significantly decrease the time needed for development cycles which leads to high pressure on the supplier industry to reduce the invested time for V&V (~which accounts for 80% of the time needed in innovation projects) whereas the complexity and quality standards are rising constantly.

- Consequently, future V&V must include the interaction between the sensor IC and all the peripherals at an early stage to save time in the process and to grant the safety and reliability of a productive system.
- This will require the extension of V&V methods and the introduction of new validation strategies at the system level with a special focus on the IC development which is difficult as they should lead to low interference for running V&V projects.
- Forward integration of automotive tier-1 validation methods must be set up including field tests, road data collection etc.
- The simulation of more complex and realistic scenarios with multiple targets and the injection of environmental data into the processing chain will be important.
- Using a multistep approach becomes necessary, meaning that different levels of Hardware-in-the-Loop tests will be performed due to the complexity of the system and the interaction between the hardware and software.

3.3.2 Contributors

Partners contributing to the UC3**:** NXP-DE, NXP-FR

3.3.3 Contributors' Roles & Evaluation Scenarios

In the demonstrator implementation status report, the workflow of the interaction between the developed tools RSES (a target simulator for radar targets in real-world driving scenarios) and the method of remote testing have already been mapped. The combination of these two parts is the centre of gravity for the innovations in the NXP V&V process as they are enabling the system validation at the IC supplier stage (see the V&V flow in Figure 3-14). Other methods, such as DDMA and smart test evaluations, are accompanying the workflow but this can be accepted as rather cross-sectional method improvements.

In addition to the new tools and methods, a radar system set up is created in which customer antenna boards can be put under test. This enables testing the IC in a system set up and thus, V&V can be also done for system integration testing and hence, forward integrate a V&V step formerly done by Tier 1 supplier. The improved system-wide test setup enabled by the new developments can be seen in Figure 3-15.

Evaluation Scenarios for the Use Case

1. VALU3S_WP1_Automotive_9 - Failure detection of Software and hardware subsystem components – Multi targets are injected at different stages of the system.
2. VALU3S_WP1_Automotive_11 - System performance – System detection range validation within test bench.
3. VALU3S_WP1_Automotive_12 – Advanced Driver-Assistance System (ADAS) must be reliable and has to comply with Safety standards– A high level of car automation leads to higher safety coverage and FTTI. The novel approach has to move from IC to a system in order to guarantee high performance.
4. VALU3S_WP1_Automotive_13 – A system-on-chip (SoC) validation with intensive use of SoC internal self-tests

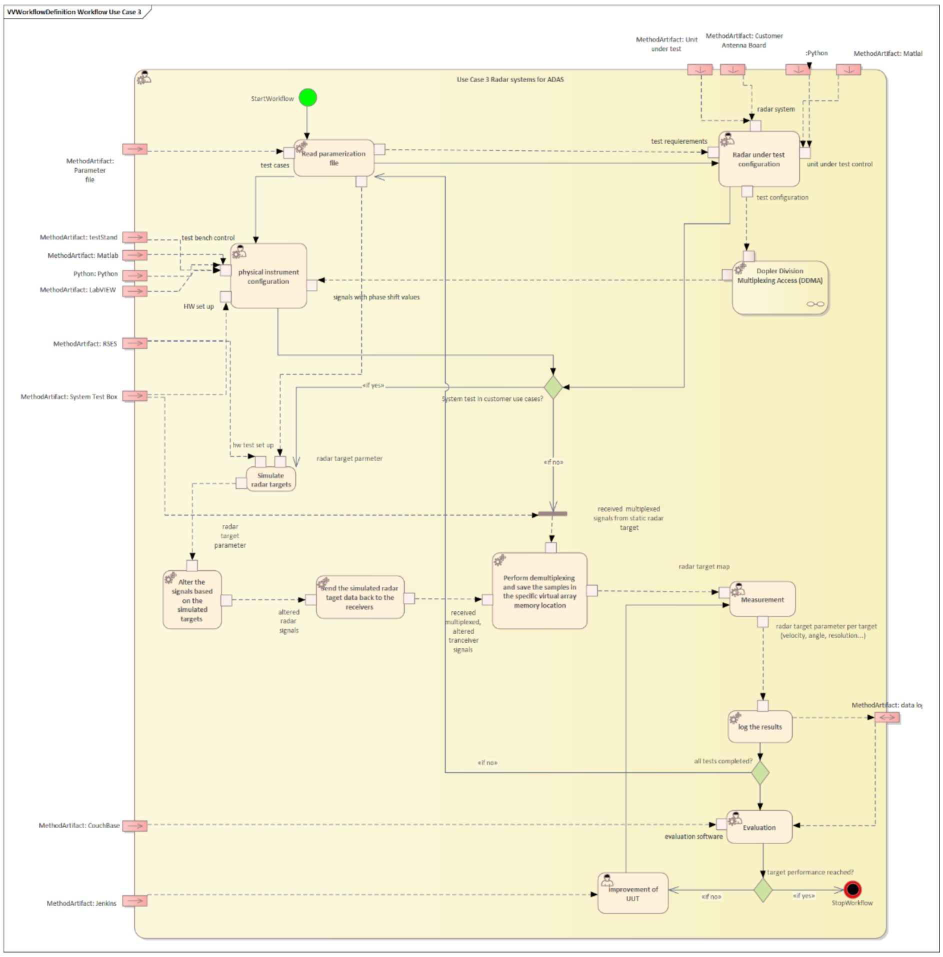

VVWorkflowDefinition Workflow Use Case 3
Use Case 3 Radar systems for ADAS
MethodArtifact: Unit under test
MethodArtifact: Customer Antenna Board
:Python
MethodArtifact: Matlal
StartWorkflow
radar system
MethodArtifact: Parameter file
Read paramerization file
test cases
test requierements
Radar under test configuration
unit under test control
test configuration
MethodArtifact: testStand
MethodArtifact: Matlab
Python: Python
MethodArtifact: LabVIEW
test bench control
physical instrument configuration
HW set up
signals with phase shift values
Dopler Division Multiplexing Access (DDMA)
MethodArtifact: RSES
MethodArtifact: System Test Box
«if yes»
System test in customer use cases?
hw test set up
radar target parmeter
«if no»
Simulate radar targets
received multiplexed signals from static radar target
radar target parameter
Alter the signals based on the simulated targets
altered radar signals
Send the simulated radar taget data back to the receivers
received multiplexed, altered tranceiver signals
Perform demultiplexing and save the samples in the specific virtual array memory location
radar target map
Measurement
radar target parameter per target (velocity, angle, resolution...)
MethodArtifact: data lo
log the results
«if no»
all tests completed?
MethodArtifact: CouchBase
evaluation software
Evaluation
target performance reached?
MethodArtifact: Jenkins
improvement of UUT
«if no»
«if yes»
StopWorkflow

Figure 3-14 V&V workflow for UC3

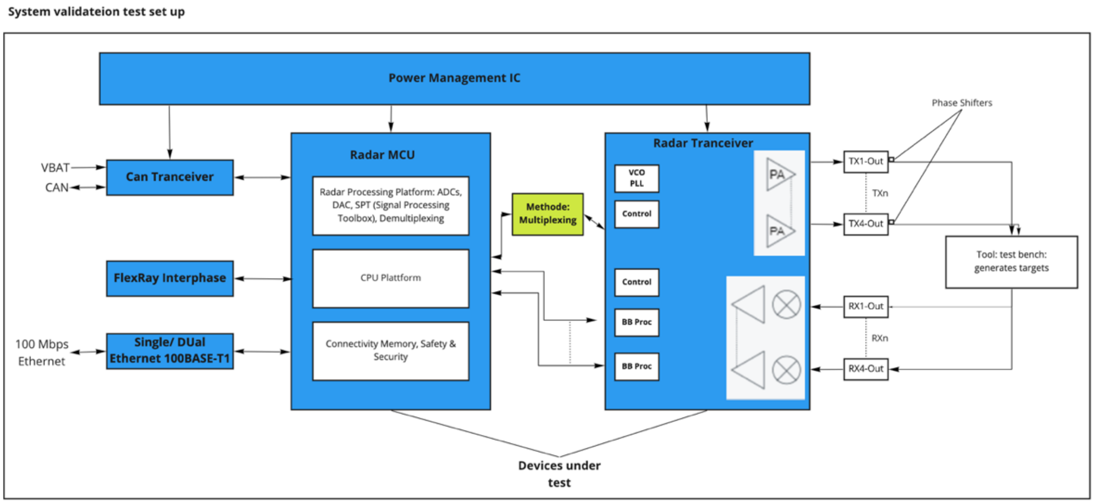


Figure 3-15 System validation test setup for UC3

3.3.4 Demonstration

Two toolchains enable system tests (normally often performed by Tier-1) at the IC supplier level. This will increase the test coverage and reduce the time for system validation that directly benefits the customers. This can be demonstrated by showing the set-up and the improved test results (see Table 3-22).

Table 3-22 Overview of demonstration prepared by UC3 partners.

Item #	Demonstration name	Description/Purpose	Format	Responsible
1	Remote-controlled radar target simulation and validation	The RSES is utilised to simulate various real-world driving scenarios. However, due to the high cost and immobility of the equipment, we aim to demonstrate our newly developed remote validation process that can be performed from the lab in Munich while being based in Porto. Our goal is to allow global competence centres to use the hardware validation equipment in the future, making the validation process more resilient to external factors, such as a pandemic situation. Additionally, this approach reduces the overall cost and time required for validation, which accounts for approximately 80% of the total radar development cycle cost. The planned scenarios simulate different moving targets, validating a radar chip in a system environment with varying speeds, angles, ranges, and temperatures.	Lead Demonstrator PowerPoint/ Poster/ Video	NXP-DE
2	Validation of silicon chips integrated into a corner radar system	As the future of autonomous driving and the success of automotive OEMs rely heavily on advancements in corner radar systems, this demonstrator aims to showcase the impact of VALU3S by improving the methods and tools used for validation. Without these improvements, complex autonomous systems may not be feasible in the near future. The hardware demonstrator particularly focuses on detecting moving objects and presenting their characteristics on a user-friendly GUI.	Complementary Demonstrator Radar validation set up (Video and presentation)	NXP-DE

3.3.5 Quantitative Results

UC3 improvements as a factor to the baseline evaluation criteria are itemised below and also depicted in Figure 3-16.

Evaluation criteria for SCP:

1. Eval_SCP_2 – Number of Safety/Security Requirement Violations: The detection of bugs which could cause safety violations is improved by the RSES and System Test Box as test coverage also covers traffic scenarios.
    - Baseline: Safety issues were going through one validation process at Tier 1.
    - Improvement: Safety issues can go through two test processes. The percentage of how this increases the prevention of an absolute number of safety violations cannot be

measured exactly at the moment but as every prevented violation counts, the impact can be quite substantial.

Evaluation criteria of the V&V process:

1. Eval_VV_2 – Coverage of Test Set: By integrating the RSES it is possible to validate radar systems in traffic scenarios.
    - Baseline: possible traffic scenarios were not validated at Tier 2.
    - Improvement: multiple targets with multiple parameters can be measured in traffic scenarios at Tier 2 (IC supplier level). As the traffic scenarios and test cases are implemented for demonstration purposes the concrete amount of additional test cases in daily operation can ultimately be defined after the project.
2. Eval_VV_8 – Effort Needed for Test 1: The remote testing capabilities developed in this project enable optimization of radar chip parameters using a teleoperated testing setup to work off-site. This is especially impactful if there are restrictions to collaborate in person such as during the pandemic.
    - Baseline: workers needed to be on-site to perform V&V; Effort = 100%
    - Improvement: workers can be off-site if required due to restrictions or if it is desired. Further, the focus time of employees could be increased leading to a boost in performance by conducting remote work. Effort = 87%
3. Eval_VV_8 – Effort Needed for Test 2 By integrating the RSES it is possible to validate radar systems in traffic scenarios.
    - Baseline: Validating radar systems took a great amount of effort as the radars had to be mounted on a car and traffic scenarios had to be simulated by staff, and governmental approvals need to be obtained; Effort = 100%
    - Improvement: traffic scenarios can be simulated by the RSES and thus, test effort can be reduced; Effort = 80%

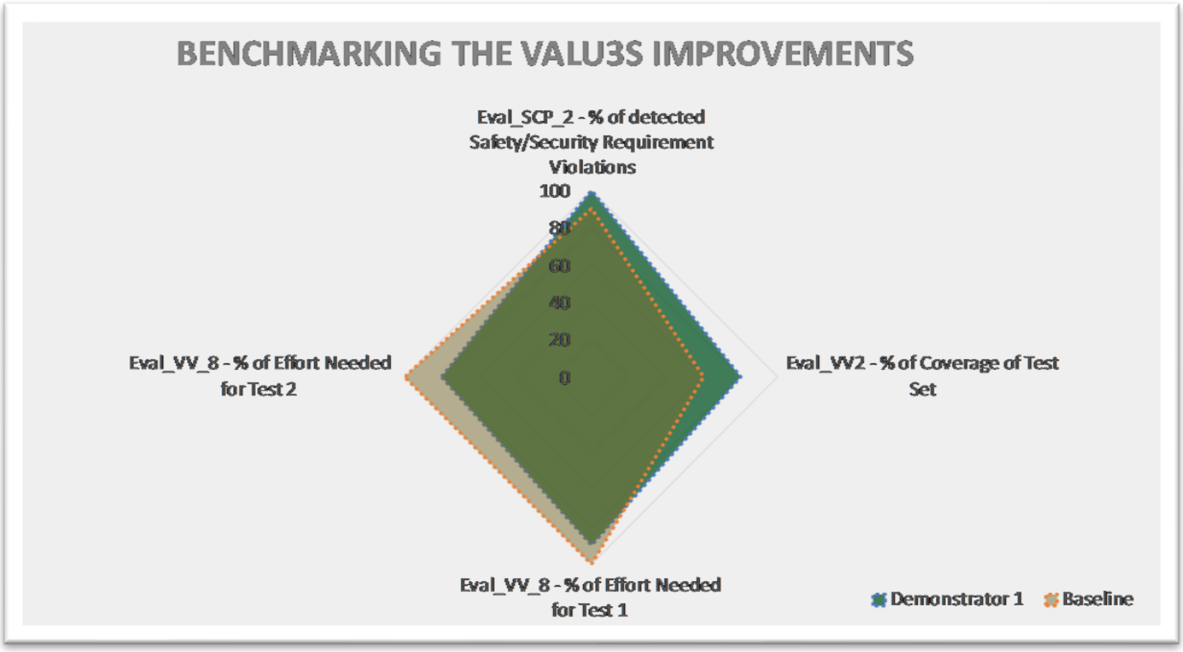


Figure 3-16 Benchmarking of the VALU3S Improvements

3.3.6 Qualitative Results

Demonstrator 1: Remote-controlled radar target simulation and validation

Participants Profile: QAM is applied to 10 subjects (10 Males) aged in the range of 24-34. The education level is relatively high as the subject pool is composed of 1 PhD researcher and 9 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "R&D engineers, researchers, Q&A, system engineers, software/hardware engineers, etc." having experience in the fields of "cyber-physical systems, automotive, health, semantic web, IoT, digital twin etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-23. The results show that PU, PEOU, MO, CO and BI are not highly correlated with each other.

Table 3-23. UC3 – Demonstrator 1 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PUSF	1										
PEOU	0.358	1									
MO	0.055	0.146	1								
CO	-0.383	-0.030	0.204	1							
ROI	0.228	-0.174	0.226	0.152	1						
PE	0.220	0.475	0.247	0.412	0.595	1					
PT	0.124	0.389	-0.179	-0.045	0.430	0.540	1				
PR	-0.031	0.067	-0.375	0.369	0.069	0.473	0.472	1			
SI	-0.159	0.158	-0.458	0.166	0.157	0.478	0.173	0.329	1		
ATU	0.234	0.339	-0.278	-0.015	0.061	0.575	0.295	0.465	0.741	1	
BI	-0.085	0.191	-0.585	-0.308	-0.170	0.160	0.280	0.364	0.522	0.553	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-24, the majority of the questions asked to subjects are sufficiently reliable. However, answers related to questions considering MO and BI seem not reliable enough. The main reason can be the concentration level of subjects might not be high or the responses to MO and BI are not correlated.

Table 3-24. UC3 - Demonstrator 1 Reliability Analysis

Cronbach-alpha values	
PU	0.522
PEOU	0.469
MO	-2.000
CO	0.359
ROI	0.364
PE	0.286
PT	0.230
PR	0.750
SI	0.566
ATU	0.510
BI	-1.064

Regression Analysis: Finally, regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-25. For this demonstrator, CO-PU and ROI-BI are proportionally right while the other pairs influence each other with an inversely proportional relation.

Table 3-25. UC3 - Demonstrator 1 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Inverse	MO = -0.12xPU + 4.80	0.055	0.880	0.156
H2	CO-PU	Right	CO = 0.61xPU + 9.04	0.382	0.275	-1.171
H3	PEoU-PU	Inverse	PEoU = -1.03xPU – 0.33	0.128	0.310	1.085
H4	PU-ATU	Inverse	PU = -0.15xATU + 4.41	0.054	0.516	0.680
H5	PEoU-ATU	Inverse	PEoU = -0.65xATU + 1.59	0.115	0.338	1.020
H6	ATU-BI	Inverse	ATU = -0.49xBI + 3.17	0.305	0.098	1.876
H7	ROI-BI	Right	ROI = 0.20xBI + 6.91	0.029	0.639	-0.488
H8	PE-BI	Inverse	PE = -0.23xBI + 3.40	0.026	0.659	0.459
H9	SI-BI	Inverse	SI = -0.52xBI + 2.94	0.272	0.122	1.731
H10	PT-BI	Inverse	PT = -0.34xBI + 3.78	0.078	0.434	0.823
H11	PR-BI	Inverse	PR = -0.21xBI + 4.69	0.132	0.301	1.105

Demonstrator 2: Corner Radar validation

Participants Profile: QAM is applied to 10 subjects (10 Males) aged in the range of 24-44. The education level is relatively high as the subject pool is composed of 1 Post-Doc or higher degree and 1 PhD researcher and 8 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "R&D engineers, Q&A, principle engineers, software/hardware engineers, etc." having experience in the fields of "radar, semantic web, IoT, digital twin etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-26. The results show that SI and ATU are not highly correlated with the other constructs whereas PR seems less correlated with SI, AUT and BI.

Table 3-26. UC3 – Demonstrator 2 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.638	1									
MO	0.414	0.292	1								
CO	0.560	0.426	0.226	1							
ROI	0.126	-0.309	0.128	0.169	1						
PE	0.276	0.286	0.240	0.123	0.140	1					
PT	0.568	0.629	0.428	0.116	-0.255	0.071	1				
PR	0.431	0.456	0.599	0.332	0.326	0.076	0.560	1			
SI	0.234	-0.049	-0.491	-0.156	-0.253	-0.152	0.292	-0.135	1		
ATU	0.038	-0.155	-0.356	-0.016	-0.243	-0.527	0.078	-0.366	0.498	1	
BI	0.628	0.332	0.140	0.606	-0.039	0.081	0.323	-0.061	0.168	0.596	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-27, the majority of the questions asked to subjects are sufficiently reliable. However, answers related to questions considering PR, SI and BI seem not reliable enough.

Table 3-27. UC3 – Demonstrator 2 Reliability Analysis

Cronbach-alpha values	
PU	0.341
PEOU	0.350
MO	0.375
CO	0.400
ROI	-0.682
PE	0.072
PT	0.375
PR	-1.295
SI	-1.125
ATU	0.435
BI	-2.000

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-28. For this demonstrator, there exists a right proportional relation between PEOU-ATU, ROI-BI and PR-BII while the other pairs influence each other with an inverse proportional relation.

Table 3-28. UC3 Demonstrator 2 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Inverse	MO = -0.52xPU + 2.72	0.172	0.234	1.288
H2	CO-PU	Inverse	CO = -0.76xPU + 1.30	0.313	0.092	1.911
H3	PEoU-PU	Inverse	PEoU = -1.04xPU – 0.29	0.407	0.047	2.345
H4	PU-ATU	Inverse	PU = -0.03xATU + 5.62	0.001	0.918	0.106
H5	PEoU-ATU	Right	PEoU = 0.19xATU + 6.91	0.024	0.668	-0.445
H6	ATU-BI	Inverse	ATU = -0.58xBI +2.66	0.355	0.069	2.099
H7	ROI-BI	Right	ROI = 0.05xBI + 6.30	0.002	0.914	-0.111
H8	PE-BI	Inverse	PE = -0.09xBI + 5.45	0.007	0.824	0.229
H9	SI-BI	Inverse	SI = -0.25xBI + 4.57	0.028	0.642	0.482
H10	PT-BI	Inverse	PT = -0.39xBI + 3.76	0.105	0.362	0.966
H11	PR-BI	Right	PR = 0.05xBI + 6.30	0.004	0.866	-0.174

3.3.7 Observed Limitations, Lessons Learnt and Best Practices

Throughout this research project, we have encountered several observed limitations, lessons learned, and best practices that have shaped our understanding of verification and validation (V&V) in the automotive industry.

Lessons Learned and Best Practices:

1. Early-stage interaction: Combining sensor ICs and peripherals for system testing in early-stage V&V is streamlining development and thus, saving time.
2. Remote testing capabilities: Developing remote testing capabilities has proven invaluable for enabling efficient, cost-effective V&V processes that can adapt to various work scenarios.
3. Real-world scenario simulation: Utilizing the Radar System Environment Simulator (RSES) has demonstrated the importance of simulating real-world traffic scenarios, improving test coverage and system performance while reducing the need for costly testing setups.
4. Continuous improvement: The project emphasized the need for ongoing refinement and expansion of V&V methods to keep pace with the evolving automotive industry.
5. Encourage effective collaboration and communication among team members and stakeholders to facilitate innovation and the development of novel solutions.

Observed Limitations:

There are four main limitations to the implemented simulation-based approach:

1. For the usage of joint simulation models with partners in European R&D projects, there is still a need to share a considerable amount of background IP. For industry partners, this IP is often the most critical part of the business and thus, it is hard to find appropriate ways to develop joint solutions. In the future, methods have to be found which enable collaboration without sharing critical background IP at any given time when working jointly on simulation-based V&V.
2. Through this simulation-based approach, the qualitative outcome of the IC manufacturer V&V process is increased and the development cycle of ADAS functions will be fastened when looking at the whole supply chain. Nevertheless, the effort spent for V&V might raise IC manufacturers significantly in the future in case Tier 1 suppliers demand more and more V&V steps to be performed at this stage already. The increased efforts will come especially from programming new traffic scenarios and the required individual HW set-up for individual customer antenna boards. This might be dealt with by closer collaboration with Tier1s on V&V processes in the future.
3. The required individual HW set-up/customer antenna board still requires a lot of manual work which hinders automation.
4. Further test equipment has to be included in the HW test set-up to enable more tests without local intervention.

By acknowledging the observed limitations and learning from the lessons and best practices, we can pave the way for the development of more efficient and effective V&V processes, ultimately contributing to safer and more reliable automotive systems.

As seen in Table 3-29 and Table 3-30, for both demonstrations experts' opinions are more or less the same. There is a strong motivation to use the V&V tools for ADAS systems as this has become a natural need of the automotive industry with recent advancements in autonomous driving. The radar-supported ADAS and the utilisation of AI-powered traffic scenario generation have a positive impact on increasing the acceptance of the technology. Since the targeted area is industry-driven, because automotive companies tend to invest in ADAS and traffic scenario management covering the V2X applications and smart solutions to decrease the safety problems in daily driving, UC3 outcomes may find a chance in the market. Moreover, social influence is well addressed in this use case as the experts see that if the proposed V&V solutions are integrated into the design and implementation of autonomous vehicles and traffic management systems, public acceptance can be higher. Nevertheless, there is still much work to increase the overall acceptance to the level of 6 or higher (out of 7.00), as the qualitative assessment has been implemented with a relatively less number of experts. This may cause some not fully interpreted results or potential mismatches with the real status of the technology acceptance.

Table 3-29 Mean and standard deviation of experts' responses to UC3 - Demonstration 1

UC3-Demonstrator-1	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,49	5,67	5,79	5,80	5,70	5,84	5,80	5,19	5,37	5,27	5,76
Std Dev	0,88	0,31	0,41	0,55	0,44	0,36	0,43	0,93	0,52	0,59	0,52

Table 3-30 Mean and standard deviation of experts' responses to UC3 - Demonstration 2

UC3/ Demonstrator-2	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,82	5,83	5,93	5,94	5,80	5,91	5,74	5,71	5,79	5,79	6,01
Std Dev	0,75	0,46	0,60	0,55	0,43	0,47	0,45	0,65	0,37	0,57	0,55

3.4 Use Case 4 - Human-Robot-Interaction in Semi-Automatic Assembly Processes (UC4)

UC4 is based on a Human-Robot-Interaction (HRI) process taking place on the shop floor of a manufacturing-like environment. The process itself involves the execution of assembly tasks by human workers, focusing on the assembly of transformer units which consist of multiple parts. HRI systems have to manage the coordination between humans and robots according to the requirements for collaborative industrial robot systems as defined in ISO 10218-2:2011 [29], ISO/TS 15066 [30], and IEC 61508:2010 [22].

3.4.1 V&V Challenges

Recognising and localising objects of interest and tracking them in the workspace to allow a smooth flow of interaction between humans and robots are of vital importance, particularly in the scope of the evaluation and test scenarios for failure injection. The envisaged use case is set up as a physical demonstrator, consisting of different IoT building blocks as sketched in Figure 3-17. These include one six-axis collaborative robot (Fanuc CR35ia) and one six-axis non-collaborative robot for medium payloads (Fanuc R2000i), an IPC (Industrial PC) with a soft-PLC (Siemens S7), an IoT gateway, a switch/hub, and two functional safe microcontrollers for doing the pre-processing of sensor data.

The IoT components are responsible for generating the positioning sensor data include a wearable, fibre-optic sensing system and an Ultra-wideband (UWB) based indoor localization system. While the fibre-optic sensor system calculates the orientation of the body limbs (arms, hands, and shoulders) of the human, the localisation system localizes the relative position of the human within the workspace. The sensor data streams of both systems are communicated via communication protocols such as UDP, MQTT, and OPC-UA. The communication among the IoT components, which are directly connected to the IPC, is performed via PROFINET connection. The corresponding test cases will consider specific failure injection methods in both physical and virtual space as defined in the evaluation scenario spreadsheet.

At the component level, it is planned to focus on the manipulation of data generated by three IoT sensor components, named from the fibre-optic sensor system as the UWB localization system, and the proximity sensor. Additionally, it is foreseen to inject failures directly into the PLC device. For the full deployment of the HRI use case, a simulation model of the behaviour and interaction of the physical components according to Figure 3-17 will be implemented in CIROS Studio and connected to FERAL - a simulation framework for virtual validation – via the communication protocol MQTT.

The main sensor systems (fibre-optic sensor system and UWB localization system) are not available as virtual models within the preferred simulation environment, thus will be complemented through approaches based on hardware-in-the-loop (HiL), as well as the creation of complementary virtual models through the support of FRAUNHOFER. To realise a failure detection and diagnosis as proposed in this use case, it is foreseen to access and analyse the manipulated data streams emerging from the IoT components by implementing a sensor data stream pipeline based on the usage of appropriate data mining and machine learning techniques. The objective here is to detect and classify failures on the cyber-physical level for HRI applications.

3.4.2 Contributors

Partners contributing to the UC: PUMACY, FRAUNHOFER IESE, UCLM.

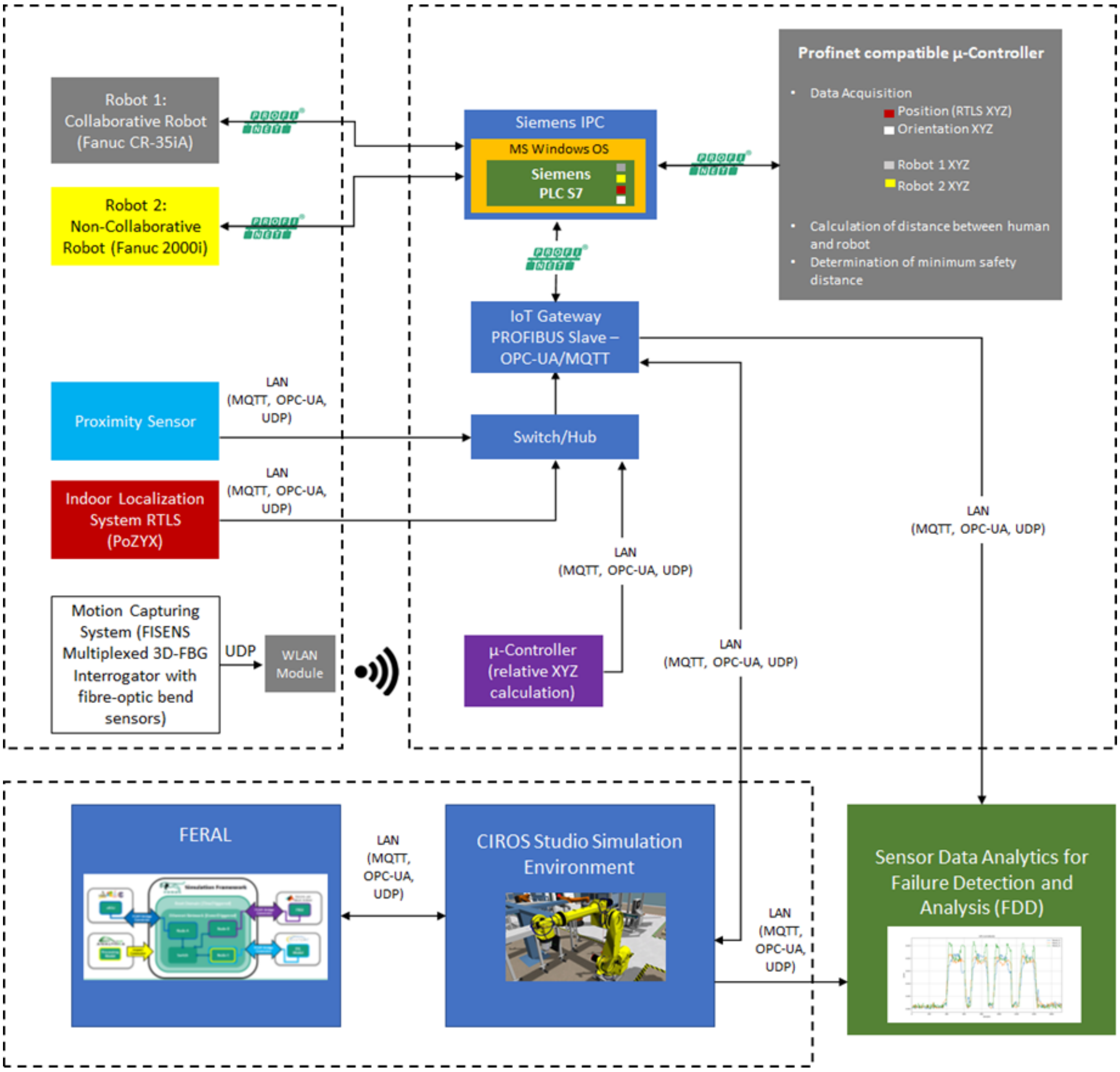

Robot 1: Collaborative Robot (Fanuc CR-35iA)
Robot 2: Non-Collaborative Robot (Fanuc 2000i)
PROFI NET
Siemens IPC
MS Windows OS
Siemens PLC S7
Profinet compatible µ-Controller
Data Acquisition
Position (RTLS XYZ)
Orientation XYZ
Robot 1 XYZ
Robot 2 XYZ
Calculation of distance between human and robot
Determination of minimum safety distance
IoT Gateway PROFIBUS Slave – OPC-UA/MQTT
Proximity Sensor
LAN (MQTT, OPC-UA, UDP)
Switch/Hub
Indoor Localization System RTLS (PoZYX)
LAN (MQTT, OPC-UA, UDP)
LAN (MQTT, OPC-UA, UDP)
LAN (MQTT, OPC-UA, UDP)
Motion Capturing System (FISENS Multiplexed 3D-FBG Interrogator with fibre-optic bend sensors)
UDP
WLAN Module
µ-Controller (relative XYZ calculation)
LAN (MQTT, OPC-UA, UDP)
FERAL
LAN (MQTT, OPC-UA, UDP)
CIROS Studio Simulation Environment
LAN (MQTT, OPC-UA, UDP)
Sensor Data Analytics for Failure Detection and Analysis (FDD)

Figure 3-17: Architecture of HRI Demonstrator

3.4.3 Contributors' Roles & Evaluation Scenario

The goal of the UC4 is to use and combine state-of-the-art simulation methods and tools to validate the architecture design of a complex manufacturing plant. This workflow focuses on one part of the use case toolchain, the FERAL framework [33], which is used for coupling different simulators and simulation components such as the robot model in CIROS studio or sensor models. Within UC4 a Combined Virtual Validation and Failure Detection Diagnosis will be applied which is described in detail in the previous deliverable (see D5.5 [11]). The applied combined validation approach requires several documents from the system development process (system specification, system architecture design model, and fault model) and existing artefacts of the test and validation activities (existing test scripts, tool adapters as well as configurations and trained ML models). Outputs comprise different reports, the updated set of test cases and test scripts and the updated test artefacts like test scripts, tool adapters and ML models.

The combined method of Data Analytics/ML and Virtual Validation is designed and developed to detect failures in the Simulation. ML-Pipeline is used as enhancement/improvement of the "Failure Detection and Diagnosis (FDD)" Method that is applied in UC4 as part of the toolchain. The tool has been used in combination with FERAL for Combined Virtual Validation and Failure Detection Diagnosis.

The list of evaluation scenarios is as follows:

1. VALU3S_WP1_Industrial_10 - Localization of Human
2. VALU3S_WP1_Industrial_11 - Handling and gripping of product/parts
3. VALU3S_WP1_Industrial_12 - Knocking off product/part from robot gripper by a human worker.
4. VALU3S_WP1_Industrial_13 - Corruption of input/output signal at robot gripper
5. VALU3S_WP1_Industrial_14 - Data manipulation in human-robot-interaction

Evaluation Criteria for SCP & VV

The updated evaluation criteria for SCP and V&V processes are summarised in Table 3-31.

Table 3-31 Updated evaluation criteria.

Evaluation criteria for SCP:	Description of evaluation
Eval_SCP_3 – Number of Malicious Attacks and Faults Detected	This metric has been used for evaluating the results from simulation testing that is performed in UC4
Eval_SCP_4 – Metrics to Evaluate AI/ML Algorithms	Identification of faults through AI data stream analytics has been evaluated by the accuracy of the AI/ML model which depends on the data quality and quantity.
Eval_SCP_10 – Software Fault Tolerance Robustness	Metric indicating the portion of defined software faults the system can handle while providing its specified functionality.
Eval_SCP_11 – Simulation-level System Robustness	This criterion has been used to analyse and assess the robustness of the system architecture design and the composition of simulation components of a complex manufacturing plant. Fault injection is applied on the levels of communication middleware and component behaviour.
Evaluation criteria of the V&V process:	**Description of evaluation**
Eval_VV_3 – Number of Test Cases	The specified test cases focus on different aspects mainly on safety but also on security. As these are the first test carried out in such a scenario it is expected that not all test cases and targeted measurements can be conducted as set out and needed to be adapted.

Eval_VV_4 – Effort for Test Creation	UC4 covers lots of effort for preparation, setup and conduction of testing. Within the project, this metric has been measured at least concerning preparation and testing.
Eval_VV_5 – Joint Management of SCP Requirements	UC4 mainly deals with safety-related issues. Accordingly, the equations above have been applied within this focus. Faults and attacks are injected into CIROS simulation and data communication using the FERAL framework. The use of FERAL enables the injection of different fault types during execution, as well as the construction and evaluation of validation scenarios that include multiple SCP requirements.
Eval_VV_6 – Cost of Finding and Fixing a Coding Bug	UC4 covers scripting activities and thus the criterion has been used for roughly analysing costs. It is unlikely to measure exact time/cost, but it is sufficient to obtain an approximate value.
Eval_VV_8 – Effort Needed for Test	The criterion has been used to measure the utilization of effort for setting up tests and preparation of the tests namely the development of models, and the connection of tools / set up of the toolchain at all partners.
Eval_VV_10 – Reduced Cost and Time for Work on Certification Process and Functional Safety	This metric has been used for evaluating the results from simulation testing that is performed in UC4

An overview of individual partners' contributions within evaluation scenarios can be found in Table 3-32.

Table 3-32 Overview of contribution to evaluation scenarios by UC4 partners

Evaluation Scenario	PUMACY	FRAUNHOFER IESE	UCLM
VALU3S_WP1_Industrial_10	X	X	X
VALU3S_WP1_Industrial_11	X	X	
VALU3S_WP1_Industrial_12	X	X	
VALU3S_WP1_Industrial_13	X	X	
VALU3S_WP1_Industrial_14	X	X	X

Figure 3-18 illustrates the V&V tools that are being developed and planned to be used/demonstrated in this use case as well as the V&V methods associated with the tools.

FRAUNHOFER has been supporting the UC4 demonstrator by extending the simulation framework FERAL regarding its support for domain-specific communication protocols (here: MQTT), the connection to the simulator CIROS studio via a Python interface, and a fault injection component that enables fault injection into (1) the simulation models of the communication protocol MQTT and (2) the robot simulation model within CIROS studio. UCLM provides a virtual reality (VR) interface that allows the interaction of a real human operator with the robot simulator. This VR interface connects to the rest of the system via MQTT and FERAL.

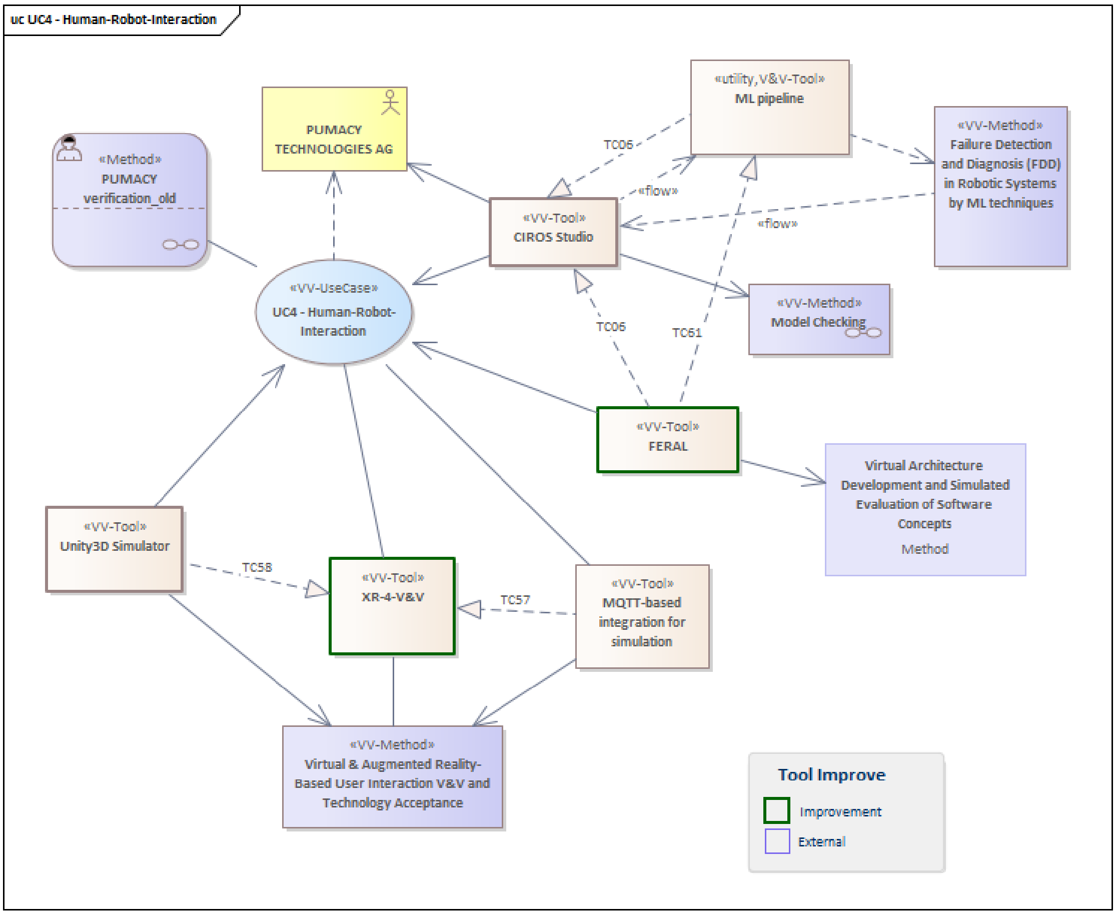


Figure 3-18 Tools and Methods for UC4 - Human-Robot-Interaction diagram

3.4.4 Demonstration

Demonstration for UC4 was planned by individual partners. Therefore, multiple demonstrable items are covering defined challenges and scenarios, and partially cover VALU3S dimensions (see Table 3-33).

Table 3-33: Overview of demonstration prepared by UC4 partners.

Item #	Demonstration name	Description/Purpose	Format	Responsible
1	Handling and gripping of product/ parts Fault injection using FERAL tool into CIROS assembly simulation / model	The demonstrator presents how the fault tolerance of the architecture design, where faults are injected at two levels, is checked. On the one hand, communication faults are injected into the virtual network component that implements the MQTT protocol. Fault types are e.g., message delay and message loss. On the other hand, different implementation faults have been defined and injected for the robot system model, such as faults that trigger misbehaviour of the air pressure gripper, which is a part of the robot arm. A special interface for the robot model has been developed for the activation and control of implementation faults: • "Remove product from simulation" (failure simulation) - (robot should stop immediately) and • "Do not grip in simulation" (failure simulation) A virtual simulation environment and dedicated test cases are constructed and used to run and evaluate validation scenarios. The tool executes the simulation scenarios and controls the simulation components and the data flow between them. Log data from the simulation run is collected and provided. A validation report is created after the execution and an evaluation of the simulation scenarios	Lead demo	PUMACY, FRAUNHOFER IESE
2	Machine Learning Pipeline	In the industrial robotics domain, faults have the potential to affect the efficiency of the underlying process, namely causing failures of internal physical components (e.g., robot, IPC, actuators), or even compromising the safety of humans interacting with the robot. When detecting a fault, usually a diagnosis process is induced in order to identify which internal components are involved. Enhancing failure detection by Machine Learning techniques analyzes real data and manipulated data streams in order to detect anomalies in the to be process. This has been achieved through process mining and pattern recognition in data from the original assembly process used to develop and train the ML model. The resulting model is used for failure detection in manipulated data streams in defined test scenarios.	Lead demo	PUMACY

Item #	Demonstration name	Description/Purpose	Format	Responsible
3	Virtual & augmented reality-based user interaction V&V	An immersive virtual reality application, namely XR-4-V&V, has been developed to facilitate early human-robot collaboration. This system allows human workers to collaborate with industrial robots in a simulated environment through the use of a head-mounted display. XR-4-V&V is developed using the Unity3D platform and focuses on handling only human interaction. Meanwhile, the robot simulation model runs on the CIROS studio, and the communication between the two is facilitated by FERAL, utilizing MQTT for message exchange. To improve realism, the XR-4-V&V provides a 3D representation of the working environment, including the robot, which enables the execution of assembly tasks for transformer units considered for this use case. The human worker can observe the robot's movements while it grips the transformer parts and brings them closer. After the robot completes its task, the human worker can assemble the parts and wait for the robot to retrieve them. Throughout the entire process, the human operator's behaviour can be monitored, enabling the analysis of human factors and technology acceptance before the system's full deployment. At the moment of writing, this is the plan regarding this demonstrator, but it is worth noting that this plan is conditioned to the successful integration of the three tools (XR-4-V&V, FERAL and CIROS).	XR-4-V&V tool demo	UCLM

3.4.5 Quantitative Results

Demonstrator-1: Handling and gripping of product/parts Fault injection using FERAL tool into CIROS assembly simulation/model

Table 3-34 reflects and explains the quantitative results from the demonstrator and Figure 3-19 displays the improvements accordingly.

Table 3-34 Evaluation criteria UC4

	Eval_VV_3 Number of Test Cases	Eval_VV_5 Joint Management of SCP Requirements	Eval_SCP_10 Simulation-Level System Robustness	Eval_SCP_3 Number of Malicious Faults Detected
Demonstrator UC4 Gripping Scenarios	8	6	8	6
Baseline UC4 Gripping Scenarios	6	3	5	5
Argumentation on the improvement obtained	In the existing approach, a set of test cases and configuration have been manually designed and implemented, which cover the main scenarios.	Test cases have been separately created for each single test goal.	The existing approach considered dedicated fault injection test for covering faults.	A larger part of the remaining faults could be detected by the existing approach, which aimed at manually creating test cases from specifications.
	By using automated and virtual test approaches, a larger set of test cases for the selected target could be generated and executed, which provided a slightly increased coverage of system requirements.	By using dedicated simulation models and fault injection, several functional and non-functional requirements (esp. Fault tolerance and robustness) can be checked.	The new approach enables the consideration of fault models and the creation of targeted fault injection tests that include faulty behaviour or faulty communication that were derived from the fault model.	With the new approach, slightly more residual errors could be detected and eliminated. The fault types described by the fault model are systematically checked in the test. Since the fault model focusses only on selected fault several residual faults of these types are currently no considered. By extending the fault model, the detection rate of additional fault types can be increased.

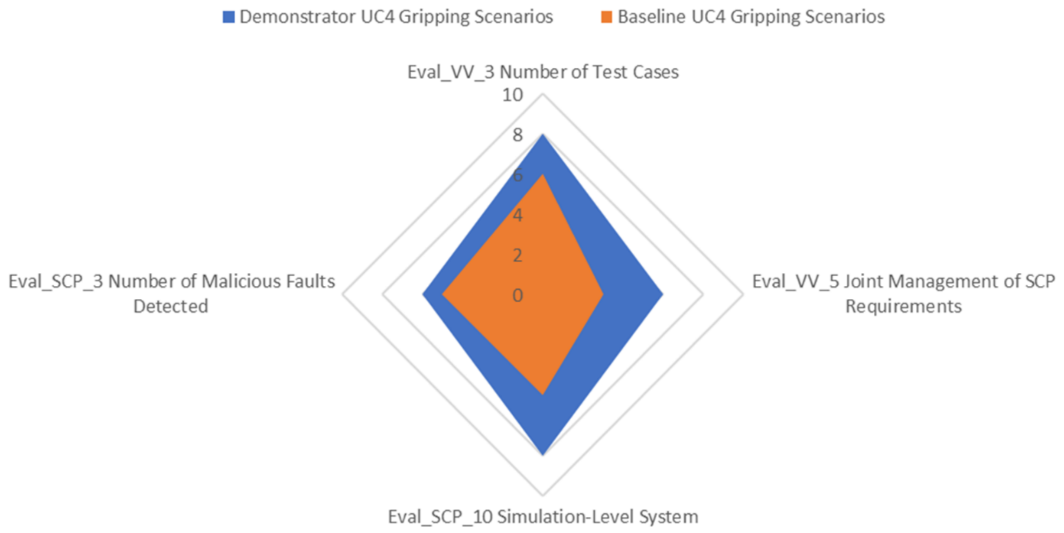


Figure 3-19 UC4 Quantitative improvements applying the FERAL tool

Demonstrator-2: ML-Pipeline

A model's accuracy is dependent on the quality and volume of the available data. Both increase over time and help to train and improve the model to detect additional faults and events. The accuracy has been improved significantly up to 8,5 (85%) currently and is targeted to reach 95-98% after additional training cycles (cf. SCR_4_accuracy in Figure 3-20. The number of faults that are injected and which can be detected have been increased in parallel (SCP_3_faults). In addition, the number of executed test cases had been increased and the jointly managed requirements too.

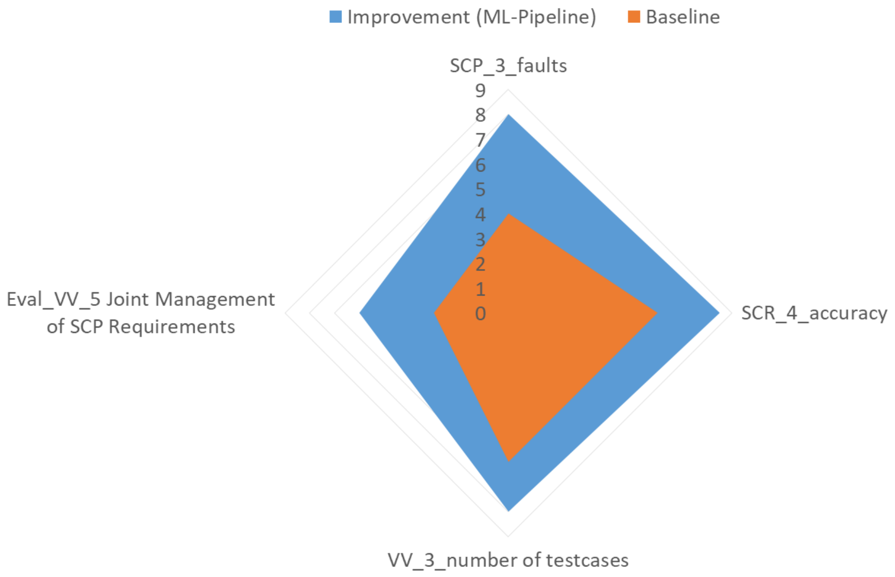


Figure 3-20 UC4 Quantitative improvements applying ML-Pipeline

3.4.6 Qualitative Results

Demonstrator 1: Handling and gripping of product/parts Fault injection using FERAL tool into CIROS assembly simulation/model

Participants Profile: QAM is applied to 12 subjects (12 female) aged in the range of 24-44. The education level is relatively high as the subject pool is composed of 1 Post-Doc or higher degree and 2 PhD researchers and 9 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "lecturers, R&D engineers, ML engineers, Q&A, embedded system engineers, managers, software/hardware engineers etc." having experience in the fields of "AI/ML, data analytics, radar, semantic web, IoT, digital twin, product lifecycle management etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-35. The results show that all constructs are correlated with each other.

Table 3-35. UC4 – Demonstrator 1 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.702	1									
MO	0.751	0.868	1								
CO	0.640	0.958	0.825	1							
ROI	0.463	0.825	0.825	0.840	1						
PE	0.723	0.903	0.884	0.876	0.905	1					
PT	0.413	0.817	0.779	0.796	0.925	0.809	1				
PR	0.615	0.816	0.695	0.891	0.788	0.797	0.774	1			
SI	0.595	0.728	0.717	0.762	0.887	0.888	0.784	0.849	1		
ATU	0.668	0.823	0.823	0.768	0.877	0.891	0.774	0.764	0.882	1	
BI	0.698	0.821	0.919	0.743	0.889	0.919	0.825	0.702	0.842	0.953	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-36, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-36. UC4 - Demonstrator 1 Reliability Analysis

Cronbach-alpha values	
PU	0.212
PEOU	0.423
MO	0.542
CO	0.214
ROI	0.483
PE	0.592
PT	0.587
PR	0.605
SI	0.873
ATU	0.243
BI	0.636

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-37. For this demonstrator, there exists a right proportional relation between all constructs influencing each other in the same direction.

Table 3-37. UC4 - Demonstrator 1 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = -0.48xPU + 3.10	0.564	0.005	3.599
H2	CO-PU	Right	CO = 0.34xPU + 3.98	0.409	0.025	2.633
H3	PEoU-PU	Right	PEoU = 0.40xPU +3.70	0.493	0.010	3.119
H4	PU-ATU	Right	PU = 1.40xATU – 3.13	0.447	0.017	2.842
H5	PEoU-ATU	Right	PEoU = 0.98xATU + 0.06	0.677	0.001	4.575
H6	ATU-BI	Right	ATU = 0.89xBI + 0.80	0.908	~0	9.933
H7	ROI-BI	Right	ROI = 1.00xBI + 0	0.791	~0	6.148
H8	PE-BI	Right	PE = 1.05xBI – 0.35	0.843	~0	7.350
H9	SI-BI	Right	SI = 0.68xBI + 1.97	0.710	~0	4.943
H10	PT-BI	Right	PT = 0.88xBI + 0.51	0.681	~0	4.622
H11	PR-BI	Right	PR = 0.69xBI + 1.52	0.493	0.011	3.118

Demonstrator 2: ML-Pipeline

Participants Profile: QAM is applied to 10 subjects (9 Males, 1 female) aged in the range of 24-44. The education level is relatively high as the subject pool is composed of 1 Post-Doc or higher degree and 2 PhD researchers and 7 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "R&D engineers, ML engineers, Q&A, embedded system engineers, software/hardware engineers, etc." having experience in the fields of "AI/ML, data analytics, radar, semantic web, IoT, digital twin etc.". Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-38. The results show that all constructs are correlated with each other.

Table 3-38. UC4 – Demonstrator 2 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.668	1									
MO	0.709	0.455	1								
CO	0.643	0.780	0.441	1							
ROI	0.528	0.738	0.383	0.790	1						
PE	0.635	0.446	0.680	0.515	0.761	1					
PT	0.538	0.827	0.329	0.856	0.716	0.436	1				
PR	0.621	0.519	0.755	0.650	0.350	0.314	0.484	1			
SI	0.424	0.415	0.565	0.749	0.595	0.481	0.630	0.771	1		
ATU	0.733	0.418	0.596	0.743	0.706	0.759	0.440	0.625	0.701	1	
BI	0.333	0.134	0.335	0.621	0.566	0.515	0.320	0.419	0.615	0.767	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-39, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-39. UC4 - Demonstrator 2 Reliability Analysis

Cronbach-alpha values	
PU	0.333
PEOU	0.812
MO	0.266
CO	0.552
ROI	0.394
PE	0.340
PT	0.640
PR	0.638
SI	0.328
ATU	1.366
BI	0.468

Regression Analysis: Regression analysis is applied to estimating the relationships among QAM constructs, as seen in Table 3-40. For this demonstrator, there exists a right proportional relation between all constructs influencing each other in the same direction.

Table 3-40. UC4 - Demonstrator 2 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO =0.43xPU + 3.29	0.503	0.022	2.844
H2	CO-PU	Right	CO = 0.255xPU + 4.52	0.414	0.045	2.376
H3	PEoU-PU	Right	PEoU = 0.28xPU + 4.36	0.446	0.035	2.537
H4	PU-ATU	Right	PU = 1.01xATU - 0.28	0.537	0.016	3.046
H5	PEoU-ATU	Right	PEoU = 0.243xATU + 4.37	0.174	0.23	1.3
H6	ATU-BI	Right	ATU = 0.84xBI + 1.52	0.589	0.01	3.383
H7	ROI-BI	Right	ROI = 0.45xBI + 3.74	0.32	0.088	1.942
H8	PE-BI	Right	PE = 0.547xBI + 3.05	0.265	0.128	1.7
H9	SI-BI	Right	SI = 0.33xBI + 4.51	0.378	0.058	2.206
H10	PT-BI	Right	PT = 0.28xBI + 4.65	0.103	0.367	0.957
H11	PR-BI	Right	PR = 0.27xBI + 4.78	0.103	0.228	1.304

3.4.7 Observed Limitations, Lessons Learnt and Best Practices

This use case showed that novel V&V technologies can be tailored and applied to improve the V&V processes of complex software-intensive systems. By using and integrating model-based system design, virtual validation, simulation coupling, and data analysis with machine learning the architectural design concept of the distributed assembly line could be validated.

The following findings have been identified:

- The connection between different tools has by nature limitations due to IPR reasons and technical obstacles. In UC4 these limitations had to be solved when connecting the tool FERAL and the commercial modelling software CIROS Studio.
- Interfaces need to be thoroughly pre-checked and aligned with each other to achieve seamless interoperability.
- Limitation in modelling due to limited libraries: Human models and specific sensors are not available and could also not be programmed or modelled themselves.
- Detailed, formalized system specifications and fault models facilitate the creation of automated validation scenarios using the given toolchain.

The mean and standard deviation values of user responses to UC4-Demonstration 1 are given in Table 3-41. The results show that users have a positive opinion especially related to the constructs: PU, MO, ROI, PE, PT, PR and BI. Responses to PEOU, CO, SI and ATU are relatively lower than the other constructs but there is no significant difference among the QAM factors. These results indicate that the users have sufficiently identified the importance of the proposed tools and they believe the demonstrated solution can enhance the failure detection and analyse the real-time data that can be collected from the robots, IPCs and actuators. However, if it is accepted that there is no problem with the statistical evaluation procedure and subjects consistently answered the questions, users seem quite sceptical about the compliance and ease of use of this tool with existing industrial settings. Such responses are normal when simulations are presented instead of real applications. When the solution is installed in real-life settings, the overall acceptance rate will most probably increase.

Table 3-41 Mean and standard deviation of experts' responses to UC4 - Demonstration 1

UC4-Demonstrator-1	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,60	4,78	5,25	4,75	4,98	5,11	5,08	5,05	4,47	4,74	5,00
Std Dev	0,92	1,62	1,44	1,71	1,58	1,57	1,67	1,82	2,22	1,92	1,79

As the user responses to UC4 – Demonstration 2 present (Table 3-42), the acceptance rate is relatively higher as compared to Demonstrator 1. This might be due to the increasing demand for using ML pipelines in industrial processes and manufacturing. Users seem more convinced and motivated to benefit from the results of Demonstration-2. However, some respondents are still sceptical about the simulation idea itself as they believe that simulators may not present the entire set of cases and there is still a need to improve the scenario creation and enrichment methodologies. Nevertheless, it is noteworthy that this is a general comment and valid for all simulators.

Table 3-42 Mean and standard deviation of experts' responses to UC4 - Demonstration 2

UC4-Demonstrator-2	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,89	5,44	6,06	5,37	5,58	5,90	5,86	5,53	5,42	5,69	6,28
Std Dev	0,49	1,17	0,81	1,24	0,92	0,70	0,86	1,15	1,39	0,68	0,74

3.5 Use Case 5 – Aircraft Engine Controller (UC5)

The system for UC5 is an aircraft engine model along with a corresponding controller set. In particular, the engine model is represented by a linear state space model of 18 internal states, 4 outputs (pressure ratio and spool speed of high-pressure compressor, as well as exit Mach number and spool speed limit of low-pressure compressor), and 3 inputs (fuel flow, nozzle area and inlet guide vane angle). A pair of controllers have been designed, one for thrust control and the other for low-pressure compressor spool speed control, along with the switching logic that activates the appropriate controller based on the engine state and pilot commands. To demonstrate resilience to sensor faults, a sensor voting mechanism has been adapted and integrated with the system model. While the current system model in UC5 is a simplified version (particularly concerning the engine part) of the actual system, it can readily serve as a testbed for the demonstration of relevant V&V methodologies, as well as a way to evaluate improvements on such methodologies.

3.5.1 V&V challenges

The verification and validation of the UC5 model pose various challenges:

- Scalability is a major issue in the verification of hybrid systems. Symbolic techniques typically scale up to a dozen variables, while the considered model has more than 20 variables.
- The heterogeneity of the properties to be verified (stability, robustness, reliability) poses challenges to having an integrated flow that addresses all of them. At the same time, properties are difficult to be expressed using temporal logic, thus it is important for the engine control engineers to co-evaluate results with formal verification experts
- Elements of the Simulink model may need to be discretised for formal modelling and verification.
- In the context of the Verifiable Formal Requirements workflow, state space explosion may occur when using model-checking tools for verification of formalised requirements. Abstraction steps may result in a formal model whose behaviour does not exactly match the associated Simulink model. The contracts generated from requirements may need to be further edited so that they are easier to relate to specific components in the Simulink model.
- In the context of the SiLVer workflow, manual translation from model and requirements to C++ may be time-consuming. Automated translation is missing currently (but parametrized templates are provided to alleviate the issue).

3.5.2 Contributors

Partners contributing to the UC: **UTRCI**, NUIM, FBK, RISE

3.5.3 Contributors' Roles & Evaluation Scenario

List of evaluation scenarios defined for this use case:

- VALU3S_WP1_Aerospace_1 - Robust and safe operation under sensor faults
- VALU3S_WP1_Aerospace_2 - Robust operation under system parameter perturbation
- VALU3S_WP1_Aerospace_3 - Robust operation under low probability hazardous events
- VALU3S_WP1_Aerospace_4 - Robust fault detection, isolation, and recovery

An overview of individual partners' contributions within the evaluation scenarios can be found in Table 3-43.

Table 3-43 Overview of contribution to evaluation scenarios by UC5 partners

Evaluation Scenario	UTRCI	NUIM	FBK	RISE
VALU3S_WP1_ Aerospace_1	X	X	X	X
VALU3S_WP1_ Aerospace_2	X	X	X	
VALU3S_WP1_ Aerospace_3	X	X		
VALU3S_WP1_ Aerospace_4	X	X		X

Individual UC partners are contributing to evaluation scenarios as follows:

- UTRCI (currently Collins Aerospace) as the UC provider, supports all UC partners and provides appropriate pieces of code and system models. In addition, through the SiLVer workflow and tool, UTRCI provides a methodology to test requirement satisfaction on the system model in the presence of faults.
- NUIM's approach to Verifying and Refactoring Formalised Requirements (VeRFoR) connects a semi-formal requirements engineering phase with formal verification of the system's design. The link between these phases also encourages traceability of the requirements through the formalised properties, against which the system's design is formally verified. Requirement violations can be detected by several verification activities including formal methods, simulation, testing and run-time monitoring. However, it is important that when violations are detected that the root cause of the conflict is identified and resolved. One way to reduce requirement violations from the outset is to follow a methodical requirements elicitation and specification process that involves formalising and slowly refining the requirements. This involves beginning with a high-level set of requirements that are gradually decomposed into more detailed, specific requirements. Typically, this kind of process would start with abstract natural language requirements and return to a larger set of formalised requirements. Here we can measure the following:
    - Number of natural language requirements: 12
    - Number of formalised requirements: 42

These numbers give an idea of the effort involved in removing ambiguities from natural language requirements and the formalised requirements can be used as direct input to other verification tools/techniques.

The NUIM tool improvement (MU-FRET) enables the semi-formal requirements to be refactored – a process of modifying the architecture of software, without altering its behaviour. Like with software refactoring, refactoring requirements can make them easier to maintain. Natural language requirements expressed in FRET are refactored with tool support in Mu-FRET and refactored requirements are verified as having the same behaviour as the original requirements while providing greater traceability between natural language requirements and verified properties of the system. The Semi-Formal requirements can then drive formal verification, both using Event-B models and using CoCoSim contracts in a Simulink diagram. Because they have formalised all the requirements, their work supports all four evaluation scenarios.

- FBK mainly contributes with a verification methodology based on a reformulation of the combined engine/controller model as a hybrid system, followed by an analysis phase which employs a mixture of numeric and symbolic tools, and finally provides symbolic guarantees on the stability of the system under specific assumptions on the input parameters. This formal verification task is integrated into a workflow that comprises other analyses including model-based safety analysis of the sensors' redundancy.
- RISE through the model-implemented fault/attack injection with pre-injection analysis workflow provides state-of-the-art fault- and attack injection capabilities in the form of pre-injection analyses aimed at reducing the error space before the V&V is conducted. The pre-injection analysis techniques inject-on-read, inject-on-write and error space pruning of signals have been implemented in an improved version of their MODIFI tool used for conducting fault- and attack injection on the aircraft engine Simulink model.

Tools and methods for this use case can be seen in Figure 3-21.

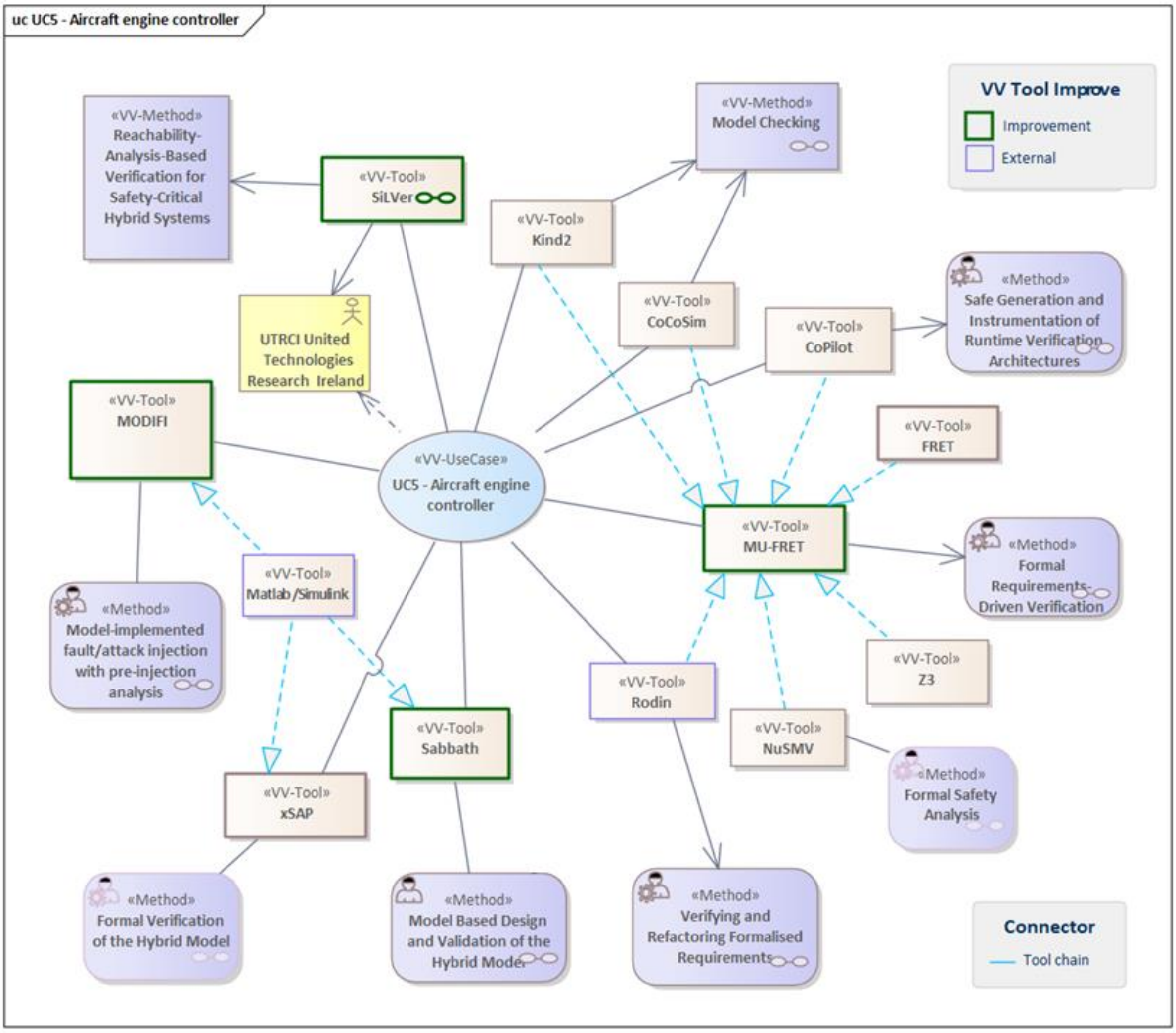

uc UC5 - Aircraft engine controller
VV Tool Improve
Improvement
External
«VV-Method» Reachability-Analysis-Based Verification for Safety-Critical Hybrid Systems
«VV-Tool» SiLVer
«VV-Tool» Kind2
«VV-Method» Model Checking
«VV-Tool» CoCoSim
«VV-Tool» CoPilot
«Method» Safe Generation and Instrumentation of Runtime Verification Architectures
UTRCI United Technologies Research Ireland
«VV-Tool» MODIFI
«VV-Tool» FRET
«VV-UseCase» UC5 - Aircraft engine controller
«VV-Tool» MU-FRET
«Method» Formal Requirements-Driven Verification
«VV-Tool» Matlab/Simulink
«Method» Model-implemented fault/attack injection with pre-injection analysis
«VV-Tool» Z3
«VV-Tool» Rodin
«VV-Tool» Sabbath
«VV-Tool» NuSMV
«Method» Formal Safety Analysis
«VV-Tool» xSAP
«Method» Formal Verification of the Hybrid Model
«Method» Model Based Design and Validation of the Hybrid Model
«Method» Verifying and Refactoring Formalised Requirements
Connector
Tool chain

Figure 3-21 Tools and Methods for UC5

List of evaluation criteria of the V&V process:

1. Eval_VV_1 – Time of test execution (UTRCI, RISE)
   a. UTRCI compares test execution time for similar coverage between Monte Carlo simulations and the developed tool for reachability analysis (SiLVer).
   b. RISE compares the execution time when injecting faults into all the relevant signals of the aircraft engine Simulink model with and without an improved workflow for model-implemented fault/attack injection with pre-injection analysis. The comparison is made for the pre-injection analysis techniques inject-on-read, inject-on-write and error space pruning of signals.
2. Eval_VV_3 – Number of test cases (UTRCI, RISE)
   a. UTRCI compares coverage of test cases requiring similar execution times between Monte Carlo simulations and the developed tool for reachability analysis (SiLVer).
   b. RISE compares the number of test cases performed when injecting faults into all the relevant signals of the aircraft engine Simulink model with and without an improved workflow for model-implemented fault/attack injection with pre-injection analysis. The

comparison is made for the pre-injection analysis techniques inject-on-read, inject-on-write and error space pruning of signals.

3. Eval_VV_5 – Joint Management of SCP Requirements (RISE)
    a. Some of the fault models used in model-implemented fault injection can be mapped to cybersecurity attack models used in model-implemented attack injection. For example, oscillation or noise faults injected into signals of Simulink models to test the safety requirements may be modelled similarly to jamming attacks used for testing cybersecurity requirements. This means that some attacks may affect the system in the same way as faults do, which implies that the lessons learned from previous fault injection campaigns could be reused when evaluating the cybersecurity of safety-critical systems. This is demonstrated by RISE with the MODIFI tool which uses both model-implemented fault- and attack injection on the aircraft engine Simulink model, e.g., where noise fault models corresponding to jamming attacks are injected into the sensor signals of the aircraft engine.
4. Eval_VV_8 – Effort needed for test (UTRCI)
    a. UTRCI reduces the effort needed for testing by reducing the effort for preparing the test cases (automatic generation of control property requirement monitors from parametrized templates).
5. Eval_VV_10 – Reduced cost and time for work on the certification process and functional safety (UTRCI)
    a. UTRCI shows that certification costs can be significantly reduced by demonstrating a reduction in testing effort (Eval_VV_8), test execution time (Eval_VV_1), as well as number of test cases required for adequate coverage (Eval_VV_3).

List of evaluation criteria for SCP:

1. Eval_SCP_1 – Error coverage (UTRCI, RISE)
    a. UTRCI injects faults into the UC5 Simulink model and demonstrates improvement in error coverage by a sensor voter integrated with the system model to implement redundancy. Without the sensor voter, observed error coverage is 0% (no errors are detected; output deviation may be observed; the system may become unstable) -- with the sensor voter, observed error coverage is 100% (all errors are detected and corrected – no deviation from nominal / expected behaviour is observed).
    b. RISE injects faults into the aircraft engine Simulink model to estimate the error coverage with and without the sensor voter integrated with the system model to demonstrate the improvement in error coverage obtained with the voter.
2. Eval_SCP_2 – Number of safety/security requirement violations (NUIM)
    a. NUIM has accurately elicited and formalised the original set of natural language requirements. There were 14 requirements originally and, after a thorough elicitation process by NUIM, this number increased to 42 requirements. This demonstrates that significant ambiguities were present in the natural-language requirements that could be identified and captured by formalising the requirements. NUIM will provide support for refactoring these natural language requirements expressed in FRET, via the

Mu-FRET tool. Here, refactored requirements are verified as having the same behaviour as the original requirements, while providing greater traceability between natural language requirements and verified properties of the system.

3. Eval_SCP_3 – Number of malicious attacks and faults detected (UTRCI)
    a. Similarly to Eval_SCP_1, UTRCI measures this before and after integration with the sensor voter. Without sensor voter, no injected faults are detected. With the sensor voter, all injected faults are detected.
4. Eval_SCP_10 – Software fault tolerance robustness (UTRCI, FBK)
    a. UTRCI evaluates the sensor voter component implementation in isolation and checks conformance with the corresponding specification (expected behaviour).
    b. FBK has evaluated the fault tolerance robustness, by performing Fault Tree Analysis on a redundant schema of the sensors architecture, using the xSAP tool.
5. Eval_SCP_11 – Simulation level system robustness (UTRCI, FBK)
    a. UTRCI evaluates overall system robustness before and after integration with the sensor voter (see Eval_SCP_1 and Eval_SCP_3).
    b. FBK obtains certified proof of the stability of the hybrid system by the use of symbolic techniques. The model considered is a hybrid system with two modes, each one consisting of an affine dynamical system. The evaluation focuses on two different aspects: the synthesis of a robust region (with fixed reference values), and robustness to reference value changes. We approach these two targets by use of the tool Sabbath, which computes and certifies quadratic Lyapunov functions.

3.5.4 Demonstration

Demonstration for UC5 was planned by individual partners. Therefore, multiple demonstrable items are covering the defined challenges and scenarios and partially cover the VALU3S dimensions (see Table 3-44).

Table 3-44: Overview of demonstration prepared by UC5 partners.

Item #	Demonstration name	Description/Purpose	Format	Responsible
1	SiLVer demonstration	Demonstrate improved V&V capabilities of the SiLVer workflow and tool on the UC5 model.	Lead Demo (video or live) showing the application of the SiLVer workflow on UC5	UTRCI
2	Mu-FRET demonstration	Demonstrate semi-formal requirements set, refactoring behaviour (in the improved tool, MU-FRET), and formal verification from the semi-formal requirements using both Event-B and CoCoSim contracts in the Use Case's Simulink model.	Complementary Demo showcasing the Mu-FRET tool with a presentation that demonstrates verifying and refactoring formalised requirements (VeRFoR) features on the UC5 requirements and Q&A session.	NUIM

Item #	Demonstration name	Description/Purpose	Format	Responsible
3	Model-based Design and Validation of the Hybrid Model Demonstration	Demonstrate improved applicability of the formal verification approaches, providing formal guarantees on the stability of the switched system with compositional reasoning.	Complementary Demo (video or live) showing the application of Sabbath	FBK
4	Model Implemented Fault and Attack Injection demonstration	Demonstrate application of the improved MIFI_MIAI workflow on the UC5 Simulink model.	Complementary Demo (video or live) showing the application of the improved workflow	RISE

3.5.5 Quantitative Results

UTRCI contributes by improving performance and usability of the in-house developed reachability analysis framework, SiLVer (SimuLation-based Verification). The framework aims to be a near-drop-in replacement for Monte Carlo simulations in the context of system testing / verification, and the core improvements are in terms of analysis coverage and test execution time. Compared to Monte Carlo, analysis coverage is better, since a single symbolic simulation run can obtain results for entire intervals of system inputs and internal states – in comparison, a Monte Carlo approach would require several hundred (or potentially thousands) of runs to obtain similar coverage. For the same reason, the total test execution time to obtain comparable coverage is lower with SiLVer. Note that this difference becomes more pronounced as the dimension of the analysed system increases. This is because increased system dimension implies that a Monte Carlo approach will reach a combinatorial explosion wall quite early. This issue also occurs in our interval-based reachability analysis framework (splitting the testing domain in several, non-overlapping intervals is typically desirable since keeping interval width small helps to keep the analysis results precise), however, due to the interval nature of the analysis, this happens much later (higher system dimension) in comparison with the Monte Carlo approach. We argue that the above mentioned improvements (increased coverage and reduced testing time – typically by 25% - 33% with these savings potentially increasing with system dimension) coupled with reduced testing effort by providing requirement monitor templates for well-known control properties (e.g. overshoot, settling time, steady state error), help considerably reduce certification costs, and this summarizes our improvements w.r.t. the V&V evaluation criteria. Concerning the SCP criteria, the focus here is not on improving the state of the art (i.e. our implementation of the sensor voter is fairly standard), but rather on showcasing that our framework is a practical option for carrying out analysis in this setting as well (i.e. with integrated sensor voter). This is in contrast with reachability analysis approaches from academia, where the main bottleneck is in computing intersections of reachable states / flowpipes with (mode transition) guards. We address this challenge by relaxing the rigorousness of the approach and temporarily falling back to gridded simulations when faced with such computations (note that once the intersection is computed we revert to interval-based analysis; also, gridding is only performed for quantities involved in the guard expressions – the rest are kept as intervals). Similar to the V&V evaluation results described earlier, the resulting approach is comparable with Monte Carlo simulations in execution speed and, at the same time, offers better coverage, due to its interval-based nature. As an example, during verification of the sensor voter component in isolation, it was possible in some cases (3 inputs are in agreement and the average value is simply returned) to reduce the number of test cases per input by more than 80%, as well as the total test execution time by more than 35%, as shown in Figure 3-22 and Figure 3-23, and still obtain better coverage (due to the interval nature of our analysis).

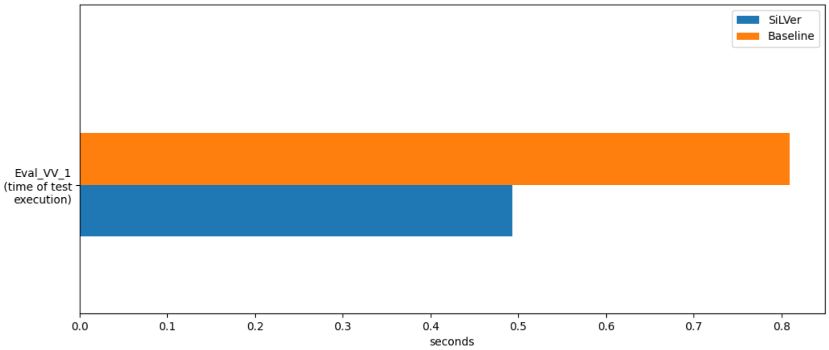


Figure 3-22 EVAL-VV1 (time of test execution) improvement as compared to the baseline.

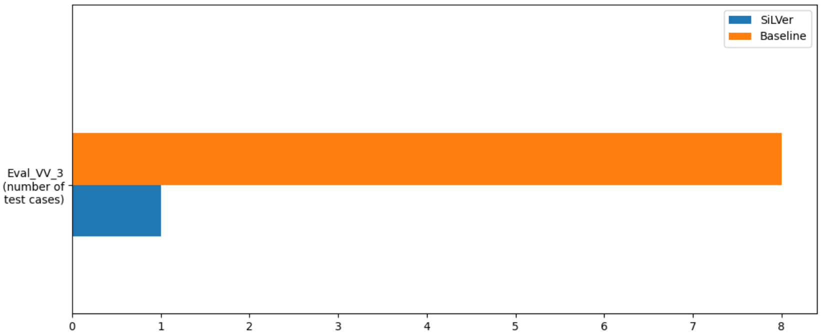
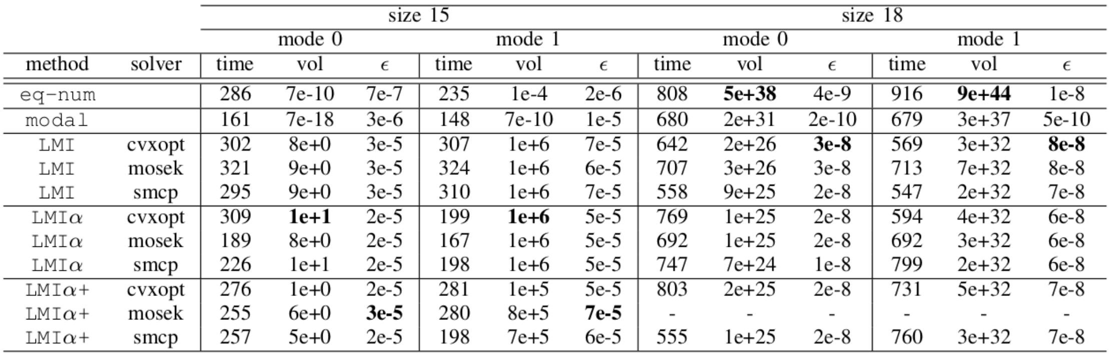


Figure 3-23 EVAL-VV3 (number of test cases) improvement as compared to the baseline.

FBK contributes by certifying the stability of hybrid systems by means of symbolic techniques. The model considered is a hybrid system with two modes, each one consisting of an affine dynamical system. The evaluation focuses on two aspects: the synthesis of a robust region (with fixed reference values), and robustness to reference value changes. We approach these two targets using the tool Sabbath, which computes and certifies quadratic Lyapunov functions.
To assess the scalability of the algorithm, we consider different sizes of the problem, obtained by Balanced Truncation Model Reduction on the full system; we also consider Lyapunov functions obtained using different algorithms, and we validate them using different SMT solvers or symbolic methods.

Out of 192 problems: 186 Lyapunov functions were validated, 2 proved incorrect, and 4 syntheses failed due to timeout. The volume of the synthesized regions and the robustness to reference values changes depends on the method used to synthesize the Lyapunov function and on the specific dynamical system considered. In Table 3-45, we report only the results regarding the two largest systems (size 15 and size 18). Column "time" reports the time in seconds needed to compute the robust region; column "vol" reports the volume of a such region; column "ϵ" reports the radius of the ball that is proved to be a robust region for reference values changes. For each problem, we highlight the maximum value for the volume of the robust region and the robustness ϵ.

Table 3-45 Comparison of methods and solvers concerning different sizes and modes.



		size 15						size 18					
		mode 0			mode 1			mode 0			mode 1		
method	solver	time	vol	ϵ	time	vol	ϵ	time	vol	ϵ	time	vol	ϵ
eq-num		286	7e-10	7e-7	235	1e-4	2e-6	808	**5e+38**	4e-9	916	**9e+44**	1e-8
modal		161	7e-18	3e-6	148	7e-10	1e-5	680	2e+31	2e-10	679	3e+37	5e-10
LMI	cvxopt	302	8e+0	3e-5	307	1e+6	7e-5	642	2e+26	**3e-8**	569	3e+32	**8e-8**
LMI	mosek	321	9e+0	3e-5	324	1e+6	6e-5	707	3e+26	3e-8	713	7e+32	8e-8
LMI	smcp	295	9e+0	3e-5	310	1e+6	7e-5	558	9e+25	2e-8	547	2e+32	7e-8
LMIα	cvxopt	309	**1e+1**	2e-5	199	**1e+6**	5e-5	769	1e+25	2e-8	594	4e+32	6e-8
LMIα	mosek	189	8e+0	2e-5	167	1e+6	5e-5	692	1e+25	2e-8	692	3e+32	6e-8
LMIα	smcp	226	1e+1	2e-5	198	1e+6	5e-5	747	7e+24	1e-8	799	2e+32	6e-8
LMIα+	cvxopt	276	1e+0	2e-5	281	1e+5	5e-5	803	2e+25	2e-8	731	5e+32	7e-8
LMIα+	mosek	255	6e+0	**3e-5**	280	8e+5	**7e-5**	-	-	-	-	-	-
LMIα+	smcp	257	5e+0	2e-5	198	7e+5	6e-5	555	1e+25	2e-8	760	3e+32	7e-8

In Figure 3-24, we present the results of the full-size case which are related to Eval_VV_3 – Number of test cases, Eval_VV_1 – Time of test execution, and Eval_SCP_11 – Simulation level system robustness. The quantitative improvement can vary, based on how many test cases belong to the synthesized region of certified stability. Therefore, the total saving depends on the density of the test cases. Implementing multiple methods instead of just one can boost the results by up to 30%, as shown in the charts above. This improvement is measured concerning the linearized volume (21st root of the volume) since we are comparing volumes of 21-dimensional regions.

NUIM has accurately elicited and formalised the original set of natural language requirements for use case 5. There were 14 requirements originally and, after a thorough elicitation process by NUIM, this number increased to 42 requirements. This demonstrates that significant ambiguities were present in the natural-language requirements that could be identified and captured by formalising the requirements, thus addressing SCP_2 by ensuring that requirements are consistent and that it is easier to determine safety/security requirements violations. NUIM provides support for refactoring these natural language requirements expressed in FRET, via the Mu-FRET tool, by adding refactoring functionality. Refactoring, when applied to software, is the process of rearranging the software's internal structure without changing its external behaviour. As presented in Table 3-46, refactoring is helpful for the maintainability of software and has similar benefits for requirements. MU-FRET enables a user to extract parts of a requirement (fragments) to a new requirement, allowing the extracted part to be reused. The number of times each fragment's definition occurs in the requirements for this use case is described in Figure 3-52.

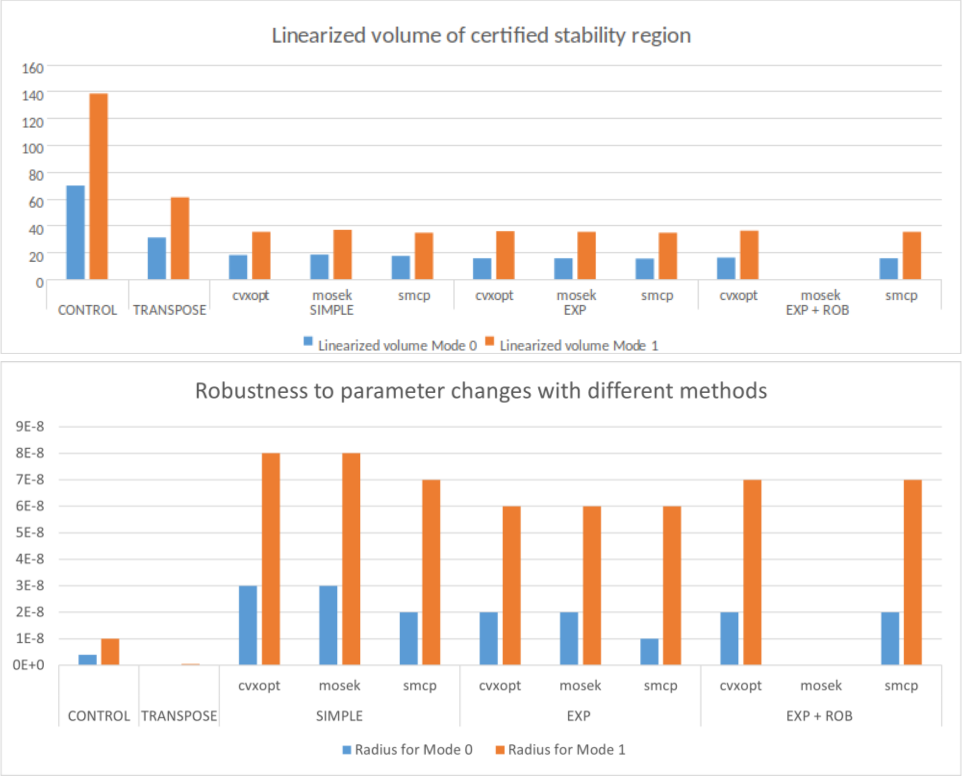
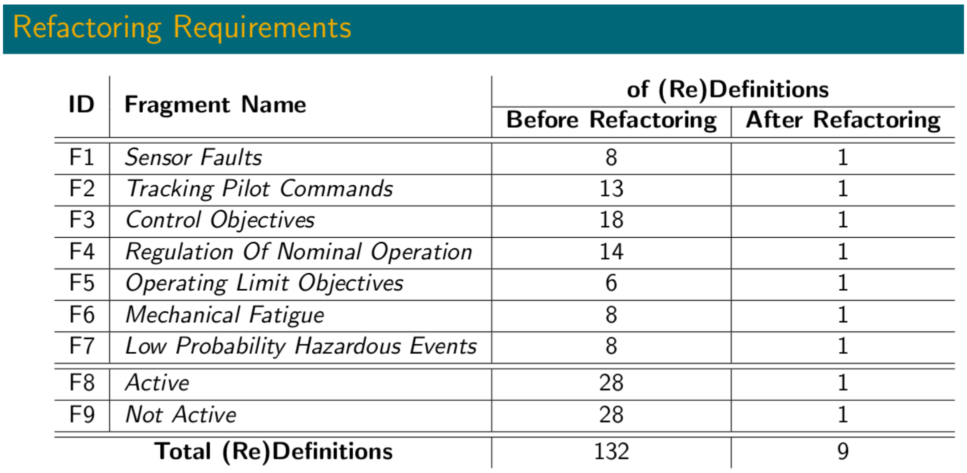


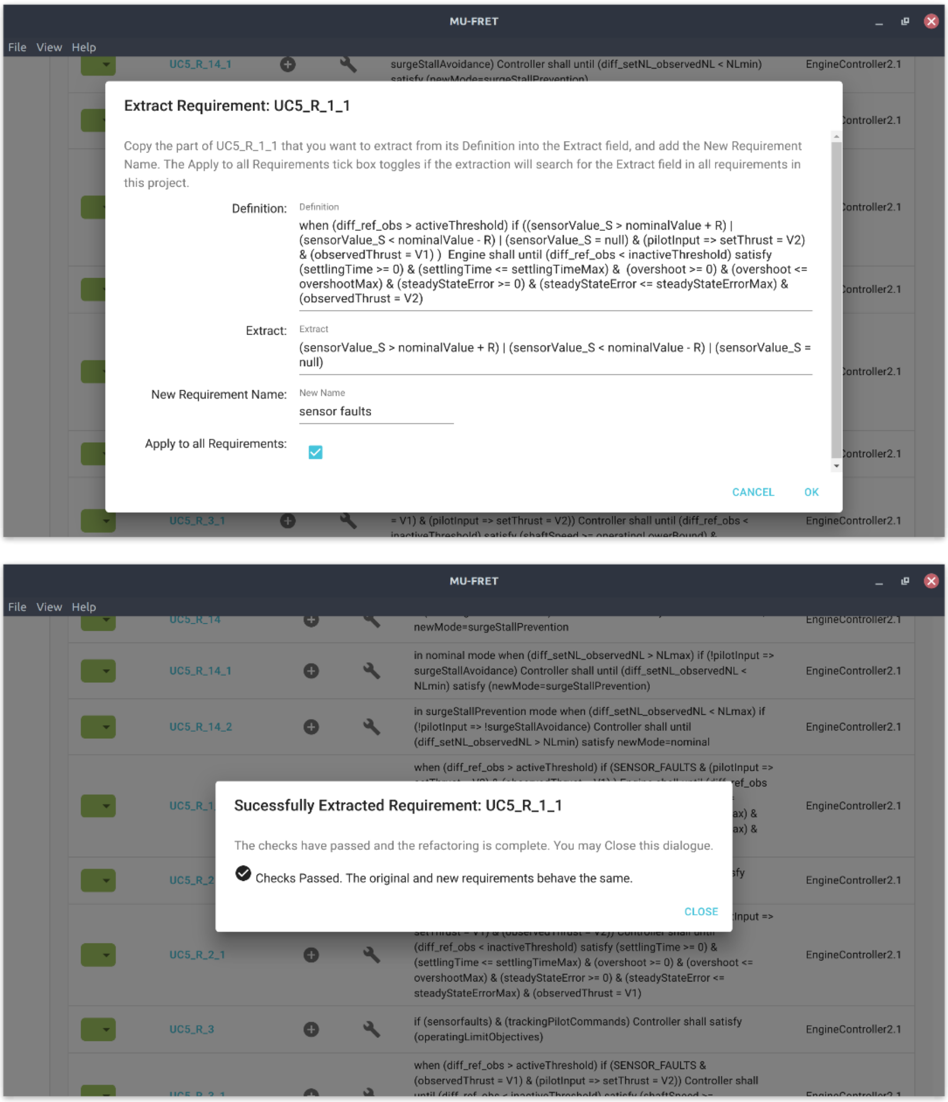
Figure 3-24 (top) Linearized volume of certified stability region; (bottom) Robustness to parameter changes with different methods.

Table 3-46 Mu-FRET Refactoring: The number of times each fragment's definition occurs in a child requirement.



Refactoring Requirements

ID	Fragment Name	of (Re)Definitions	
		Before Refactoring	After Refactoring
F1	*Sensor Faults*	8	1
F2	*Tracking Pilot Commands*	13	1
F3	*Control Objectives*	18	1
F4	*Regulation Of Nominal Operation*	14	1
F5	*Operating Limit Objectives*	6	1
F6	*Mechanical Fatigue*	8	1
F7	*Low Probability Hazardous Events*	8	1
F8	*Active*	28	1
F9	*Not Active*	28	1
	Total (Re)Definitions	132	9



Extract Requirement: UC5_R_1_1

Copy the part of UC5_R_1_1 that you want to extract from its Definition into the Extract field, and add the New Requirement Name. The Apply to all Requirements tick box toggles if the extraction will search for the Extract field in all requirements in this project.

Definition: when (diff_ref_obs > activeThreshold) if ((sensorValue_S > nominalValue + R) | (sensorValue_S < nominalValue - R) | (sensorValue_S = null) & (pilotInput => setThrust = V2) & (observedThrust = V1)) Engine shall until (diff_ref_obs < inactiveThreshold) satisfy (settlingTime >= 0) & (settlingTime <= settlingTimeMax) & (overshoot >= 0) & (overshoot <= overshootMax) & (steadyStateError >= 0) & (steadyStateError <= steadyStateErrorMax) & (observedThrust = V2)

Extract: (sensorValue_S > nominalValue + R) | (sensorValue_S < nominalValue - R) | (sensorValue_S = null)

New Requirement Name: sensor faults

Apply to all Requirements: ☑

CANCEL OK

ID	Requirement	Component
UC5_R_14_1	in nominal mode when (diff_setNL_observedNL > NLmax) if (!pilotInput => surgeStallAvoidance) Controller shall until (diff_setNL_observedNL < NLmin) satisfy (newMode=surgeStallPrevention)	EngineController2.1
UC5_R_14_2	in surgeStallPrevention mode when (diff_setNL_observedNL < NLmax) if (!pilotInput => !surgeStallAvoidance) Controller shall until (diff_setNL_observedNL > NLmin) satisfy newMode=nominal	EngineController2.1
UC5_R_2_1	(diff_ref_obs < inactiveThreshold) satisfy (settlingTime >= 0) & (settlingTime <= settlingTimeMax) & (overshoot >= 0) & (overshoot <= overshootMax) & (steadyStateError >= 0) & (steadyStateError <= steadyStateErrorMax) & (observedThrust = V1)	EngineController2.1
UC5_R_3	if (sensorfaults) & (trackingPilotCommands) Controller shall satisfy (operatingLimitObjectives)	EngineController2.1

Sucessfully Extracted Requirement: UC5_R_1_1

The checks have passed and the refactoring is complete. You may Close this dialogue.

✓ Checks Passed. The original and new requirements behave the same.

CLOSE

Figure 3-25 Snapshots from the MU-FRET tool

As presented in Figure 3-25, MU-FRET also formally verifies that the refactored requirement (including the extracted parts) has the same behaviour as the original requirement. This gives confidence that the tool has not inadvertently introduced new (possibly incorrect) behaviour.

RISE contributes with pre-injection analysis techniques for model-implemented fault- and attack injection applied on the Simulink model of the UC5 aero engine controller using the MODIFI tool.

<u>Results related to Eval_VV_3 – Number of test cases:</u>

When no pre-injection analysis is performed, the fault- and attack-injection experiments can target 958 signals in the model. Using inject-on-read and inject-on-write pre-injection analyses, the number of signals targeted is 408 and 550, respectively, corresponding to a 43% and 57% reduction of the entire error space.

Detailed results obtained for the inject-on-read and inject-on write pre-injection analyses suggest that to have more complete data about the sensitivity of the target system to faults and attacks on all locations, the results obtained for inject-on-read should be accompanied by those obtained for the inject-on-write technique. However, as this would result in an error space identical to the case when no pre-injection analysis is done, another technique called error space pruning of signals has been developed and evaluated. The number of signals targeted for fault injection is reduced to 702 signals for the error space pruning of signals technique corresponding to a 27% reduction of the total error space.

Results related to Eval_VV_1 – Time of test execution:

Since the time for performing the pre-injection analysis is negligible compared to the time for conducting the experiments, the reduction in time of execution for performing the experiments corresponds to the error space reduction achieved, i.e., 43%, 57% and 27% reduction for inject-on-read, inject-on-write and error space pruning of signals, respectively.

Results related to Eval_VV_5 – Joint Management of SCP Requirements:

The noise faults injected into the sensor signals of the aero engine controller model with MODIFI are equivalent to jamming attacks allowing both safety and cybersecurity requirements to be verified jointly when either of these fault/attack models is used.

Results related to Eval_SCP_1 – Error coverage:

Permanent offset faults were injected into the sensor signals of the aircraft engine Simulink model to estimate the error coverage with and without the sensor voter integrated with the system model. The results from the fault injection experiments demonstrate an improvement of the observed error coverage from 9% without the voter to 61% using the voter. It should be noted that noise disturbances were included in the state and output of the target system model for these experiments which were not compensated for directly by the voter. Thus, the error coverage is expected to be even higher without the noise applied.

3.5.6 Qualitative Results

Demonstrator 1: Mu-FRET

Participants Profile: QAM is applied to 12 subjects (9 Males, 1 female, 2 unknown) aged in the range of 24-64. The education level is relatively high as the subject pool is composed of 3 Post-Doc or higher-degree and 3 PhD researchers and 6 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as “professors, lecturers, researchers, R&D engineers, software/hardware engineers, etc.” having experience in the fields of “formal verification, model-driven engineering, healthcare, automotive, machinery, semantic web, IoT, digital twin etc.”.

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-47. The results show that all constructs are highly correlated with each other.

Table 3-47. UC5 – Demonstrator 1 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.608	1									
MO	0.945	0.725	1								
CO	0.835	0.776	0.874	1							
ROI	0.891	0.777	0.934	0.972	1						
PE	0.876	0.769	0.946	0.911	0.950	1					
PT	0.869	0.716	0.909	0.963	0.961	0.901	1				
PR	0.887	0.781	0.918	0.971	0.988	0.919	0.966	1			
SI	0.818	0.755	0.905	0.874	0.916	0.940	0.888	0.884	1		
ATU	0.775	0.885	0.906	0.870	0.914	0.918	0.868	0.910	0.928	1	
BI	0.893	0.778	0.950	0.875	0.938	0.963	0.877	0.896	0.965	0.919	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-48, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-48. UC5 - Demonstrator 1 Reliability Analysis

Cronbach-alpha values	
PU	0.847
PEOU	0.847
MO	0.849
CO	0.703
ROI	0.745
PE	0.784
PT	0.689
PR	0.753
SI	0.124
ATU	0.260
BI	0.650

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-49. For this demonstrator, there exists a right proportional relation between all constructs influencing each other in the same direction.

Table 3-49. UC5 - Demonstrator 1 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.63xPU + 2.16	0.893	~0	9.123
H2	CO-PU	Right	CO = 0.579xPU + 2.86	0.697	0.001	4.794
H3	PEoU-PU	Right	PEoU = 0.39xPU + 3.65	0.37	0.036	2.423

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H4	PU-ATU	Right	PU = 1.76xATU - 5.59	0.6	0.003	3.874
H5	PEoU-ATU	Right	PEoU = 1.295xATU - 1.76	0.782	~0	5.998
H6	ATU-BI	Right	ATU = 0.95xBI + 0.81	0.845	~0	7.371
H7	ROI-BI	Right	ROI = 1.4xBI - 1.77	0.881	~0	8.592
H8	PE-BI	Right	PE = 1.316xBI - 1.94	0.927	~0	11.228
H9	SI-BI	Right	SI = 0.96xBI + 0.26	0.931	~0	11.64
H10	PT-BI	Right	PT = 1.17xBI - 0.86	0.769	~0	5.777
H11	PR-BI	Right	PR = 1.123xBI - 1.19	0.769	~0	6.376

Demonstrator 2: SMT-based verification of stability

Participants Profile: QAM is applied to 10 subjects (6 Males, 1 female, 3 unknown) aged in the range of 24-54. The education level is relatively high as the subject pool is composed of 4 Post-Doc or higher-degree and 2 PhD researchers and 4 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "professors, lecturers, principal engineers, R&D engineers, etc." having experience in the fields of "formal verification, aerospace, cyber-physical systems, semantic web, IoT, digital twin etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-50. The results show that all constructs are correlated with each other.

Table 3-50. UC5 – Demonstrator 2 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.867	1									
MO	0.794	0.892	1								
CO	0.224	0.577	0.495	1							
ROI	0.270	0.626	0.607	0.786	1						
PE	0.311	0.660	0.641	0.748	0.980	1					
PT	0.204	0.582	0.608	0.921	0.908	0.894	1				
PR	0.652	0.756	0.654	0.582	0.668	0.597	0.541	1			
SI	0.355	0.739	0.722	0.827	0.915	0.946	0.939	0.545	1		
ATU	0.149	0.446	0.417	0.656	0.865	0.833	0.704	0.582	0.705	1	
BI	0.745	0.761	0.693	0.542	0.636	0.628	0.523	0.899	0.543	0.588	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-51, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-51. UC5 - Demonstrator 2 Reliability Analysis

Cronbach-alpha values	
PU	0.712
PEOU	0.222
MO	0.170
CO	0.573
ROI	0.555
PE	0.644
PT	0.485
PR	0.343
SI	0.132
ATU	0.409
BI	0.463

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-52. For this demonstrator, there exists a right proportional relation between all constructs influencing each other in the same direction.

Table 3-52. UC5 - Demonstrator 2 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.43xPU + 2.89	0.631	0.006	3.699
H2	CO-PU	Right	CO = 0.156xPU + 4.43	0.05	0.534	0.65
H3	PEoU-PU	Right	PEoU = 0.59xPU + 2.27	0.752	0.001	4.922
H4	PU-ATU	Right	PU = 0.29xATU + 2.8	0.022	0.68	0.427
H5	PEoU-ATU	Right	PEoU = 0.575xATU + 1.48	0.199	0.197	1.409
H6	ATU-BI	Right	ATU = 0.43xBI + 3.54	0.346	0.074	2.057
H7	ROI-BI	Right	ROI = 0.58xBI + 2.6	0.405	0.048	2.332
H8	PE-BI	Right	PE = 0.539xBI + 2.77	0.395	0.052	2.285
H9	SI-BI	Right	SI = 0.41xBI + 3.58	0.295	0.105	1.829
H10	PT-BI	Right	PT = 0.54xBI + 2.96	0.273	0.121	1.734
H11	PR-BI	Right	PR = 0.63xBI + 2.24	0.273	~0	5.792

3.5.7 Observed Limitations, Lessons Learnt and Best Practices

The following findings have been identified as a result of applied demonstrations:

- Close collaboration, between use-case providers and technology providers, clarified ambiguous text in the requirements and test cases. This iterative process produced a set of detailed, formalised requirements that we are confident correspond to the intent of the natural-language requirements. These requirements are now ready for use in formal verification activities.
- Providing traceability from requirements through to their formalisation is essential to ensure confidence that the system that is verified corresponds to the originally specified system.

- During this work, we identified improvements that could be made to FRET. Our formalised requirements contain quite a lot of repetition, so if changes were needed, we often had to make the change manually in several places. This was very time-consuming and motivated us to add automatic requirement refactoring.
- The usage of advanced formal verification approaches requires considerable expertise from a user perspective. It will be crucial in the next years to increase the usability of formal methods tools by increasing the explainability of the results to non-expert formal methods users, which will allow an immediate acceleration of comprehending verification results; such a fact will decrease engineering time and design cycles while it will disseminate the usage of verification techniques horizontally to all engineering steps – from requirements till prototype integration and deployment.
- Work is required to improve the interoperability of formal verification tools. For example, adding a translator to the input language of a theorem prover to avoid the state-explosion faced by model checkers (like Kind2, which is used to verify CoCoSpec contracts); or outputting the requirement to a parse tree; or integrating Simulink models with tools that can provide support for formal modelling and model checking properties.

The qualitative assessment results, as shown in Table 3-53 and Table 3-54, indicate that the level of acceptance of the proposed solution stack is at a moderate level. Although PU and MO results show that the experts are still positive, the demonstrated technologies can be more convincing. Since the factors like CO and PE are still below 5.00 (out of 7.00), experts may have question marks in their minds about the scalability and interoperability of the solutions (relatively low ATU especially for Demonstration-1). These results are quite normal because the aerospace industry has very strict rules and V&V processes are very complex in general. Nevertheless, the user responses are still promising, especially for Demonstration-2. The proposed solutions may have a better chance if the total time needed for the V&V processes is reduced and the scope of the verifiable system components is extended.

Table 3-53 Mean and standard deviation of experts' responses to UC5 - Demonstration 1

UC05-Demonstrator-1	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,33	4,31	5,00	4,28	4,43	4,84	4,52	5,01	4,36	3,82	4,44
Std Dev	1,02	1,59	1,52	1,47	1,61	1,76	1,80	1,92	2,42	2,33	2,41

Table 3-54 Mean and standard deviation of experts' responses to UC5 - Demonstration 2

UC5-Demonstrator-2	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
mean	5,10	4,84	5,09	4,27	4,79	4,83	4,44	4,97	4,37	4,26	5,38
Std Dev	0,92	1,36	1,68	1,31	1,40	1,49	1,23	1,83	1,69	1,75	1,28

3.6 Use Case 6 - Agricultural Robot (UC6)

The UC6 development system is an agricultural multi-utility robot that integrates parallel autonomous guidance feature (See Figure 3-26).

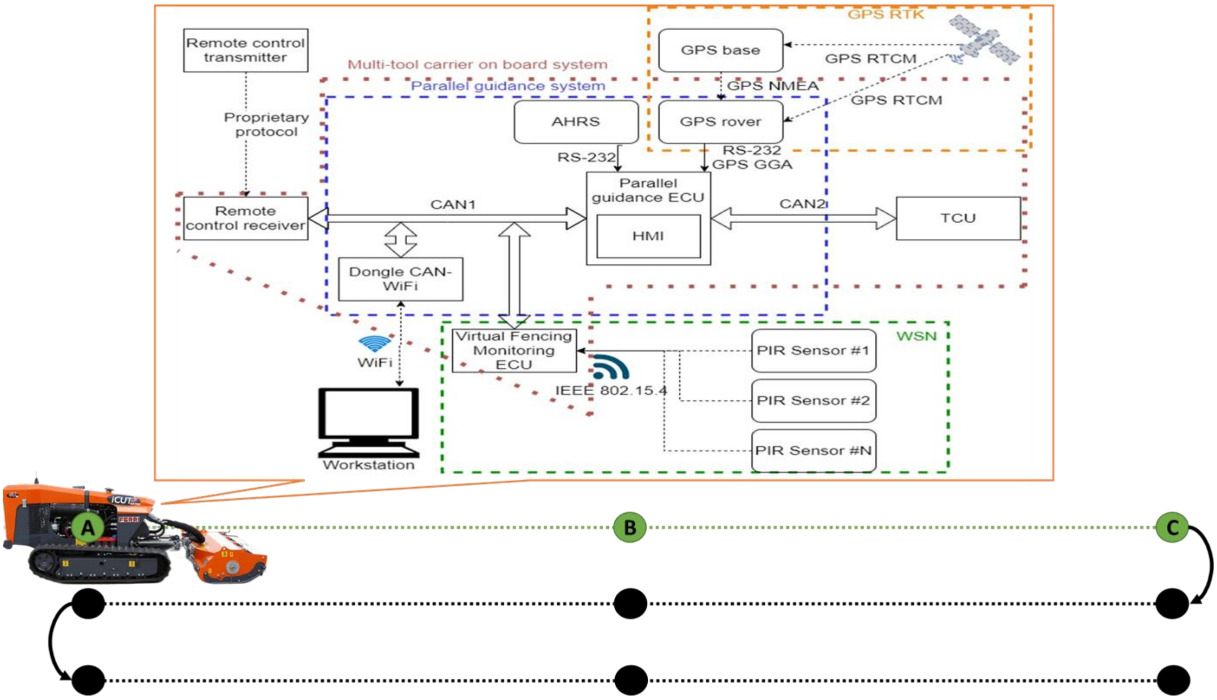


Figure 3-26 Parallel guidance system

The agricultural robot (agri-robot) is equipped with a parallel guidance system, which receives the location from GPS unit and the direction from the AHRS unit, The operator selects a couple of points A and B that the robot uses to calculate a straight line to keep the robot on the way to the end of the field. At the end of the field (point C), the robot automatically turns around and autonomously continues the parallel line next to the previous one with a defined offset.

The system also integrates a proximity sensors network to enhance the safety of the robot that stops its parallel guidance when the sensor reveals the presence of an operator in its vicinity. The system also implements a user interface with a PC Tool to command the robot and enable the usage of runtime error detector methods and tools running on PC. The PC user interfaces is enabled by a dongle WiFi to CAN.

3.6.1 V&V challenges

The V&V challenges in the design and development of the parallel autonomous guidance system, to be integrated into a multi-utility robot normally controlled by remote control; are related to defining and addressing safety and cybersecurity critical aspects in a new application field, in compliance with standard related to utilization of robots and automated systems in the agriculture domain.

This use case "Agricultural Robot" (UC6) looks at safety, cybersecurity, and privacy challenges from the perspective of verification and validation in the domain of the system integrator. The UC6 V&V

challenges are related to fill the identified gaps to existing V&V methods developing or improving the existing tool. The targeted challenges are:

- Safety/security requirement violations (EVAL_SCP_2).
- Malicious Attacks And Faults Detected (EVAL_SCP_3).
- Estimation of potential impact of incidents and attacks (EVAL_SCP_3).
- Cybersecurity and cyber-physical resilience problems (EVAL_SCP_6).
- Accidents in actual operation of the system (Eval_SCP_7).
- Robustness of the attacked system against false packet (such as command) injection through the radio wireless interface (Eval_SCP_10).
- Reference models lacking attack/incident typologies to be examined (Eval_SCP_12).
- Complexity, coverage and total time or effort needed for the test execution (Eval_VV_1, Eval_VV_2, Eval_VV_3).
- Joint Management of SCP Requirements is cumbersome and complex (Eval_VV_5).
- Cost and time for work on certification process and functional safety (Eval_VV_10).
- Effort needed for test and effort required to the user for prepare and running the tool (Eval_VV_8, Eval_VV_12).

3.6.2 Contributors

Partners contributing to the UC: ESTE, STAM, UNIVAQ, UNIGE, RULEX, INTECS.

3.6.3 Contributors' Roles & Evaluation Scenario

List of evaluation scenarios defined for this use case:

1. VALU3S_WP1_Agriculture_1 -Vehicle switching from parallel guidance to manual mode
2. VALU3S_WP1_Agriculture_2 -Vehicle switching from manual mode to parallel guidance
3. VALU3S_WP1_Agriculture_3 -Transmission line disturbances
4. VALU3S_WP1_ Agriculture_4-Disturbances of IMU communication

An overview of individual partners' contributions within the evaluation scenarios can be followed in Table 3-55.

Table 3-55 Overview of contribution to evaluation scenarios by UC6 partners

Evaluation Scenario	ESTE	STAM	UNIVAQ	UNIGE	RULEX	INTECS
VALU3S_WP1_ Agriculture_1	X	X				
VALU3S_WP1_ Agriculture_2	X	X				
VALU3S_WP1_ Agriculture_3	X	X	X	X	X	X
VALU3S_WP1_ Agriculture_4	X	X			X	X

The following partners are contributing to the demonstrator implementation (see also Figure 3-27)**:**

- ESTE, as the use case provider, is designing and implementing the parallel guidance algorithm, collecting data to train the partners' machine learning methods and providing data for model definition.
- STAM has developed the tool RAMSES that improves the risk assessment during the design phase of the agriculture robot considered within the use case. RAMSES supports the end-user (in this case, ESTE has a designer of the final system) in identifying risk scenarios according to ISO 12100 [45] standard and evaluating them in terms of likelihood and severity using the so-called "risk graph" foreseen in the standard. Finally, the tool allows users to find and add needed safety measures to decrease risk in different scenarios.
- INTECS has improved the CHESS tool [35], developing a plugin (CHESS-FLA) that can automatically generate the Failure Mode and Effect Analysis (FMEA) table and Fault Trees (FTs), and also automatically compute the probability of occurrence of the top events. INTECS is also working on the model of the agricultural robot, implementing a Failure Logical Analysis and, thus, applying the improved tool.
- UNIVAQ developed an improved method of intrusion detection for wireless sensor networks. UNIVAQ also created a new data-driven fault detector to detect deviations from normal system operations. UNIVAQ worked on the test of such methods on the agricultural robot operational data.
- UNIGE is working on radio-link security of agricultural robots, evaluating the security of the proprietary protocol, testing different attacks on the communication channel, and evaluating online intrusion detection strategies.
- RULEX is using machine learning methods based on rules to improve model-checking effectiveness in the implementation of safety requirements in agricultural robots. Rulex will make use of data derived from the real-world and simulated scenarios to build the models and to test their ability to identify critical regions in the input space.

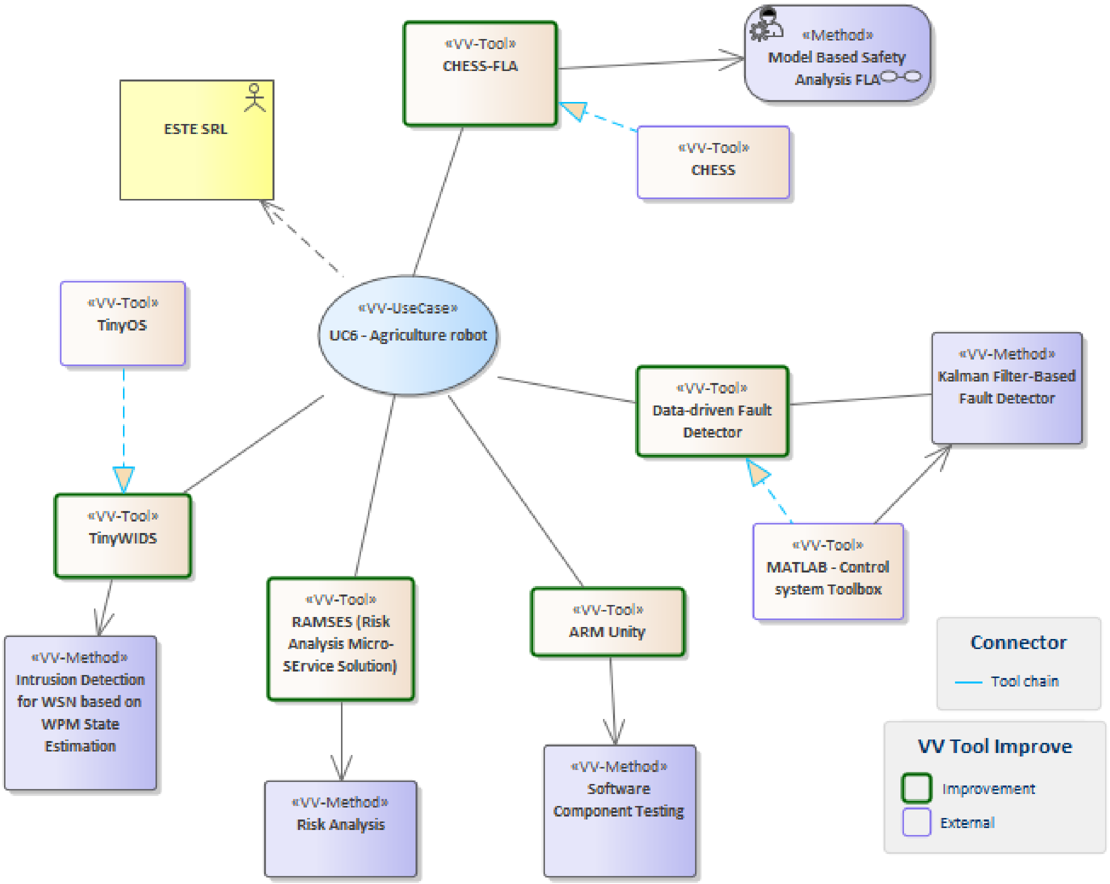


Figure 3-27 Tools and Methods for UC6

3.6.4 Demonstration

The demonstrator plan includes the following steps:

- Improved design based on VALU3S V&V tools are demonstrated as tools demonstration with a focus on VALU3S evaluation criteria.
- V&V Workflow driven by tools developed or improved:
  - Risk analysis tool
  - Model-based safety analysis tool
  - Model-based design verification
- The agricultural robot parallel guidance system developed in the VALU3S project has demonstrated adherence to SCP requirements.
- Selection and testing of one SCP requirement on UC6:
  - SW Component testing on the parallel guidance system
  - Kalman filter-based Fault detector
  - Intrusion detection for WNS
  - Wireless interface network security assessment

Therefore, there are multiple demonstrable items that are covering defined challenges and scenarios, and partially covering VALU3S dimensions (see Table 3-56).

Table 3-56 Overview of demonstration prepared by UC6 partners.

Item#	Demo Name	Description/Purpose	Type	Responsible
1	Risk analysis with RAMSES tool	Identifying risk scenarios according to ISO 12100 standard, to evaluate them in terms of likelihood and severity using the so-called "risk graph" foreseen in the standard and to define safety measures	Lead demonstrator Demo using RAMSES tool	STAM
2	MSA-FLA with CHESS-FLA	Demonstration of the application of the CHESS-FLA tool on the UC6 system. Starting from the designed functional model of the systems, we show how to apply the Failure Logical Analysis and to automatically compute the FMEA table and the Fault Trees.	Lead demonstrator Demo of the CHESS-FLA plug-in	INTECS
3	SW component testing for functional software behaviour	Validate the functional behaviour of wifi to CAN dongle	Complementary demonstrator Report	ESTE
4	IEE 802.15.4 wireless sensor network – Intrusion Detection	Intrusion detection for Wireless Sensor Network. The WSN detects intrusion attempts, notifying via CAN bus and eventually stopping the AgriRobot actions.	Lead demonstrator (on-bench setup)	UNIVAQ
5	Data-driven Fault Detector	Use of data-driven modelling on historical dataset collected from nominal behaviour of the AgriRobot, thus generating dynamical models based on the responses. Setup Kalman filter where its state consists on the parameters of the nominal model, and check the deviation of such parameters from the nominal behaviour to detect faults.	Complementary demonstrator Report	UNIVAQ
6	Machine learning methods based on rules	Improve model checking effectiveness in the implementation of safety requirements in agricultural robot.	Complementary demonstrator Report	RULEX
7	Radio-link security of agricultural robot	Evaluating the security of the proprietary protocol, testing different attacks on the communication channel and evaluating online intrusion detection strategies	Complementary demonstrator Video	UNIGE

3.6.5 Quantitative Results

List of evaluation criteria of the SCP process:

1. Eval_SCP_3 –Number of Malicious Attacks and Faults Detected. This criterion is used to quantify the number of malicious attacks among the ones carried out that have been

successfully detected and did not cause any malfunction to the attacked system. (ESTE, UNIVAQ)

- The data collected from the system are used to derive a linear or nonlinear model of the system to describe its dynamics. The dynamical behaviour of such a system is used together with a Kalman filter to check if a change in the evolution of the dynamics occurred. In particular, a PCA model based on ARX modelling, extended with Poly-Exponential modelling if necessary, is built based on the collected data, and the ARX parameters describing the system are used as the state of a Kalman filter. Then, the analysis of the state evolution of the Kalman filter corresponds to the analysis of the dynamic behaviour of the system. If a fault/attack is occurring on the system, it produces a change in the dynamical behaviour, and thus in the Kalman filter evolution. Thus, by monitoring the state of the filter, and comparing it with thresholds defined ad-hoc based on the system under study, it is possible to detect whether a fault/attack is occurring. Tests to detect attacks and faults have been designed and conducted on the use case data. No previous results are available. Results have shown that the faults simulated on the agrirobot (GPS and AHRS) have been detected. (UNIVAQ)
- In the UC6, the agrirobot is equipped with a WSN node which is connected to the CAN interface. During intrusions/attacks, the robots should stop their activity to ensure that the safety and security requirements are satisfied. To evaluate an attack scenario and the correct response of the system, we simulate intrusion/attacks to the WSN motes located outside the robot operation area. When intrusions/attacks are detected, the WSN motes send a notification to the robots' WSN gateway which sends an alert via its CAN interface. Since there is not always possible to test a real operation scenario, we also simulate the agrirobot robot using an on-bench replica of the robot. 0% of attacks undetected (UNIVAQ)

2. Eval_SCP_5–Potential Impact of Incidents and Attacks. This criterion will be used to evaluate the potential impact of the implemented attacks on the normal operations of the attacked system. (ESTE, INTECS, UNIGE, STAM)
    - It is possible to estimate the potential impact of attacks or incidents on what we have defined as system outputs (robot movement, etc) through the analysis of the FMEA table and FTs generated from the Failure Logical Analysis. Also, it is possible to compute the probability that specific attacks or incidents can occur. (INTECS)
3. Eval_SCP_12–Number of Attack/Incident Typologies Examined. This criterion will be used to evaluate the risk analysis method. (ESTE, INTECS, STAM).
    - Evaluation has been done by measuring the hazards considered with ESTE traditional risk analysis tool (i.e., an Excel file) vs the hazards implemented in RAMSES taxonomy. The number of attack/incident typologies examined within the risk analysis process of the agriculture robot has been recorded before and after the usage of the RAMSES tool. Before VALU3S, indeed, risk analysis was done manually with an Excel file while now it is undertaken through the RAMSES web app. The number of hazards considered has been increased from 50 to 84 (on average), i.e., +68%.
    - Different injected faults and failures are estimated for the modelled system in terms of impact on the system outputs (INTECS).

List of evaluation criteria for V&V:

1. Eval_VV_8 –Effort Needed for Test. This metric will be used for evaluating the impact of SW component test tools (ESTE)
    - Software component testing using open-source SW framework adapted during the VALU3S project to be executed directly in the target device instead of being executed on PC (that means testing a code compiled for a different HW). This permits to execution SW test and HW/SW integration test (requested by ISO 26262 [34] and ISO 25119 [35]) improving and reducing the effort needed for testing VALU3S' VV evaluation criteria. The estimation of the minimum effort required by the user to perform software component testing and hardware-software integration test in the same test step is evaluated in –10% of the time based on less testbench to be prepared and less code modification needed to perform the component testing directly on the device under test;
2. Eval_VV_10 –Reduced Cost and Time for Work on Certification Process and Functional Safety. This metric will be used for evaluating the impact of model-based safety analysis tools (ESTE, INTECS)
    - The estimation of the minimum effort required by the user for the generation of the safety analysis artefacts (FMEA table and FTs) will be provided with a short demo. The automatic artefact generation will facilitate and improve the safety analysis in the early stages of development (INTECS).
3. Eval_VV_12–Effort Required to the User for Prepare and Running the Tool. This criterion will be used to evaluate the risk analysis method (ESTE, INTECS, STAM)
    - Evaluation has been done by recording the effort (in terms of man-hours) needed to carry out risk analysis through the ESTE traditional tool (i.e. an Excel file) vs using the RAMSES tool. The effort needed to conduct a risk analysis of the agriculture robot has been recorded before and after the implementation of the RAMSES tool. Before VALU3S, risk analysis was done with an Excel file, while now RAMSES is used to create scenarios, evaluate risk and apply safety measures to mitigate risk. A reduction of the effort needed to carry out risk analysis has been recorded, namely from 120 to 80 (-33%)
    - The estimation of the minimum effort required by the user for the generation of the safety analysis artefacts (FMEA table and FTs) will is provided with a short demonstration (INTECS).

The bar charts in Figure 3-28 show the overall UC6 results regarding the evaluation criteria, based on the result described in the section above. The usage of the risk analysis tool RAMSES increases the number of risks analysed decreasing the effort needed to perform the risk analysis. ArmUnity tool merges two different test steps in one reducing the effort needed for the test.

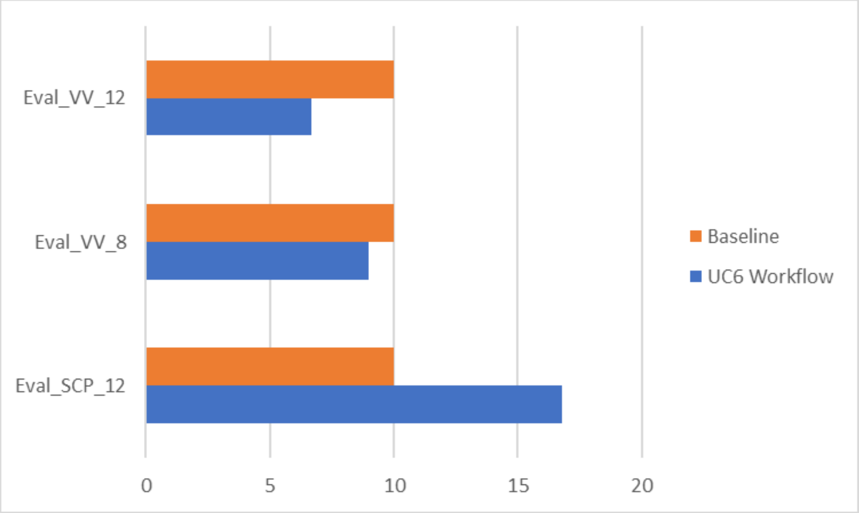


Figure 3-28 Quantitative results chart for UC6 demonstrators (as a whole)

3.6.6 Qualitative Results

Demonstrator 1: MSA-FLA with CHESS-FLA

Participants Profile: QAM is applied to 10 subjects (9 Males, 1 female) aged in the range of 24-44. The education level is relatively high as the subject pool is composed of 1 PhD researcher and 9 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as “financial safety engineers, R&D engineers, software/hardware engineers, etc.” having experience in the fields of “automotive, agricultural electronics, functional safety, semantic web, IoT, digital twin etc.”.

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-57. The results show that all constructs are correlated with each other.

Table 3-57. UC6 – Demonstrator 1 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.526	1									
MO	0.712	0.910	1								
CO	0.600	0.878	0.878	1							
ROI	0.706	0.907	0.968	0.846	1						
PE	0.479	0.600	0.684	0.607	0.750	1					
PT	0.577	0.840	0.907	0.726	0.931	0.576	1				
PR	0.270	0.455	0.546	0.459	0.563	0.450	0.600	1			
SI	0.261	0.567	0.495	0.225	0.528	0.339	0.675	0.468	1		
ATU	0.317	0.512	0.615	0.466	0.537	0.335	0.530	0.333	0.198	1	
BI	0.710	0.792	0.903	0.675	0.850	0.662	0.755	0.301	0.477	0.694	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-58, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-58. UC6 - Demonstrator 1 Reliability Analysis

Cronbach-alpha values	
PU	0.134
PEOU	0.612
MO	0.208
CO	0.780
ROI	0.719
PE	0.362
PT	0.934
PR	0.260
SI	0.644
ATU	0.568
BI	0.102

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-59. For this demonstrator, there exists a right proportional relation between all constructs influencing each other in the same direction.

Table 3-59. UC6 - Demonstrator 1 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = -0.08xPU + 4.78	0.003	0.878	0.159
H2	CO-PU	Right	CO = 0.058xPU + 5.42	0.008	0.81	-0.248
H3	PEoU-PU	Right	PEoU = -0.23xPU + 4.09	0.041	0.576	0.583
H4	PU-ATU	Right	PU = -0.15xATU + 4.04	0.047	0.548	0.627
H5	PEoU-ATU	Right	PEoU = -0.13xATU + 4.2	0.027	0.65	0.471
H6	ATU-BI	Right	ATU = -0.79xBI + 1.21	0.344	0.075	2.048
H7	ROI-BI	Right	ROI = -0.42xBI + 3.17	0.2	0.196	1.412
H8	PE-BI	Right	PE = -0.183xBI + 4.06	0.029	0.636	0.492
H9	SI-BI	Right	SI = -0.56xBI + 2.42	0.398	0.05	2.301
H10	PT-BI	Right	PT = -0.88xBI + 0.47	0.541	0.015	3.069
H11	PR-BI	Right	PR = -0.234xBI + 3.98	0.541	0.458	0.78

Demonstrator 2: Arm Unity

Participants Profile: QAM is applied to 10 subjects (9 Males, 1 unknown) aged in the range of 24-44. The education level is relatively high as the subject pool is composed of 1 Post-Doc or higher degree and 1 PhD researcher and 7 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as “CTO, embedded system engineer, software/hardware engineer etc." having experience in the fields of "cloud, embedded systems, functional safety, semantic web, IoT, digital twin etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-60. The results show that PU is not correlated with, PEOU, MO, CO, ROI, PE and PT.

Table 3-60. UC6 – Demonstrator 2 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	-0.244	1									
MO	-0.437	0.953	1								
CO	-0.218	0.984	0.950	1							
ROI	-0.204	0.980	0.958	0.984	1						
PE	-0.218	0.987	0.956	0.994	0.985	1					
PT	-0.124	0.952	0.916	0.971	0.982	0.967	1				
PR	0.003	0.922	0.846	0.947	0.947	0.956	0.961	1			
SI	0.188	0.819	0.703	0.870	0.859	0.858	0.916	0.957	1		
ATU	0.082	0.899	0.801	0.926	0.925	0.931	0.957	0.993	0.979	1	
BI	0.004	0.936	0.845	0.963	0.936	0.960	0.957	0.977	0.950	0.979	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-61, the questions asked to subjects are sufficiently reliable as understood from subject responses, except answers related to PU.

Table 3-61. UC6 - Demonstrator 2 Reliability Analysis

Cronbach-alpha values	
PU	-0.712
PEOU	0.703
MO	0.848
CO	0.118
ROI	0.277
PE	0.555
PT	0.154
PR	0.112
SI	0.012
ATU	0.568
BI	0.123

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-62. For this demonstrator, there exists a right proportional relation between all constructs influencing each other in the same direction.

Table 3-62. - UC6 - Demonstrator 2 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.08xPU + 5.16	0.191	0.206	-1.375
H2	CO-PU	Right	CO = 0.039xPU + 5.04	0.048	0.545	-0.632
H3	PEoU-PU	Right	PEoU = 0.05xPU + 5.06	0.059	0.498	-0.71
H4	PU-ATU	Right	PU = 0.44xATU + 0.34	0.007	0.821	0.233
H5	PEoU-ATU	Right	PEoU = 0.933xATU - 0.06	0.808	0	5.807
H6	ATU-BI	Right	ATU = 0.99xBI + 0.17	0.958	0	13.451
H7	ROI-BI	Right	ROI = 0.93xBI - 0.17	0.877	0	7.542
H8	PE-BI	Right	PE = 0.866xBI - 0.08	0.921	0	9.647
H9	SI-BI	Right	SI = 0.84xBI + 0.46	0.903	0	8.623
H10	PT-BI	Right	PT = 0.96xBI + 0	0.916	0	9.33
H11	PR-BI	Right	PR = 0.986xBI + 0.1	0.916	0	12.887

Demonstrator 3: SDR-based Radio link interface Security Assessment

Participants Profile: QAM is applied to 10 subjects (10 Males) aged in the range of 24-34. The education level is relatively high as the subject pool is composed of 1 PhD researcher and 9 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "R&D engineers, software/hardware engineers etc." having experience in the fields of "AI/ML, cyber-security, embedded systems, 3D visualisation, agriculture etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-63. The results show that PEOU-CO, CO-ROI, SI-ROI, ATU-ROI and ATU-PEOU pairs are not highly correlated with each other.

Table 3-63. UC6 – Demonstrator 3 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.631	1									
MO	0.535	0.720	1								
CO	0.264	-0.140	0.359	1							
ROI	0.367	0.315	0.070	-0.267	1						
PE	0.518	0.379	0.565	0.719	0.199	1					
PT	0.129	0.012	0.004	0.196	0.390	0.463	1				
PR	0.502	0.074	0.142	0.686	0.196	0.739	0.638	1			
SI	0.486	0.220	0.228	0.481	-0.013	0.737	0.161	0.473	1		
ATU	0.196	-0.101	0.503	0.728	-0.177	0.460	0.097	0.204	0.238	1	
BI	0.670	0.312	0.294	0.423	0.498	0.783	0.717	0.746	0.656	0.324	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-64, the questions asked to subjects are mostly reliable as understood from subject responses except for responses to PU and MO questions.

Table 3-64. UC6 – Demonstrator 3 Reliability Analysis

Cronbach-alpha values	
PU	-0.261
PEOU	0.113
MO	-0.915
CO	0.252
ROI	0.738
PE	0.436
PT	0.872
PR	0.514
SI	0.394
ATU	0.371
BI	0.026

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-65. For this demonstrator, there exists a right proportional relation between all constructs influencing each other in the same direction except the MO-PU pair.

Table 3-65. UC6 – Demonstrator 3 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Inverse	MO = -0.61xPU + 2.03	0.286	0.111	1.79
H2	CO-PU	Right	CO = 0.315xPU + 3.62	0.07	0.461	0.774
H3	PeoU-PU	Right	PeoU = 0.82xPU + 0.61	0.398	0.051	2.298
H4	PU-ATU	Right	PU = 0.11xATU + 4.53	0.039	0.587	0.566
H5	PeoU-ATU	Right	PeoU = 0.076xATU + 5.53	0.01	0.781	-0.287
H6	ATU-BI	Right	ATU = 0.4xBI + 3.17	0.105	0.36	0.97
H7	ROI-BI	Right	ROI =0.47xBI + 2.65	0.248	0.143	1.626
H8	PE-BI	Right	PE = 0.565xBI + 2.33	0.613	0.007	3.562
H9	SI-BI	Right	SI = 0.81xBI + 0.87	0.431	0.039	2.461
H10	PT-BI	Right	PT = 0.54xBI + 2.32	0.514	0.02	2.908
H11	PR-BI	Right	PR = 0.628xBI + 2.03	0.514	0.013	3.169

Demonstrator 4: RAMSES tool for Risk Analysis of Agriculture Robot following ISO12100 standard

Participants Profile: QAM is applied to 10 subjects (9 Males, 1 female) aged in the range of 24-34. The education level is relatively high as the subject pool is composed of 1 PhD researcher and 9 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "R&D engineers, software/hardware engineer, Q&A etc." having experience in the fields of "AI/ML, digital twin, semantic web, IoT, agriculture, automotive etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-66. The results show that PU is not fully correlated with other constructs except PEOU, MO and CO.

Table 3-66. UC6 – Demonstrator 4 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.324	1									
MO	0.172	0.885	1								
CO	0.252	0.641	0.634	1							
ROI	-0.267	0.604	0.683	0.768	1						
PE	-0.168	0.511	0.664	0.669	0.810	1					
PT	-0.454	0.263	0.485	0.493	0.783	0.891	1				
PR	-0.452	0.289	0.527	0.579	0.833	0.893	0.972	1			
SI	-0.537	0.291	0.526	0.536	0.873	0.864	0.944	0.976	1		
ATU	-0.477	0.283	0.535	0.546	0.848	0.895	0.964	0.984	0.993	1	
BI	-0.474	0.325	0.567	0.582	0.882	0.867	0.922	0.975	0.993	0.987	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-67, the questions asked to subjects are sufficiently reliable except for the construct PU.

Table 3-67. UC6 - Demonstrator 4 Reliability Analysis

Cronbach-alpha values	
PU	-0.345
PEOU	0.433
MO	0.232
CO	0.512
ROI	0.445
PE	0.233
PT	0.381
PR	0.461
SI	0.345
ATU	0.522
BI	0.222

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-68. For this demonstrator, there exists a right proportional relation between all constructs influencing each other in the same direction except the MO-PU pair. There is a nearly direct relation between SI and BI in this demonstrator.

Table 3-68. UC6 - Demonstrator 4 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Inverse	MO = -0.24xPU + 4.87	0.03	0.634	0.495
H2	CO-PU	Right	CO = 0.172xPU + 5.21	0.063	0.483	0.735
H3	PEoU-PU	Right	PEoU = 0.31xPU + 4.36	0.105	0.362	0.968
H4	PU-ATU	Right	PU = 1.73xATU + 15.64	0.228	0.163	-1.537
H5	PEoU-ATU	Right	PEoU = 0.987xATU - 0.75	0.08	0.428	0.835
H6	ATU-BI	Right	ATU = 0.97xBI + 0.07	0.974	0	17.426
H7	ROI-BI	Right	ROI = 2.8xBI - 9.88	0.778	0.001	5.3
H8	PE-BI	Right	PE = 1.868xBI - 4.66	0.752	0.001	4.926
H9	SI-BI	Right	SI = 1xBI + 0	0.986	0	23.628
H10	PT-BI	Right	PT = 0.9xBI + 0.61	0.849	0	6.716
H11	PR-BI	Right	PR = 0.934xBI + 0.21	0.849	0	12.374

Demonstrator 5: Data-driven Fault Detector

Participants Profile: QAM is applied to 10 subjects (10 Males) aged in the range of 25-51. The education level is relatively high as the subject pool is composed of 2 Post-Doc or higher-degree and 1 PhD researcher and 7 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "academicians, researchers, R&D engineers, software/hardware engineers, etc." having experience in the fields of "AI/ML, cyber-security, embedded systems, 3d visualisation etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-69. The results show that PCO is not correlated with ROI, PT, or SI. MO and SI pairs are not correlated either.

Table 3-69. UC6 – Demonstrator 5 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.606	1									
MO	0.110	0.255	1								
CO	0.101	0.704	0.197	1							
ROI	0.190	0.737	0.184	0.680	1						
PE	0.202	0.279	0.850	0.153	0.442	1					
PT	0.132	0.184	0.499	-0.009	0.255	0.619	1				
PR	0.594	0.463	0.514	0.213	0.530	0.742	0.405	1			
SI	0.472	0.474	-0.068	-0.093	0.217	0.068	0.574	0.142	1		
ATU	0.493	0.906	0.379	0.699	0.841	0.546	0.369	0.599	0.443	1	
BI	0.457	0.677	0.373	0.606	0.618	0.562	0.573	0.576	0.457	0.871	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-70, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-70. UC6 – Demonstrator 5 Reliability Analysis

Cronbach-alpha values	
PU	0.492
PEOU	0.38
MO	0.418
CO	0.444
ROI	0.99
PE	0.098
PT	0.093
PR	0.26
SI	0.034
ATU	0.272
BI	0.167

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-71. For this demonstrator, there exists a right proportional relation between all constructs influencing each other in the same direction except the MO-PU pair. There is a nearly direct relation between SI and BI in this demonstrator.

Table 3-71. UC6 - Demonstrator 5 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Inverse	MO = -0.18xPU + 5.1	0.012	0.761	0.314
H2	CO-PU	Right	CO = 0.096xPU + 5.7	0.01	0.781	0.287
H3	PEoU-PU	Right	PEoU = 0.65xPU + 2.26	0.367	0.063	2.155
H4	PU-ATU	Right	PU = 0.54xATU + 2.41	0.243	0.148	1.601
H5	PEoU-ATU	Right	PEoU = 1.065xATU - 0.74	0.82	0	6.046
H6	ATU-BI	Right	ATU = 0.63xBI + 1.96	0.759	0.001	5.014
H7	ROI-BI	Right	ROI = 0.42xBI + 3.08	0.382	0.057	2.224
H8	PE-BI	Right	PE = 0.585xBI + 1.98	0.315	0.091	1.92
H9	SI-BI	Right	SI = 0.49xBI + 2.75	0.209	0.184	1.453
H10	PT-BI	Right	PT = 0.48xBI + 2.74	0.328	0.083	1.978
H11	PR-BI	Right	PR = 0.576xBI + 2.27	0.328	0.081	1.995

3.6.7 Observed Limitations, Lessons Learnt and Best Practices

Concerning the Model-based Safety Analysis with Failure Logical Analysis method, supported by the CHESS-FLA tool, INTECS, together with ESTE, agree that the greatest effort on the application of the method falls back on the design of the functional model of the system and on the enrichment of this functional model with the decorations required to apply the Failure Logical Analysis. Once the extended functional model of the system has been designed, there is a huge reduction in the time and effort needed to compute the Failure Mode and Effect Analysis (FMEA) table and the Fault Trees, which are commonly used artefacts in the context of the safety analysis. CHESS-FLA automatically computes these artefacts in a couple of seconds, compared to a couple of weeks usually need to manually compute them. Furthermore, the automatic computation allows also the reduction of the error processes and the possibility to miss some relevant error propagation paths.

Arm Unity is based on Serial Wired Debug (SWD) interface with a Serial Wired Output extension, developed by ARM for its microcontroller. Other vendors' microcontrollers, that do not implement the SWD interface, are not able to use the Arm Unity tool.

Methods based on a data-driven model require data collected on the field in all possible operating conditions; in the VALU3S project, ESTE collected the data in an ideal environment and the model developed may be integrated with other operation conditions (e.g., off-road, on a slope, after heavy rains) to improve the model; after that, the methods based data-driven model has to be evaluated and in case re-design the way to be applied in the agricultural robot.

The mean and standard deviation of experts' responses to UC6 demonstrations are given in Table 3-72, Table 3-73, Table 3-74, Table 3-75 and Table 3-76. The overall results show that the respondents are confident with the developed technology as they fairly accept the proposed solution stack. Especially for the Demonstration 3, 4, and 5, relatively high QAM results are reported. These results indicate that the proposed fault detection, SDR-based security assessment and risk analysis techniques are well-adopted. For the first two demonstrations, the QAM factors are still widely accepted, not as much as the last three demonstrations, but the overall assessment indicates that the users may still have a positive attitude toward using the solution stack. One of the main reasons for the variation between the highly- and fairly-adopted demonstrations can be the diverse nature of the use case. UC6 is one of the most complex use cases and five demonstrations are evaluated by the experts. Especially external experts might not understand the functionalities and benefits of the entire set of innovations. There exist evaluation scenarios and criteria that should be digested in a limited time. Nevertheless, the hardware and software utilities strengthen the level of user awareness. Such tangible outputs yield a certain positive attitude towards using the solution stack.

Table 3-72 Mean and standard deviation of experts' responses to UC6 - Demonstration 1

UC6-Demonstrator-1	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,17	4,66	4,92	4,39	4,42	5,21	5,17	4,42	4,60	4,81	5,02
Std Dev	0,82	0,71	0,59	1,24	0,82	0,71	0,64	0,87	0,85	0,57	0,76

Table 3-73 Mean and standard deviation of experts' responses to UC6 - Demonstration 2

UC6-Demonstrator-2	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	4,93	4,49	4,76	5,05	4,73	5,74	5,04	4,98	4,78	4,50	4,54
Std Dev	0,46	0,33	0,63	0,57	1,55	0,90	1,00	1,39	1,84	1,44	1,06

Table 3-74 Mean and standard deviation of experts' responses to UC6 - Demonstration 3

UC6-Demonstrator-3	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,18	5,54	5,19	4,95	5,40	5,08	5,28	5,04	5,36	5,11	5,20
Std Dev	0,57	0,44	0,50	0,48	0,42	0,55	0,53	0,47	0,32	0,33	0,40

Table 3-75 Mean and standard deviation of experts' responses to UC6 - Demonstration 4

UC6-Demonstrator-4	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	6,17	5,79	5,49	5,55	5,27	5,11	5,25	5,60	5,49	5,67	4,88
Std Dev	0,49	0,51	0,36	0,72	0,55	0,81	0,66	0,51	0,32	0,58	1,75

Table 3-76 Mean and standard deviation of experts' responses to UC6 - Demonstration 5

UC6-Demonstrator-5	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	6,24	6,12	6,33	5,58	5,88	6,14	5,96	5,73	5,77	5,78	5,58
Std Dev	0,70	0,65	0,43	0,74	0,81	0,53	0,67	0,55	0,52	0,77	0,55

3.7 Use Case 7 – Human-Robot Collaboration in a Disassembly Process with Workers with Disabilities (UC7)

UC7 from ALDAKIN targets a collaborative robotic cell for the removal of refrigerators' magnetic gaskets in a human-robot interaction context, in which the co-workers are assumed to have a kind of disability. In a first proof-of-concept, within an identical simulated setting but with simplified disassembly parts, the system applies machine learning techniques, namely reinforcement learning for grasping and removing a peg representing the gasket, and employs two cameras, with zenithal and pitched angles to identify the workspace and track the worker's position relative to the robot, respectively. Figure 3-29 shows the actual and simulated disassembly plant. In the latter, the corresponding evaluations are carried out.

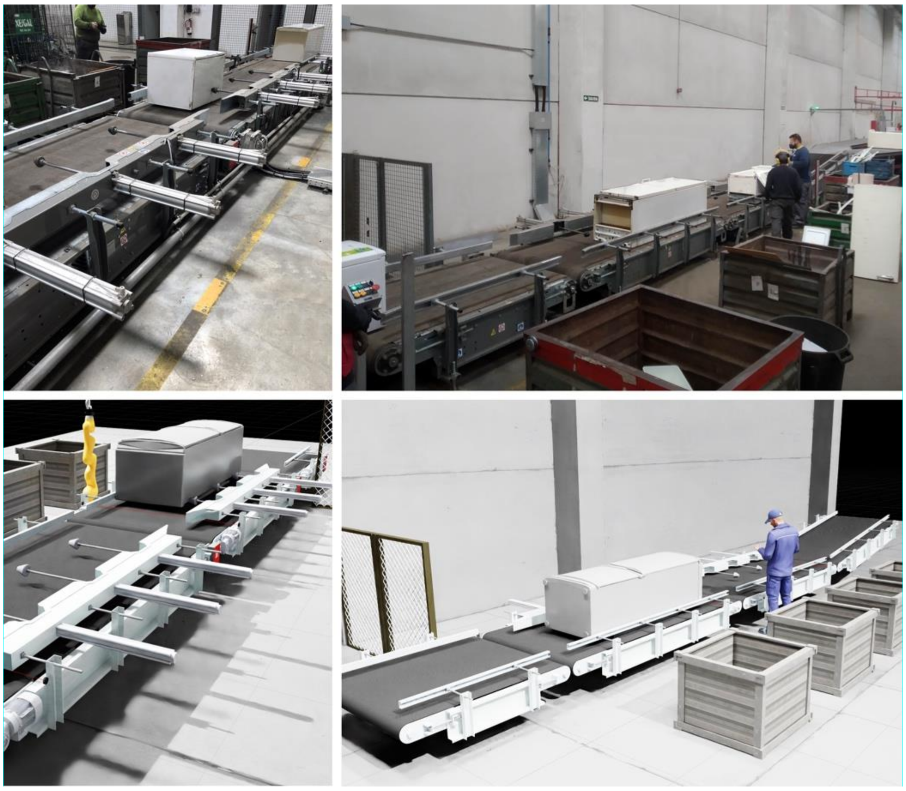


Figure 3-29 Actual disassembly plant (top) and simulated scenario (bottom).

3.7.1 V&V challenges

Within this use case, some V&V challenges arise based on the defined evaluation scenarios and the corresponding baseline.

1. Mapping of safety requirements to each oracle to provide a real-time and automated verdict during the test and validation of critical requirements related to different standards.
   - Such situations can be covered by establishing a list of those requirements in advance to be fulfilled by the system to perform the validation of each safety requirement.
2. Connection of nodes, interoperability among systems and coordination of testing and evaluation through oracle.
   - A specific library has been developed that connects and coordinates the execution of tests and the validator (oracle), also enabling the automatic selection of validation constraints without human input.

3.7.2 Contributors

Partners contributing to the UC: ALDAKIN, MGEP

3.7.3 Contributors' Roles & Evaluation Scenario

Evaluation scenarios defined for this use case are listed below; the assignments of the contribution of all UC7 partners can be found in Table 3-77.

1. VALU3S_WP1_Industrial_15 – Worker position/action monitoring.
2. VALU3S_WP1_Industrial_18 – AI capabilities to work in the system.

Table 3-77 Overview of contribution to evaluation scenarios by UC7 partners

Evaluation scenario	ALDAKIN	MGEP
VALU3S_WP1_Industrial_15	X	X
VALU3S_WP1_Industrial_18	X	X

Individual UC7 partners have contributed are contributing to the evaluation scenarios in the following way:

- ALDAKIN, as a use case provider, supports MGEP and provides appropriate pieces of code. Also, cooperation on the evaluation process has been done.
- MGEP carries out an improvement of the V&V methods: Coordination of Test Generation and Validation in simulation-based Human-Robot Collaborative environments (CTGV-HRC) and Simulation-Based Testing for Human-Robot Collaboration employing Constraint-Based Oracles (SBT-CBO). Both methods are part of the HuRoCTest tool that has been used within UC7 in each of the demonstrator evaluation scenarios.

This use case deals with testing and verification of the safety of model-free intelligent agents (reinforcement learning). Figure 3-30 shows the V&V tools that have been developed and used/demonstrated in this use case as well as the V&V methods associated with the tools (see D4.5 [4] for more details).

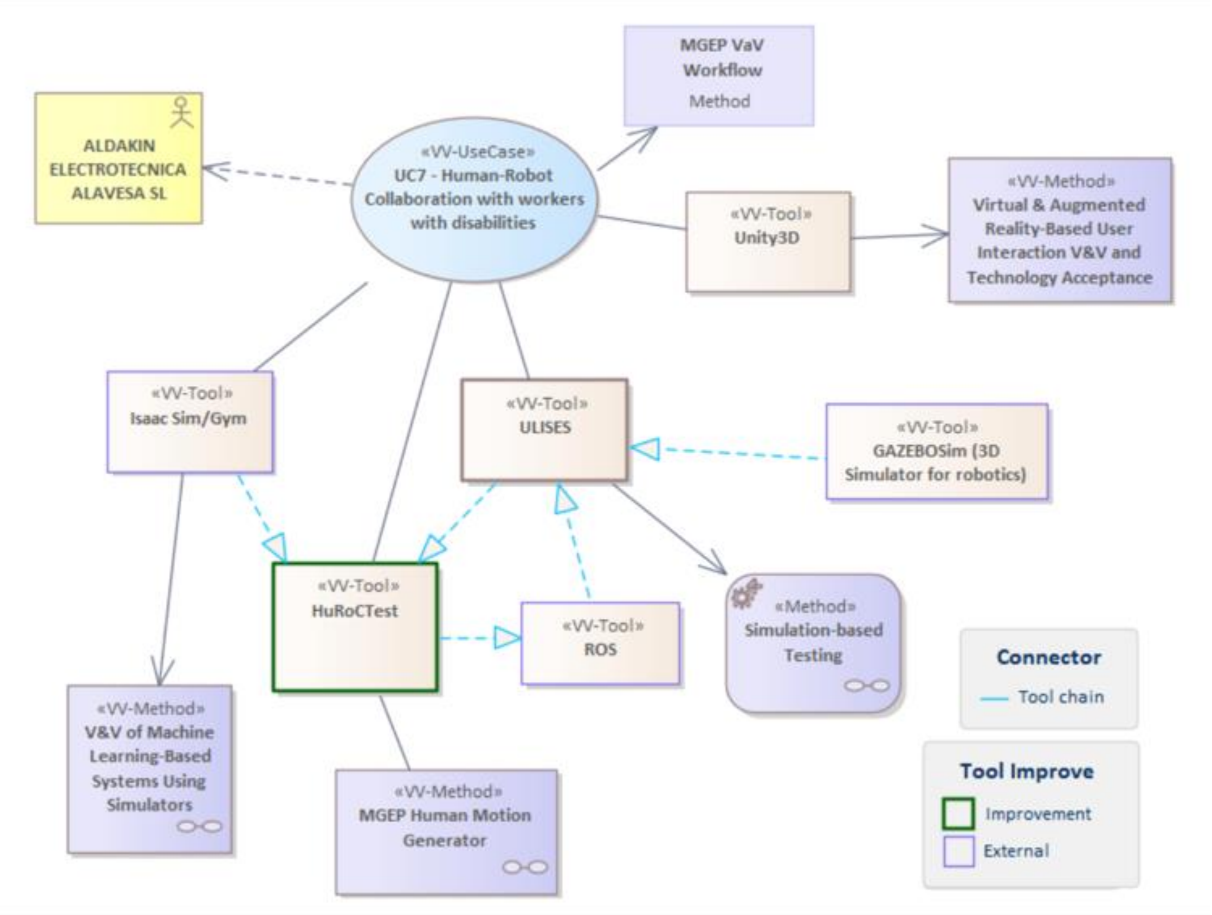

MGEP VaV Workflow
Method
ALDAKIN ELECTROTECNICA ALAVESA SL
«VV-UseCase» UC7 - Human-Robot Collaboration with workers with disabilities
«VV-Tool» Unity3D
«VV-Method» Virtual & Augmented Reality-Based User Interaction V&V and Technology Acceptance
«VV-Tool» Isaac Sim/Gym
«VV-Tool» ULISES
«VV-Tool» GAZEBOSim (3D Simulator for robotics)
«VV-Tool» HuRoCTest
«VV-Tool» ROS
«Method» Simulation-based Testing
Connector
Tool chain
«VV-Method» V&V of Machine Learning-Based Systems Using Simulators
«VV-Method» MGEP Human Motion Generator
Tool Improve
Improvement
External

Figure 3-30 Tools and Methods for UC7 – Human-Robot Collaboration in a Disassembly Process with Workers with Disabilities.

3.7.4 Demonstration

UC7 demonstration employs the HuRoCTest tool. This tool coordinates simulation-based testing activity in human-robot interaction environments. Specifically, the HuRoCTest tool provides an automated real-time verdict of test execution of human-robot interaction simulation environments employing constrained-based oracles. To coordinate testing with the constrained-based oracle, HuRoCTest leverages a ROS package to seamlessly align the execution of the test, the simulation environment, and the oracle. Therefore, the UC7 demonstrator is split into three sub-demos (see Table 3-78).

Table 3-78. Overview of demonstration prepared by UC7 partners.

Item #	Demonstration name	Description/Purpose	Format	Responsible
1	Disassembly plant scenario in MGEP simulator sub-demo	Show the system functionalities and settings of environment/scenario parameters to support V&V requirements	Lead Demo using NVIDIA ISAAC SIM	MGEP, ALDAKIN

Item #	Demonstration name	Description/Purpose	Format	Responsible
2	Human-disassembly collaboration scenario in MGEP simulator sub-demo	Show how we coordinated a batch of tests with different human motions to later evaluate the performance of the reinforcement learning agent under different situations	Complementary Demo using NVIDIA ISAAC SIM	MGEP, ALDAKIN
3	Constraint-based oracle sub-demo	Verify and validate the functionalities of the generated environment, as well as the actions taken by the reinforcement learning agent to support V&V requirements initially defined	Complementary Demo using NVIDIA ISAAC SIM/ULISES	MGEP

3.7.5 Quantitative Results

The objective of this demonstrator was to validate that by using the HuRoCTest tool, the effort cost of validating in simulation the robot control algorithm of a fridge disassembly system could be reduced. In addition, the use case has been used to validate the safety requirements concerning avoiding collisions between the robot and the human operator. The quantitative results, related to both the improvement of the V&V process and the improvement in the evaluation of SCP requirements due to the use of HuRoCTest are summarized below.

List of evaluation criteria of the V&V process:

1. Eval_VV_8 – Effort needed for test. This evaluation criterion aims to reduce human intervention in the test generation, selection, and prioritisation (if applicable), and the test execution and evaluation in simulations. The criterion is to facilitate and automate the definition of test artefacts to generate, execute and determine the outcome of each test (MGEP, ALDAKIN).
    - Before HuRoCTest, an engineer had to run the simulation and the different tests on a regular basis and then supervise the whole process. Now the system only has to be launched once and all tests are run automatically. ULISES coordinates the simulation through a communication library and provides the results.
    - During the project, eighteen tests, each lasting approximately one minute, have been generated. The developed tool reduces testing efforts by ~90% (sixteen minutes) as no human-in-the-loop is required. The remaining 10% is the time needed by the engineer to manage the setup and launch of the tests.

List of evaluation criteria for SCP (see details in D5.2 [8]):

1. Eval_SCP_2 – Number of safety/security requirements violations. Violations of safety requirements are counted. This value is the metric to analyse to what extent the robot controller policy (using reinforcement learning) as well as the workspace of the disassembly plant facilities comply with safety requirements in the created simulation environment (MGEP, ALDAKIN).

- Six out of the test cases defined, and fourteen out of the eighteen tests conducted assess the safety of the working environment. In all these tests, the safety is 100% assured. Regarding the reinforcement learning control policy, the success rates are 87.82% and 76.31% for extraction and collision avoidance simultaneously with static obstacles and moving workers, respectively. In both cases, around 97% of failed episodes are due to violations in the peg disassembly and less than 3% due to violations of safety requirements.

Figure 3-31 shows how the cost of testing effort has been reduced by 90%. On the other hand, at the beginning of the project, 30% of the safety requirements of the test case were evaluated, whereas now 100% of the requirements of the use case have been evaluated, increasing the coverage of the security requirements. However, it is worth mentioning that even though the validated safety requirements have increased, no new safety violations have been found. Therefore, Figure 3-31 shows the increase in the number of validated safety requirements, in other words, the increase in the coverage of safety requirements rather than the number of new safety violations found.

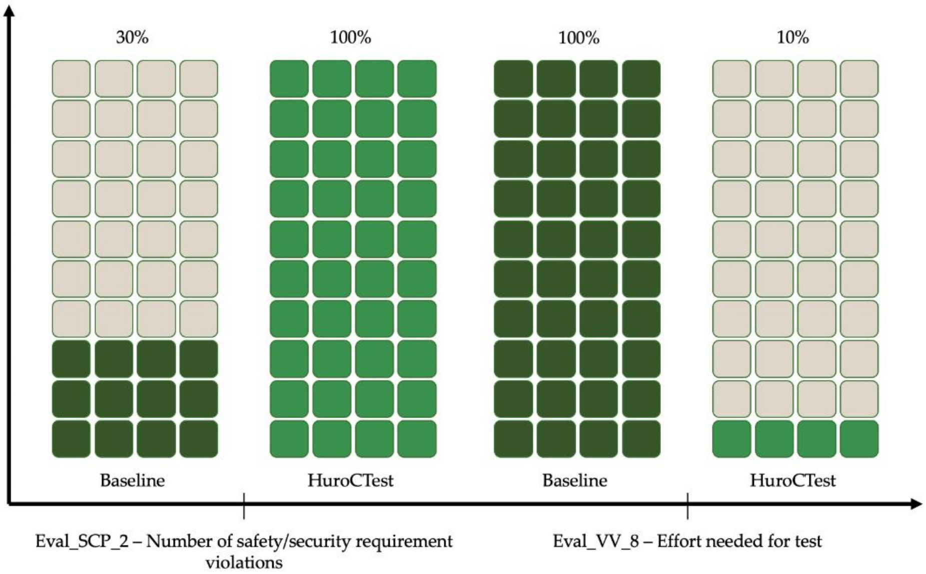


Figure 3-31: UC7 Demonstrator Evaluation Criteria Quantitative results

3.7.6 Qualitative Results

The qualitative assessment is applied at the overall use case level. The subjects' profile and the statistical analysis results are as follows:

Participants Profile: QAM is applied to 10 subjects (8 Males, 2 females) aged in the range of 24-44. The education level is relatively high as the subject pool is composed of 4 Post-Doc or higher-degree and 1 PhD researcher and 5 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "R&D engineers, researchers, professors, academicians, Q&A, etc." having experience in the fields of "human-robot interface, cyber-physical security, robotics etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-79. The results show that PU and SI are not correlated with the majority of the other constructs. PE-CO, BI-{ROI, SI, ATU} pairs are also not very correlated.

Table 3-79. UC7 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.673	1									
MO	0.224	0.483	1								
CO	-0.247	0.130	0.330	1							
ROI	-0.223	0.223	0.553	0.735	1						
PE	-0.114	0.168	0.531	-0.108	0.493	1					
PT	-0.198	0.276	0.245	0.752	0.809	0.253	1				
PR	-0.356	0.179	0.372	0.607	0.886	0.518	0.768	1			
SI	-0.176	-0.172	-0.664	-0.641	-0.399	0.037	-0.154	-0.163	1		
ATU	-0.156	0.411	0.220	0.580	0.573	0.034	0.709	0.580	-0.017	1	
BI	0.191	0.005	0.452	-0.076	0.291	0.636	0.183	0.318	-0.197	-0.229	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-80, the questions asked to subjects are mostly reliable as understood from subject responses, except the responses to the questions related to MO, CO and ATU.

Table 3-80. UC7 Reliability Analysis

Cronbach-alpha values	
PU	0.512
PEOU	0.222
MO	-1.969
CO	-1.570
ROI	0.589
PE	0.549
PT	0.495
PR	0.426
SI	0.358
ATU	-0.031
BI	0.514

Regression Analysis: Regression analysis is applied to estimating the relationships among QAM constructs, as seen in Table 3-81. For this demonstrator, there exists a right proportional relation between CO-PU, PU-ATU, ATU-BI and SI-BI pairs whereas other pairs are inversely influenced by each other.

Table 3-81. - UC7 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Inverse	MO = -0.13xPU + 5.14	0.05	0.533	0.651
H2	CO-PU	Right	CO = 0.202xPU + 6.84	0.061	0.491	-0.071
H3	PEoU-PU	Inverse	PEoU= -0.51xPU + 3.18	0.454	0.033	2.577
H4	PU-ATU	Right	PU = 0.26xATU + 6.45	0.024	0.667	-0.446
H5	PEoU-ATU	Inverse	PEoU= -0.511xATU + 2.27	0.169	0.237	1.277
H6	ATU-BI	Right	ATU = 0.27xBI + 7.03	0.053	0.524	-0.667
H7	ROI-BI	Inverse	ROI= -0.49xBI + 3.02	0.085	0.415	0.86
H8	PE-BI	Inverse	PE= -0.77xBI + 1.46	0.405	0.048	2.333
H9	SI-BI	Right	SI = 0.23xBI + 6.83	0.039	0.585	-0.569
H10	PT-BI	Inverse	MO = -0.21xBI + 4.72	0.034	0.612	0.527
H11	PR-BI	Inverse	MO = -0.277xBI + 4.29	0.034	0.37	0.949

3.7.7 Observed Limitations, Lessons Learnt and Best Practices

The use of HuRoCTest allows performing the validation of the control algorithm of a robot in a simulated human-robot collaboration system without the supervision of engineers. The main work of this project has been to integrate the results of the ULISES diagnostic system as assertions in the test suites and automate the execution and validation of test suites. By using the HuRoCTest the needed effort for the test has been reduced by 90% (Eval_VV_8 - Effort needed for test). In addition, the tests created have allowed validating in a simulation environment, without risk for the human operator, the safety requirement concerning obstacle avoidance (Eval_SCP_2- Number of safety/security requirements violations).

Although it has been shown that the HuRoCTest tool reduces the validation effort, it is important to note that when the simulation signals to be analysed change (e.g., a new use case) the ULISES tool must be adapted to receive these new signals. In the future, this should be possible through configuration files, so that it will not be necessary to create extensions or modify the ULISES tool code.

One of the initial goals of the project was also to generate a large number of different human movements and then integrate them into the test suites. Since the simulation tool, NVIDIA ISAAC SIM does not yet contain this functionality, the different human movements were generated using real motion captures. Finally, a set with the 5 basic movements performed by a person in the fridge disassembly process was recorded. This allowed the validation of the 5 movements of a specific person. In the future, it is necessary to extrapolate these movements to different typologies of people and create new movements using the basic set of movements as a base. Also, to create more realistic movements, in the future it is planned to record human movements using augmented reality in the simulation itself.

The qualitative assessment results present a relatively high acceptance of the technology. As seen in Table 3-82, PT and SI are relatively high. This is normal because UC7 encounters the needs of disabled and disadvantaged persons which can be linked with the social inclusion policies. The demonstration results show that the proposed solution can be trusted as it addresses (event) the adverse effects and ultimate needs of disadvantaged people. However, the attitude toward using the system is relatively lower than the other factors. Such a result can be explained by saying that there is no disabled person among the respondents, and they may not have sufficient empathy for the disabled workers.

Table 3-82 Mean and standard deviation of experts' responses to UC7

UC7	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,83	5,24	5,50	5,03	5,43	5,51	5,73	5,10	5,89	4,95	5,70
Std Dev	0,45	0,59	0,80	0,54	0,51	0,71	0,77	0,99	0,74	0,74	0,86

3.8 Use Case 8 – Infusion Controller of NMT (UC8)

An intelligent infusion controller for Vital Signs is a medical device that monitors the specific Vital Signs Parameter (e.g., Blood Pressure BP or Neuromuscular transmission NMT) to be regulated and infuses at regular intervals an updated drug dose value, to achieve a specific target value for the physiological value under control. VALU3S use case on Vital Signs Controller employing drug infusion is, by itself, a technological breakthrough in line with the robotic and automation of tasks within the O.R. (Operating Room). RGB, as the Use Case provider, is working on a family of controllers with innovative functionalities to be incorporated into a multi-parameter monitor. One of the objectives of VALU3S is to develop a HiL/SiL testbench platform that will allow the system to be verified under laboratory conditions. Figure 3-32 shows a schematic of the hardware components of the closed-loop controller for NMT and a simulation of the working environment.

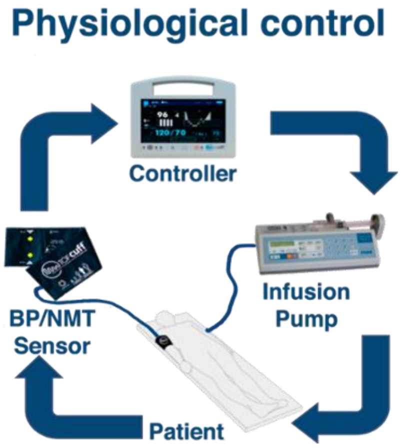


Figure 3-32 Closed Loop Controller of NMT

3.8.1 V&V challenges

The challenges that appeared during the implementation of the tool have been essentially the difficulty of making tests under varying conditions, and types of patients.

Solutions: Making tests of different types of tests. For this purpose, we have developed a Test Case Manager that allows us to introduce variables such as gender, age, sensitivity to the drug, etc.

The Test Case Manager statistically evaluates the time series of state variables outcoming from the Patient Model. Special care is taken for moments when the neuro-muscular blockade is less intense than required. Unexpected movements by the patients may interfere with the surgical process and must therefore be suppressed by providing an adequate level of NMT. A secondary goal is to minimize the total drug consumption of rocuronium, which beyond its economic relevance, it also may shorten the time needed for total recovery.

In addition, the use case has studied the potential benefit of new support for V&V tasks related to safety analysis, specification quality analysis, traceability management, and compliance management. Although RGB successfully performs these activities nowadays, the use of different methods and tools could lead to, e.g., a higher efficiency.

3.8.2 Contributors

Partners contributing to the UC: RGB, BUT, QRTECH, TRC, UCLM, INTECS.

3.8.3 Contributors' Roles & Evaluation Scenario

Contributors' roles are as follows:

- RGB, as the Use Case coordinator, provides information about the Use Case, gives support to other partners, and follows the activities carried out by partners. RGB develops an Industrial concept for the NMT controller and provides NMT technology for the use case.
- BUT contributes to VALU3S evaluation scenarios "HiL and SIL benchmark platform" and "Patient modelling with NMT drugs" by providing TCM (test case manager) to analyse the safety of the system. Together with RGB, BUT develops a HiL/SiL testbench tool for automated experimenting -at the laboratory level- with Patient Model and NMT Controller to assure the safety, efficiency and stability of the Controller.
- QRTECH cooperated in the TCM of the first demonstrator.
- TRC, in collaboration with UCLM, has worked on the application of knowledge-centric systems engineering solutions for system artefact quality analysis and traceability management, considering risk analysis results, requirements, and design models of the NMT controller.
- UCLM has dealt with process compliance aspects of the NMT controller by exploiting model-based techniques, in addition to the collaborative work with TRC mentioned above. For compliance, the IEC 62304 standard has been considered.
- INTECS: is dealing with customising the MBSA (Model-Based Safety Analysis) V&V method, in which the system and safety engineers share a common system model created using a model-based development process. By extending the system model with a fault model as well as relevant portions of the physical system to be controlled, automated support can be provided for safety analysis, in particular for FMEA and FTs generation.

The evaluation scenarios are as follows:

1. VALU3S_WP1_Healthcare_2 - Safety analysis and certification
2. VALU3S_WP1_Healthcare_3 - Certification needs of the NMT device

3. VALU3S_WP1_Healthcare_4 - HiL and SIL benchmark platform
4. VALU3S_WP1_Healthcare_5 - Patient modelling with NMT drugs
5. VALU3S_WP1_Healthcare_6 - Assurance needs of the NMT device

An overview of individual partners' contributions within evaluation scenarios can be found in Table 3-83.

Table 3-83 Overview of contribution to evaluation scenarios by UC8 partners

Evaluation Scenario	RGB	BUT	QRTECH	UCLM	INTECS	TRC
VALU3S_WP1_HealthCare_2 Safety analysis and certification	X				X	
VALU3S_WP1_HealthCare_3 Certification needs of the NMT device	X			X		X
VALU3S_WP1_HealthCare_4 HiL and SIL benchmark platform	X	X	X			
VALU3S_WP1_HealthCare_5 Patient modelling with NMT drugs	X	X	X			
VALU3S_WP1_HealthCare_6 Assurance needs of the NMT device	X			X		X

V&V Evaluation Criteria for evaluation of workflow activities or the whole workflow (process):

1. Eval_VV_2 – Coverage of Test Set – Aiming to enable better coverage of test set of infusion controller based on incorporating different types of patients and controlling conditions – no data for measurement are currently available.
2. Eval_VV_3 – Number of Test Cases – In order that results can be considered significative – data measurements are available, modifying the critical parameters of the control are currently available: A script file makes it possible to configure the patient´s parameters (weight, distribution volume, sensitivity to the drug EC50, etc) as well as the control characteristics (Initial dose, NMT target, duration of the operation...)
3. Eval_VV_5 – Joint Management of SCP Requirements – To identify and mitigate the main risk causes – no data for measurement are currently available. Some of the new methods and tools allow a user to deal with safety and security requirements at the same time. This has been considered when using the final prototype with an external board implementing the control algorithm, the infusion pump´s control and the alarm configurations.
4. Eval_VV_6 – Cost of Finding and Fixing a Coding Bug to make the Testbench platform most effective and efficient – estimations have been done when comparing the use of new tools with the previous situation, in which experimental work had to be done from the beginning.

5. Eval_VV_8 – Effort Needed for Test to identify ways of minimizing the effort required – estimations have been done when comparing the use of new tools with the previous situation, in which experimental work had to be done from the beginning.
6. Eval_VV_10 – Reduced Cost and Time for Work on Certification Process and Functional Safety as preparation for future needs (out of the project scope) – no data for measurement are currently available. Some of the new methods and tools can enable effort reduction in this context. Estimations have been done when comparing the use of new tools with the previous situation, in which experimental work had to be done from the beginning.
7. Eval_VV_13 – Reliability Measures of Decisions to get the best outcome in performance terms – data measurements are available, modifying the critical parameters of the control are currently available: A script file makes it possible to configure the patient´s parameters (weight, distribution volume, sensitivity to the drug EC50, etc) as well as the control characteristics (Initial dose, NMT target, duration of the operation...)

SCP Evaluation Criteria may be used for some of the workflow artefacts (products):

1. Eval_SCP_1 – Error Coverage – the metric is used for evaluating the percentage of time that "good" control is achieved – – data measurements are available, modifying the critical parameters of the control are currently available: A script file makes it possible to configure the patient´s parameters (weight, distribution volume, sensitivity to the drug EC50, etc) as well as the control characteristics (Initial dose, NMT target, duration of the operation...)
2. Eval_SCP_2 – Number of Safety/Security Requirement Violations - the metric is planned to be used for evaluating the level of deviation when the control is not within pre-established targets – Some of the new methods and tools allow users to identify safety- or security-related issues.
3. Eval_SCP_7 – Number of Prevented Accidents will be used to select the proper test cases to cope with specific operating conditions that can be the main cause of accidents –Some of the new methods and tools allow users to identify safety-related issues that could lead to accidents.

3.8.4 Demonstration

Application of (new / improved) V&V methods and workflows in relation to – see Table 3-84.

Table 3-84. Overview of demonstration prepared by UC8 partners.

Item #	Demonstration name	Description/Purpose	Format	Responsible
1	Testbench HiL for Virtual Simulation and control of NMT	Custom made tool to E**valuate** the system with expected values of NMT when dose infusion is delivered to the patient. Configuration of the patient parameters is carried out in each test case. Testbench platform where we can perform a large number of automated tests. The objective is to be able to select the control algorithm with best performance, in order to incorporate it in the controller prototype.	Lead Demo (Report and Demo) BUT simulation tool BUT controller tool RGB adapted Vital Sign Monitor including Infusion Pump controller Alarm settings	RGB BUT
2	Model-based Assurance and Certification	Compliance of the NMT controller has been analysed against part 5 (software development process) of the IEC 62304 standard. This part of the standard includes safety and security considerations and has been modelled with the SIAM tool. In total, 108 quality practices defined in the standard have been taken into account for compliance gap analysis.	Complementary SIAM tool demo	UCLM RGB
3	Knowledge-Centric Traceability Management	This demonstration has dealt with advanced traceability specification and automated trace discovery and verification. These activities have exploited ontologies and semantic information Risks analysis results, system requirements, and design models of the NMT controller have been taken into account. The artefacts address both safety and security considerations. System requirements and design models have been traced (44 traces). The traces addressed from the risk analysis results are those between types of hazard, causes, effects, and actions (428 traces).	Complementary Traceability Studio tool demo	UCLM TRC RGB

Item #	Demonstration name	Description/Purpose	Format	Responsible
4	Knowledge-Centric System Artefact Quality Analysis	The suitability of system artefacts in different formats has been assessed by exploiting ontologies and semantic information. Risks analysis results, system requirements, and design models of the NMT controller have been taken into account. The artefacts address both safety and security considerations. The system requirements were elicited from the actions of the risk analysis results. The analyses have covered: 21 types of hazard, 51 causes, 46 effects, and 84 actions from the risk analysis results; 42 textual system requirements, and; seven models in the form of Capella diagrams.	Complementary RQA tool demo	UCLM TRC RGB
5	MSA-FLA with CHESS-FLA	Demonstration of the application of the CHESS-FLA tool on the UC8 system. Starting from the designed functional model of the systems, we will show how to apply the Failure Logical Analysis and automatically compute the FMEA (Failure Mode and Effect Analysis) table and the Fault Trees.	Complementary Report and Demo	INTECS

3.8.5 Quantitative Results

The Testbench HiL for Virtual Simulation and control of NMT is presented in Figure 3-33.

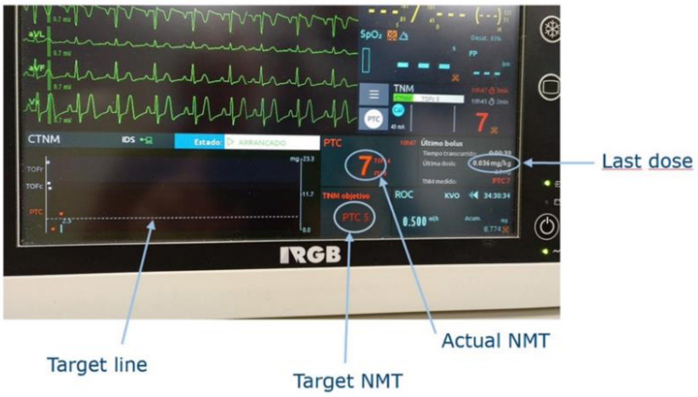


Figure 3-33 Simulation Of NMT Closed-Loop Control

Patient Simulator Tests

The Testbench platform tool has been successfully tested and shows its validity in the definition of the performance of a control algorithm. It has been carried out a considerable number of test cases to prove its validity, first in the patient´s model behaviour and then in the control algorithm. For this purpose, the best control algorithm can be identified and tested. There will be ongoing work on the ultimate definition of the algorithm, outside the VALU3S project, but the objectives of VALU3S in the development of this tool have been fully achieved.

Patient Simulations

To verify the correct behaviour of the simulator, the simulator has been run with different patient configurations. In these tests, an initial dose of relaxant has been applied to the patient and has been allowed to evolve (without applying additional doses) until complete recovery.

The tests have been carried out in series where a single parameter of the simulation is varied, to verify that the effects of said parameter are consistent with clinical experience.

Volume of Distribution (Vd): When the drug is injected into the patient, it is dispersed in a volume known as the "volume of distribution" and is expressed in ml/kg. As presented in Figure 3-34 and Figure 3-35, we see how both the concentration of the drug in plasma (CP) and the neuromuscular transmission (NMT) evolve depending on the different Vd of the patient.

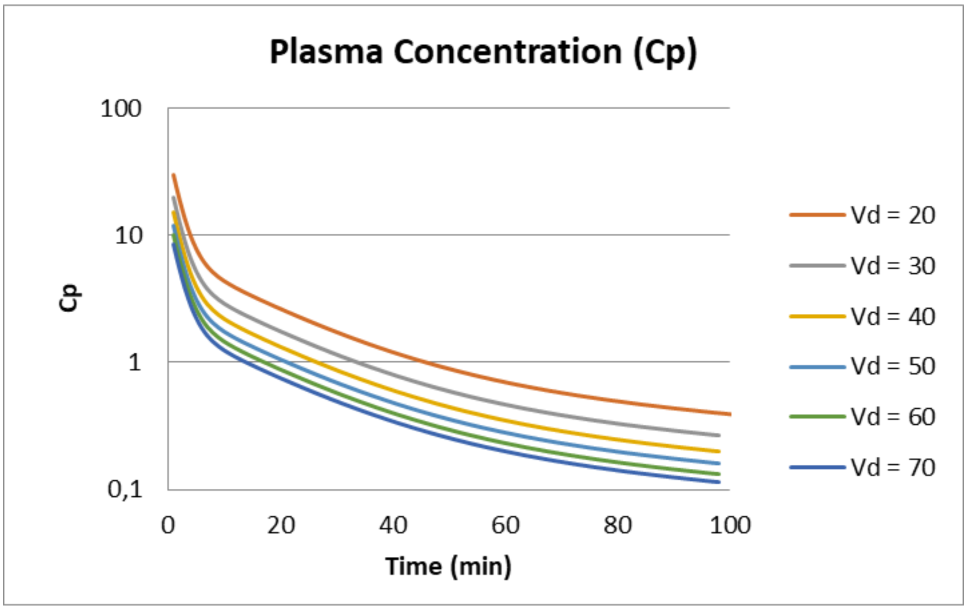


Figure 3-34 Plasma Concentration

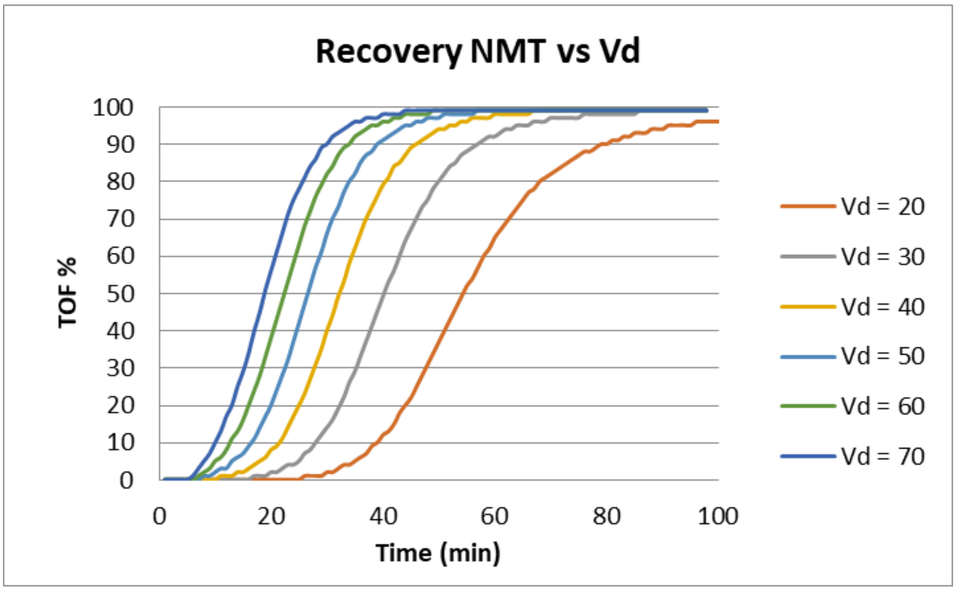


Figure 3-35 Recovery of NMT vs Vd

We can see how as Vd increases the concentration of the drug in the plasma decreases. Logically, its concentration is lower as the volume where the drug is distributed is greater. The behaviour of the NMT recovery is also logical. As Vd increases (and therefore Cp decreases) the time over which the drug takes effect decreases.

Plasma Concentration Cp50: Cp50 is the plasma concentration which would, at a steady state, produce 50% depression of NMT response. This means that the lower the Cp50, the more sensitive the patient is to the drug. As presented in Figure 3-36, we see how neuromuscular transmission (NMT) evolves depending on the different Cp50 of the patient.

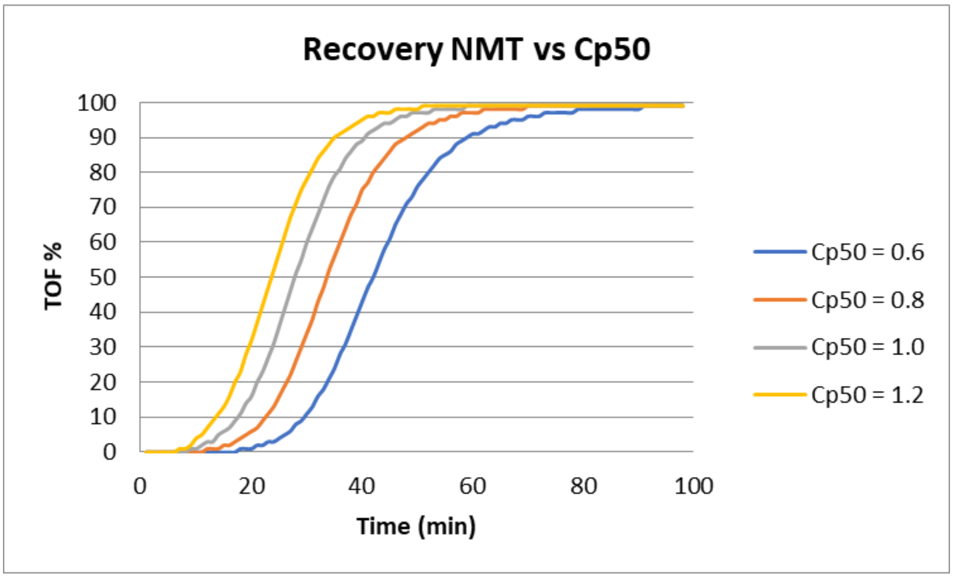


Figure 3-36 Recovery of NMT vs Cp50

In the graph, it can be seen that the recovery time increases with a lower Cp50. This is consistent with the actual behaviour of patients, since the patient who is more sensitive to the drug (<Cp50) needs less drug plasma concentration to have the same effect on NMT and therefore takes longer to recover the NMT.

Initial Dose: As presented in the following graphs (Figure 3-37 and Figure 3-38) we see how both the plasma drug concentration (CP) and the neuromuscular transmission (NMT) evolve as a function of different initial drug doses.

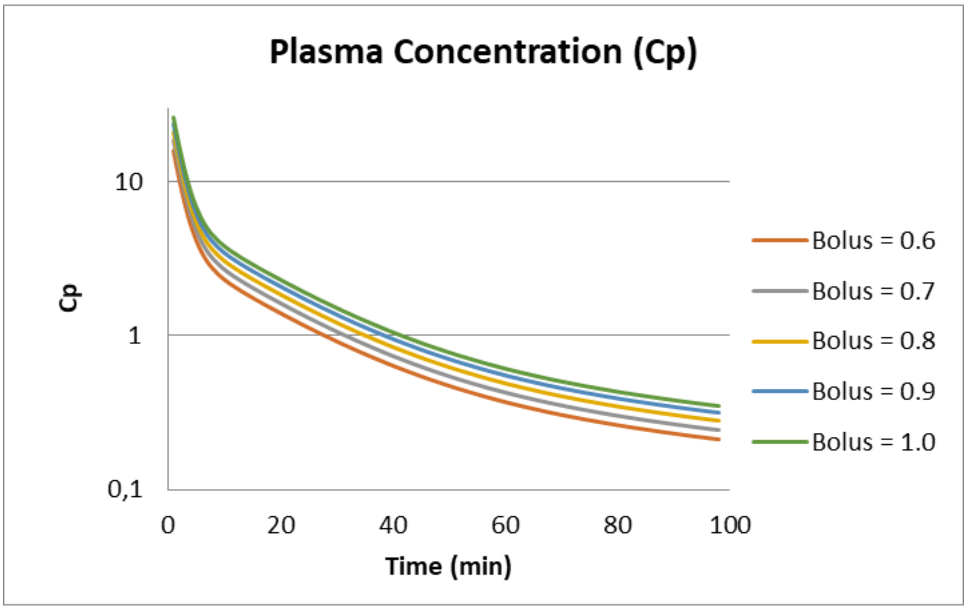


Figure 3-37 Plasma Concentration

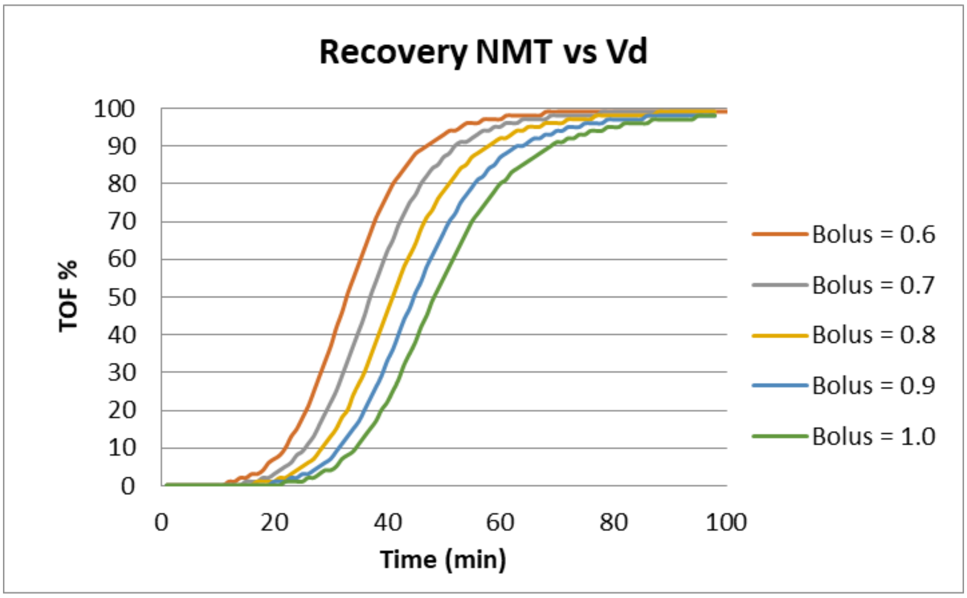
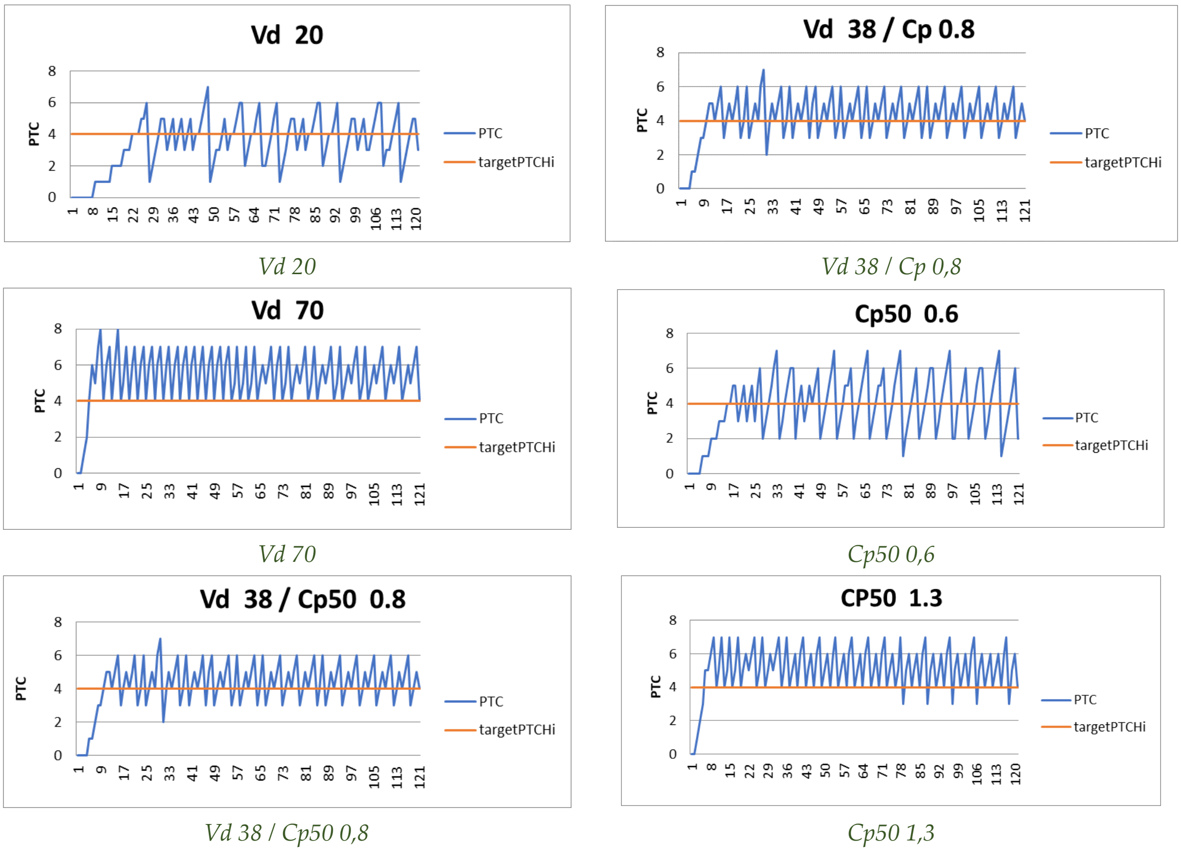


Figure 3-38 Recovery NMT vs Vd

The higher the dose, the higher the concentration in plasma and therefore the longer the drug remains in the patient, prolonging his recovery.

NMT Control Algorithm 1: Once the correct operation of the patient simulator has been verified, we apply it to evaluate the NMT control strategy.



Vd 20

Vd 38 / Cp 0,8

Vd 70

Cp50 0,6

Vd 38 / Cp50 0,8

Cp50 1,3

Figure 3-39 NMT control algorithm 1 outputs

In order to assess the effectiveness of the strategy, several controls are performed by varying the patient's parameters of Vd and Cp50. The graphs in Figure 3-39 show the results obtained.

Table 3-85 Control data with the control algorithm 1

	Target PTC	Media PTC	(sd)	PTC Above (%)	PTC Below (%)	PTC Target (%)
Vd = 20 ml/kg	4	3.94	1.4	31	48	21
Vd = 38 ml/kg (*)	4	4.5	1.05	45	25	30
Vd = 70 ml/kg	4	5.53	1.46	71	3	26
Cp50 = 0.6 µg/mL	4	4.22	1.56	41	42	17
Cp50 = 0.8 µg/mL (*)	4	4.5	1.05	45	25	30
Cp50 = 1.3 µg/mL	4	5.25	1.15	67	8	25

(*) Are the same control.

As presented in Table 3-85, with this control algorithm, the control strategy is best achieved when the patient's configuration matches standard values (Vd = 38 mL/kg, Cp50 = 0.8 µg/mL). In this case, the standard deviation is the lowest (1.05) and the time spent on target is also the highest (30%). When the patient's configuration is not very sensitive to the drug (high Vd or high Cp50), then the dose applied by this strategy is insufficient and the NMT value remains above the target for a long time (71% and 67%). When the patient's configuration is very sensitive to the drug (low Cp50) then the dose applied by this strategy causes large variations in the NMT value (its standard deviation is very large 1.56).

Model-based Assurance and Certification

Compliance of the NMT controller has been analysed with the SIAM tool, considering part 5 (software development process) of the IEC 62304 standard. This part of the standard includes safety and security considerations.

SCP2 - Number of Safety/Security Requirement Violations: Quality practices with automatic compliance gap analysis

- Automatic compliance gap analysis enabled 108 quality practices

VV5 - Joint management of SCP requirements: Quality practices in joint automatic safety & security compliance gap analysis

- Joint automatic compliance gap analysis enabled 20 quality practices that refer directly to safety or security

VV10 - Reduced cost and time for work on the certification process and functional safety: Estimated effort and cost reductions thanks to automated compliance gap analysis (average project as a reference)

- At least a 25% effort reduction in compliance gap analysis
- At least a 25% cost reduction in compliance gap resolution thanks to early gap detection (instead of late detection)

Knowledge-Centric Traceability Management

This evaluation has considered the use of Traceability Studio for trace specification, trace discovery, and change impact analysis for the NMT controller, considering risks analysis results, textual system requirements, and Capella design models. The artefacts address both safety and security considerations.

SCP2 - Number of Safety/Security Requirement Violations: Support for trace discovery and change impact analysis, for risks analysis results, textual system requirements, and Capella design models

- Automatic trace discovery of a set of 44 valid safety or security requirements traces (70% precision)
- Tool-supported identification of 330 potential change impacts on risk analysis information affecting several trace chains

SCP7 - Number of prevented accidents: Support for the discovery of safety requirement traces and of traces whose disregard could lead to accidents or whose identification confirms that possible accidents have been addressed

- Automatic trace discovery for 49% of safety requirements
- Automatic impact analysis for a set of 428 risk analysis traces

VV5 - Joint management of SCP requirements: Support for trace discovery and change impact analysis, considering both safety and security requirements, for risks analysis results, textual system requirements, and Capella design models

- Joint automatic trace discovery of a set of 44 valid safety or security requirements traces (70% precision)
- Joint automatic impact analysis for a set of 428 risk analysis traces (including security aspects)
- A joint tool supported the identification of 330 potential change impacts on risk analysis information (including security aspects) affecting several trace chains

VV10 - Reduced cost and time for work on the certification process and functional safety: Estimated effort and cost reductions thanks to automated traceability management (required by standards, e.g., trace specification and change impact analysis; average project as a reference), for risks analysis results, textual system requirements, and Capella design models

- At least 40% effort reduction in change impact analysis
- At least 25% cost reduction in traceability issue resolution thanks to early issue detection (instead of late detection)

Knowledge-Centric System Artefact Quality Analysis

The suitability of system artefacts of the NMT controller in different formats has been assessed with the RQA tool. Risks analysis results, textual system requirements, and Capella design models have been taken into account. The artefacts address both safety and security considerations. The guidance provided by RQA enabled system artefact revision and improvement.

SCP2 - Number of Safety/Security Requirement Violations: Support for the detection and resolution of safety- or security-related requirements issues in risks analysis results, textual system requirements, and Capella design models

- 38 new automatic quality analyses enabled.
- Use of 6 compliance checklists (1 for IEC 62304 [36] and 5 for ISO 14971 [37]).
- Improvement on requirements specification, reducing low quality from 31% (13 system requirements) to 12% (5 system requirements).

SCP7 - Number of prevented accidents: Support for the resolution of issues in safety requirements (which could lead to accidents)

- Improvement on the specification of 41 safety requirements

VV5 - Joint management of SCP requirements: Support to the joint management and resolution of safety- or security-related requirements issues

- Joint improvement of 42 safety or security requirements
- Use of 6 compliance checklists (1 for IEC 62304 and 5 for IEC 14971) for safety- or security-related requirements aspects

VV10 - Reduced cost and time for work on the certification process and functional safety: Estimated effort and cost reductions thanks to automated system artefact analysis (required by standards, e.g., as a verification activity; average project as a reference) for risks analysis results, textual system requirements, and Capella design models

- At least 20% effort reduction in system artefact analysis
- At least 25% cost reduction in system artefact issue resolution thanks to early issue detection (instead of late detection)

The joint application of Knowledge-Centric Traceability Management and Knowledge-Centric System Artefact Quality Analysis corresponds to the overall demonstrator named Early V&V in Knowledge-Centric Systems Engineering. Figure 3-40 provides an overview of the main improvements enacted, also considering the baseline situation before VALU3S started. The percentage of low-quality requirements has been reduced by following the base recommendations that RQA provides, automatic trace discovery with the default mechanisms of Traceability Studio has resulted in trace discovery for around half the requirements, the joint improvement of safety and cybersecurity requirements has been enabled, and effort for change impact analysis has been reduced by 40% (estimation). All in all, as summarised in Figure 3-40, the outcome of the demonstrator leads to wider system artefact quality analysis, more precise traceability management, better system artefacts, lower effort in the addressed V&V tasks thanks to automated support, and lower cost in issue resolution thanks to early issue detection.

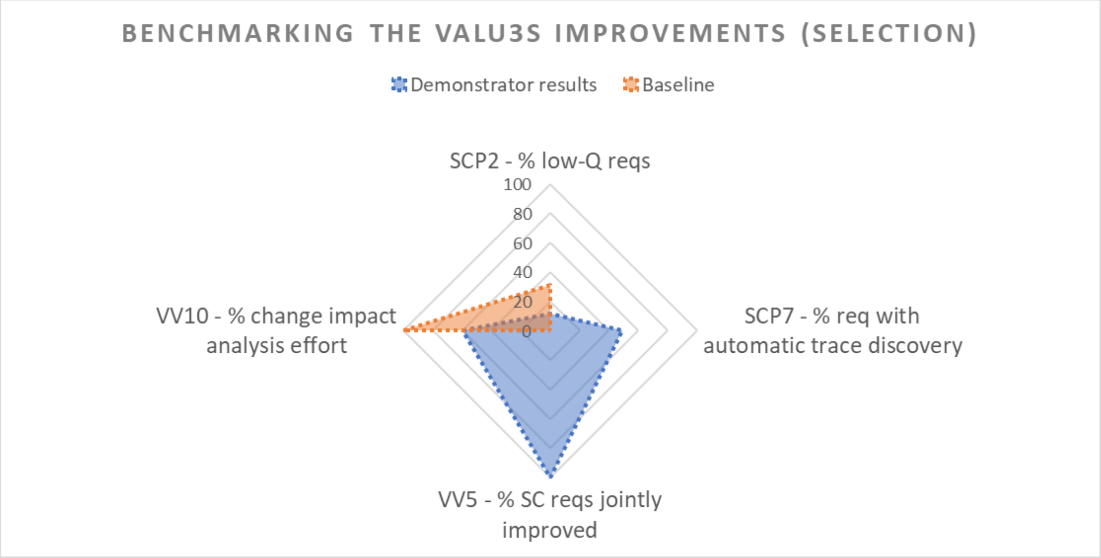


Figure 3-40 Overview of improvement thanks to Early V&V in Knowledge-Centric Systems Engineering

MSA-FLA with CHESS-FLA

The Model-based Safety Analysis with Failure Logical Analysis (MSA-FLA) performed with the CHESS-FLA tool allowed obtain the following quantitative results:

- 9 potential hazard situations deriving from the erroneous behaviour of the Controller have been identified.
- 72 sequences or combinations of events that may cause a hazardous situation have been identified.

- The time needed to analyse the performance of different potential strategies has been reduced by a factor of 0.6.
- 6 different characteristics that could affect the safety of the Controller have been analyzed.

3.8.6 Qualitative Results

In the medical sector, one of the key issues is to cope with access limitations in site evaluation. Clinical trials represent a complex and regulated process that starts with the need to obtain technical compliance with applicable safety and security standards. Laboratory tests regarding 60601 standards, for example, must be conducted, as well as full documentation on risk analysis and ergonomic aspects, in conformance to the everyday more complex regulatory process.

Even when all these aspects have been fulfilled, clinical trials are costly and time-consuming. Sometimes, the patient´s profile does not match the required type, etc. Given this, the tools that have been developed in this UC8 have proved to be most valuable. On one side, the system can be tested at a preliminary stage, thus anticipating problems at an earlier stage. Also, the system can be tested by taking into consideration different parameters for the patient model, so the performance of the controller can be tested in very different conditions. The verification tools have identified performance errors from the beginning and have proved to be a valuable means to detect the control algorithm that performs best. In the development process, before experimental trials, the monitor has been tested with a patient´s model and this has allowed additional reassurance about the correct performance since errors could be detected at the laboratory level.

Demonstrator 1: MSA-FLA with CHESS-FLA

Participants Profile: QAM is applied to 10 subjects (7 males, 3 females) aged in the range of 23-34. The education level is relatively high as the subject pool is composed of 2 PhD researchers and 8 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as “R&D engineers, project managers, Q&A etc.” having experience in the fields of “automotive, software/hardware engineering, embedded systems etc.”.

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-86. The results show that the majority of the constructs are correlated with each other except PE-CO, PT-MO, ROI-PT and BI-{PU, PEOU} pairs.

Table 3-86. UC8 – Demonstrator 1 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.504	1									
MO	0.283	0.606	1								
CO	0.235	0.268	0.493	1							
ROI	0.127	0.202	0.542	0.069	1						
PE	0.676	0.253	0.128	-0.172	0.213	1					
PT	0.423	0.278	-0.097	0.278	-0.162	0.429	1				
PR	0.526	0.620	0.884	0.560	0.490	0.263	0.234	1			

SI	0.540	0.443	0.707	0.237	0.604	0.681	0.261	0.683	1		
ATU	0.445	0.441	0.472	0.472	0.167	0.435	0.230	0.363	0.580	1	
BI	-0.056	-0.092	0.207	0.415	0.319	0.034	0.552	0.310	0.380	0.281	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-87, the questions asked to subjects are sufficiently reliable as understood from subject responses except for BI.

Table 3-87. UC8 – Demonstrator 1 Reliability Analysis

Cronbach-alpha values	
PU	0.221
PEOU	0.623
MO	0.111
CO	0.444
ROI	0.230
PE	0.141
PT	0.013
PR	0.408
SI	0.123
ATU	0.650
BI	-0.969

Regression Analysis: Regression analysis is applied to estimating the relationships among QAM constructs, as seen in Table 3-88. For this demonstrator, there exists a right proportional relation between PU-{MO, CO, PEOU}, PU-ATU and PEOU-ATU influencing each other in the same direction whereas BI and {ATU, ROI, PE, SI, PT, PR} are inversely influenced.

Table 3-88. UC8 – Demonstrator 1 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.17xPU + 4.6	0.08	0.428	0.835
H2	CO-PU	Right	CO= 0.147xPU + 4.78	0.055	0.513	0.684
H3	PEoU-PU	Right	PeoU= 0.42xPU + 3.36	0.254	0.137	1.65
H4	PU-ATU	Right	PU= 0.56xATU + 1.9	0.198	0.198	1.404
H5	PeoU-ATU	Right	PeoU= 0.468xATU + 2.65	0.194	0.202	1.389
H6	ATU-BI	Inverse	ATU= -0.18xBI + 4.32	0.079	0.432	0.827
H7	ROI-BI	Inverse	ROI= -0.15xBI + 4.51	0.102	0.368	0.954
H8	PE-BI	Inverse	PE= -0.035xBI + 5.03	0.001	0.926	0.096
H9	SI-BI	Inverse	SI= -0.3xBI + 3.69	0.145	0.278	1.163
H10	PT-BI	Inverse	PT= -0.36xBI + 3.33	0.304	0.098	1.87
H11	PR-BI	Inverse	PR= -0.108xBI + 4.72	0.304	0.384	0.922

Demonstrator 2: NMT Controller TestBench Platform

Participants Profile: QAM is applied to 10 subjects (9 males, 1 female) aged in the range of 24-60. The education level is relatively high as the subject pool is composed of 2 Post-Doc or higher-degree and 1 PhD researcher and 7 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "CTOs, technical managers, Q&A, sales managers, R&D engineers, medical instrumentation engineers, etc." having experience in the fields of "academia, health, medical instruments, embedded systems etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-89. The results show that all constructs are correlated with each other.

Table 3-89. UC8 – Demonstrator 2 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.874	1									
MO	0.791	0.959	1								
CO	0.817	0.984	0.960	1							
ROI	0.668	0.914	0.940	0.963	1						
PE	0.669	0.908	0.957	0.950	0.989	1					
PT	0.623	0.880	0.920	0.935	0.993	0.992	1				
PR	0.661	0.896	0.939	0.954	0.990	0.988	0.985	1			
SI	0.575	0.858	0.894	0.927	0.991	0.976	0.991	0.986	1		
ATU	0.778	0.970	0.980	0.990	0.977	0.973	0.957	0.968	0.942	1	
BI	0.620	0.869	0.932	0.916	0.979	0.992	0.991	0.977	0.975	0.944	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-90, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-90. UC8 – Demonstrator 2 Reliability Analysis

Cronbach-alpha values	
PU	0.312
PEOU	0.121
MO	0.831
CO	0.155
ROI	0.700
PE	0.285
PT	0.031
PR	0.352
SI	0.537
ATU	0.398
BI	0.231

Regression Analysis: Regression analysis is applied to estimating the relationships among QAM constructs as presented in Table 3-91. For this demonstrator, there exists a right proportional relation between all QAM constructs influencing each other in the same direction.

Table 3-91. UC8 – Demonstrator 2 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.26xPU + 4.7	0.625	0.006	3.654
H2	CO-PU	Right	CO = -.283xPU + 4.73	0.668	0.004	4.012
H3	PEoU-PU	Right	PEoU = 0.35xPU + 4.21	0.763	0.001	5.078
H4	PU-ATU	Right	PU = 2.23xATU - 8.47	0.605	0.008	3.502
H5	PEoU-ATU	Right	PEoU = 1.13xATU - 0.9	0.942	~0	11.358
H6	ATU-BI	Right	ATU = 0.93xBI + 0.76	0.891	~0	8.072
H7	ROI-BI	Right	ROI = 1xBI + 0.28	0.959	~0	13.7
H8	PE-BI	Right	PE = 0.99xBI + 0.1	0.983	~0	21.731
H9	SI-BI	Right	SI = 1xBI + 0.22	0.95	~0	12.351
H10	PT-BI	Right	PT = 1xBI + 0.12	0.981	~0	20.487
H11	PR-BI	Right	PR = 1.012xBI + 0.27	0.981	~0	12.873

Demonstrator 3: Early V&V in Knowledge-Centric Systems Engineering

Participants Profile: QAM is applied to 15 subjects (13 males, 2 females) aged in the range of 24-44. The education level is relatively high as the subject pool is composed of 3 Post-Doc or higher-degree and 5 PhD researchers and 7 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as “professors, researchers, software/hardware engineers, directors, managers, etc.” having experience in the fields of “CPS, V&V, Q&A, big data, computer vision, AI/ML etc.”.

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-92. The results show that all constructs are correlated with each other.

Table 3-92. UC8 – Demonstrator 3 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.573	1									
MO	0.557	0.779	1								
CO	0.572	0.815	0.892	1							
ROI	0.537	0.858	0.793	0.929	1						
PE	0.322	0.723	0.862	0.782	0.786	1					
PT	0.581	0.844	0.768	0.891	0.921	0.761	1				
PR	0.327	0.750	0.632	0.807	0.878	0.732	0.769	1			
SI	0.470	0.252	0.351	0.363	0.340	0.186	0.344	0.064	1		
ATU	0.390	0.858	0.897	0.879	0.903	0.934	0.818	0.805	0.277	1	
BI	0.568	0.895	0.792	0.864	0.950	0.841	0.931	0.822	0.294	0.909	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-93, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-93. UC8 – Demonstrator 3 Reliability Analysis

Cronbach-alpha values	
PU	0
PEOU	0.2
MO	0.219
CO	0.252
ROI	0.598
PE	0.305
PT	0.684
PR	0.232
SI	0.660
ATU	0.186
BI	0.744

Regression Analysis: Regression analysis is applied to estimating the relationships among QAM constructs as presented in Table 3-94. For this demonstrator, there exists a right proportional relation between all QAM constructs influencing each other in the same direction.

Table 3-94. UC8 – Demonstrator 3 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.31xPU + 4.23	0.31	0.031	2.417
H2	CO-PU	Right	CO = 0.28xPU + 4.46	0.327	0.026	2.512
H3	PEoU-PU	Right	PEoU = 0.25xPU + 4.69	0.329	0.025	2.523
H4	PU-ATU	Right	PU = 1.16xATU - 1.76	0.152	0.151	1.526
H5	PEoU-ATU	Right	PEoU = 1.128xATU - 0.46	0.736	~0	6.023
H6	ATU-BI	Right	ATU = 0.6xBI + 2.64	0.826	~0	7.863
H7	ROI-BI	Right	ROI = 0.77xBI + 1.72	0.902	~0	10.917
H8	PE-BI	Right	PE = 0.575xBI + 2.38	0.707	~0	5.606
H9	SI-BI	Right	SI = 0.26xBI + 5.63	0.086	0.288	1.107
H10	PT-BI	Right	PT = 1.01xBI + 0.17	0.866	~0	9.162
H11	PR-BI	Right	PR = 1.006xBI + 0.17	0.866	~0	9.162

3.8.7 Observed Limitations, Lessons Learnt and Best Practices

As far as the testbench platform is concerned, VALU3S results have proved to be a valuable tool to evaluate the control algorithm performance and to select the best approach among several alternatives. Once the algorithm is defined, it can be easily incorporated into the NMT controller monitor, and initiate the experimental and clinical trials. It is also an excellent way to conduct test cases for certification purposes.

Concerning Model-based Assurance and Certification, Knowledge-Centric Traceability Management, and Knowledge-Centric System Artefact Quality Analysis, the application of the corresponding methods and tools requires an initial effort for configuration, such as the creation of models of standards, the development of ontologies, the selection of quality aspects to analyse, or the selection of specific tool behaviours among different alternatives (e.g., for trace discovery). This has been considered when estimating effort and cost reductions in the use case. It is also important to note that RGB is an SME whose staff are highly specialised in and knowledgeable about the systems that they develop and how to address system artefact quality, traceability, and compliance for them. Their experience helps them perform the underlying activities effectively and efficiently. The estimated gains will arguably be greater in companies developing a wider range of systems or larger systems, or where the staff are less experienced.

Concerning the Model-based Safety Analysis with Failure Logical Analysis method, supported by the CHESS-FLA tool, INTECS, together with RGB, agree that the greatest effort on the application of the method falls back on the design of the functional model of the system and on the enrichment of this functional model with the decorations required to apply the Failure Logical Analysis. Once the extended functional model of the system has been designed, there is a huge reduction in the time and effort needed to compute the Failure Mode and Effect Analysis (FMEA) table and the Fault Trees, which are commonly used artefacts in the context of the safety analysis. CHESS-FLA automatically computes these artefacts in a couple of seconds, compared to a couple of weeks usually need to manually compute them. Furthermore, the automatic computation allows also to reduce the error proneness and the possibility to miss some relevant error propagation paths.

The qualitative assessment results of UC8 demonstrations 1, 2, and 3 are given in Table 3-95, Table 3-96 and Table 3-97, respectively. The user responses show that the acceptance of the technology is high as the responses to the majority of the QAM factors are higher than 5.00 out of 7.00. These numbers show that the proposed techniques are very promising, especially in the healthcare domain, where the subjects are always very sceptical. It seems that UC8 providers present a very strong quantification of results, in terms of graphics and well-prepared videos, as this increased the awareness and consciousness of respondents.

Table 3-95 Mean and standard deviation of experts' responses to UC8 - Demonstration 1

UC8 Demo1	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,48	5,00	5,18	4,77	4,90	5,24	5,30	4,58	5,15	4,99	5,22
Std Dev	0,44	0,52	0,73	0,70	0,79	0,34	0,55	1,02	0,46	0,56	0,36

Table 3-96 Mean and standard deviation of experts' responses to UC8 - Demonstration 2

UC8 Demo2	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	6,14	5,44	5,44	5,00	5,32	5,58	5,50	5,30	5,40	5,25	5,63
Std Dev	0,73	1,79	2,18	2,10	2,00	2,06	2,03	1,98	2,00	2,09	2,05

Table 3-97 Mean and standard deviation of experts' responses to UC8 - Demonstration 3

UC8 Demo3	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,95	4,98	5,58	5,34	5,17	5,81	5,52	5,22	5,33	5,16	5,72
Std Dev	0,62	1,40	1,12	1,26	1,49	1,77	1,12	1,23	0,00	1,84	1,21

3.9 Use Case 9 – Autonomous Train Operation (UC9)

The CV&AI-enhanced algorithms for (driverless) autonomous train operation will need a further substantial effort to increase the TRL before bringing it to the market. CV&AI-enhanced technology must fulfil strict standards and safety regulation in order to be certified. In addition, regarding the certification process of railway systems and according to EN-5012x standards[3] (EN 50126 (IEC 62278) – Reliability, Availability, Maintainability, and Safety (RAMS); EN 50128 (IEC 62279) – Software (Signaling Systems); EN 50129 (IEC 62425) – System safety (Subsystem Software)), CV&AI-enhanced techniques are not currently recommended, so the adoption of this kind of solutions in such a domain is still a challenge. For this reason, virtual environment for V&V will reduce costs of AI-enhanced algorithms and it will support an easier marketing process avoiding important first barriers.

UC9 focuses on the CV&AI-based railway signal detector/identifier techniques as depicted in Figure 3-41. After several data recorded in the field (real railway journeys), CAF Signalling trains different CV&AI based object detectors/identifiers. Light signals (green, red, orange), static speed restrictions panels, platform stopping point signals, and platform proximity signals have been labelled in different video databases in order to train these custom models. Although, the resulting models show accurate performances in nominal scenarios, they must be tested in higher variety of situations, extreme conditions and hazard situations in order to consider them really validated and verified.

[3] https://www.tuvsud.com/en-us/services/functional-safety/en-5012x-railway

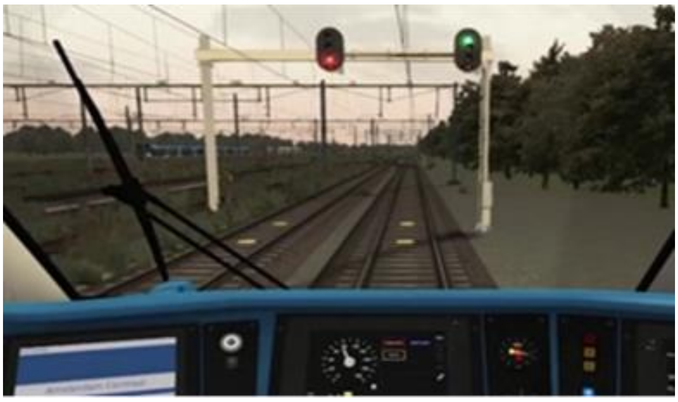


Figure 3-41 Image of the simulation tool.

3.9.1 V&V challenges

European standardization group of Shift2Rail IP2 - X2RAIL-4 [46] in which CAF Signalling is involved, is currently working on a future GoA4 (driverless) Autonomous Train Operation (ATO) system definition. Because of this ATO system, CAF Signalling is facing up different Verification and Validation (V&V) challenges for the CV&AI-enhanced autonomous train operations that are based on non-deterministic algorithms. It is not easy to collect a real database containing different realistic scenarios to validate computer vision-based AI techniques. There is a need to use simulation scenarios to ensure reliability and fasten the system validation.

Specific challenges targeted in UC9 are mainly based on the following intended or unintended cases or the problems related to the IT infrastructure:

- Safety violations on signal (light) detection where the CV/AI algorithms may fail, especially in harsh situations where the lighting conditions are not appropriate.
- Safety violations on speed restriction signs detection are also encountered as vital challenges as the camera-based systems may tend to serious detection failures resulting from CV/AI algorithm design or the problems related to the datasets used for training these algorithms.
- Computer vision system operating limitations are other challenges as signalling, monitoring and surveillance systems are very complex, heterogeneous, sparse and geographically spread. This may cause delays in operations, inefficiencies in actuating systems at edges and throughput problems in real-time.

3.9.2 Contributors

Partners contributing to the UC9: CAF, IKER

3.9.3 Contributors' Roles & Evaluation Scenario

The demonstrator focuses on the validation process for CAF's new railway sign and signal detection system, called Polaris. The aim is to carry out a preliminary validation of the system in a laboratory environment since the realisation in a real environment is not feasible due to its complexity and high cost. For the implementation of this demonstrator, IKER has prepared a set of validation tests using

Data Generation for Validation (DaGe4V) and the Train Simulator tools. The execution of the validation tests is carried out by using the Polaris system provided by CAF. Results obtained during the executions of the tests are collected and analysed by using Validation Test Result Analysis (VaTRA) tool, to detect safety-related issues and identify limitations of the Polaris system.

The requirements of the tools were decided between CAF and IKER in the design phase. The defined requirements of the tools were carried out by IKER, i.e., DaGe4V for data generation and validation and VaTRA for analysis of system validation test results. The preparation of the AI algorithms and models was carried out by CAF. The global analysis of the tools has been done by both, CAF and IKER.

List of evaluation scenarios defined for this use case:

- VALU3S_WP1_Railway_4 – Safety violations on signal (light) detection.
- VALU3S_WP1_Railway_5 – Safety violations on speed restriction signs detection.
- VALU3S_WP1_Railway_6 – Computer vision system's operating limitations identification.

The assignments of the contribution of all UC partners can be found in Table 3-98**:**

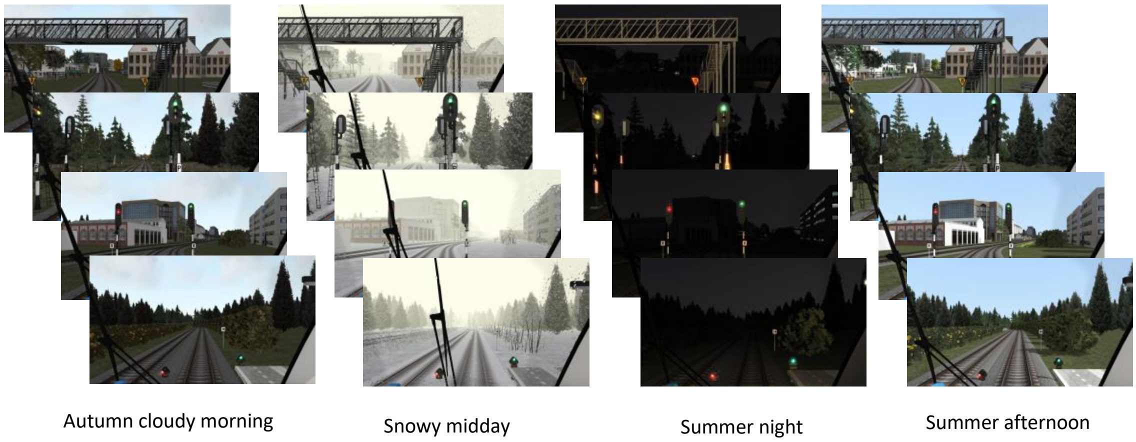
Table 3-98 Overview of contribution to evaluation scenarios by UC9 partners

Evaluation Scenario	CAF	IKER
VALU3S_WP1_Railway_4	X	X
VALU3S_WP1_Railway_5	X	X
VALU3S_WP1_Railway_6	X	X

Recently, the process of capturing railway frames is expensive in terms of time and resources, as many parties take part in operations. The track and the training need to be allocated, with the bureaucratic obstacles that this entails. A driver to drive the train and a person to take frames are needed to record the runs. Therefore, just for the obtention of the images, a) train and track allocation time and cost, b) the time and cost of the driver, and c) the time and cost of the recorder need to be considered. In today's operations, only two driving sessions are performed each month, while each session contains six runs. These runs do not differ much in meteorological or lighting conditions, which makes each of these run's images less valuable for the validation of the model, as they are very similar.

Once the frames are obtained, manual labelling must be performed. This is the most time-consuming task as each of the objects of each frame must be labelled correctly. In this second phase, the d) labelling time is measured. Once the ground truth is created, inference needs to be performed and the output of the model must be compared with the ground truth. In this last phase the e) inference time is measured.

All these factors need to be considered to estimate the time and cost of the process, so a time and cost estimation of obtaining and validating a thousand images could be performed after adding the respective times and costs of each step. The approximation of tasks a, b, and c will be estimated in the future, but the mean value per each bunch of thousand images, more or less 15 hours are needed to label and validate the labels and model validation (tasks d and e). In addition, even if the quantitative improvement is enormous, the most important change comes in qualitative matters, as the diversity of the created scenarios is impossible to replicate in real life, as the meteorological and lighting conditions cannot be chosen when obtaining the frames (see Figure 3-42).



Autumn cloudy morning

Snowy midday

Summer night

Summer afternoon

Figure 3-42 Example of framesets for Polaris validation

3.9.4 Demonstration:

The demonstration for UC9 was planned to be just one holistic demonstrator. Therefore, their junction of multiple demonstrable items that are covering defined challenges and scenarios, and partially cover VALU3S dimensions (see Table 3-99) are joined.

Table 3-99 Recently updated requirements and test cases of UC9

Item #	Demonstration name	Description/Purpose	Format	Responsible
1	Data Generation and Validation for Railway Domain	A demonstration of the tools generated in VALU3S will be shown. DaGe4V (Data Generation 4 Validation) and VaTRA (Validation Test Result Analysis) will be shown, as to how they work the functionalities they have. It will also be shown how they are connected to the offline simulation and the functionalities of each of them.	Demo/Video: VaTRA – DaGe4V	IKER/CAF

3.9.5 Quantitative Results

List of evaluation criteria of the V&V process (see the details in D5.2 [8])

1. Eval_SCP_2 – Number of Safety/Security Requirement Violations: Polaris CV System detects railway signals and semaphores. Incorrect object identification can cause a safety risk, e.g., not identifying a red semaphore that can cause an accident. The number of Safety/Security requirement violation metrics is used for evaluating the accuracy of the object detection system, as well as to assess whether the signal detection system is having problems with some signals whose misidentification can lead the system to safety problems. (CAF, IKER)

- Within the baseline, testing has been done. It measures safety in different contexts of detection and the sequence of the detections. These metrics (to be verified) are a sample of new metrics that are included in the User Interface (UI) to enhance the times while analysing the model, but no number of enhancements could be measured. Two ways of Safety/Security metrics were carried out. On the one hand, the detection side is regarding how well the detection is performing. On the other hand sequence detection is to know whether the object which is not detected is a major or minor risk.

2. Eval_SCP_4 – Metrics to Evaluate AI/ML Algorithms: This criterion aims to measure the performance of the object detection and identification algorithms. Having these metrics allows evaluation of the improvements obtained each time the algorithms are modified. Polaris CV system is validated using classification metrics, such as accuracy, precision and recall, and Computer Vision metrics, such as Intersection over Union (IoU). (CAF, IKER)
    - Within the baseline, testing has been done. These metrics are a sample of metrics that are included in the UI (not to run different code each time there is a need to analyse the model) to enhance the times while analysing the model, but no number of enhancements could be measured. These measurements give a view of the computer vision system and help in the analysis of the system.

The Polaris system is a new system under development by CAF. Eval_SCP_4 which provides the metrics to evaluate the AI/ML algorithms, such as Eval_SCP_2 which allows numerical analysis of risk situations derived from the incorrect identification of signals and signs, are metrics that allow a numerical evaluation of the system that is being developed.

These metrics would make it possible for a future comparison between different releases of the detection system, but at this moment this information is not available since there is only one interim release of the system.

Evaluation criteria of the V&V process:

1. Eval_VV_3 – Number of Test Cases: This criterion aims to evaluate the number of misidentified signs and the potential safety violations that have been identified thanks to the ability to easily generate many validation test cases, and therefore, to increase the coverage obtained by the validation tests. A comparison between a hand-made non-automated validation process, where validation test inputs are obtained in the field, and the semi-automated process based on synthetic images generated using simulators will be carried out. (CAF, IKER)
    - The number of validation test cases has been increased by a factor of 10, and the great diversity of these test cases is also worth emphasising. Trying to achieve the diversity obtained from the generation of the scenarios would have been a very arduous task in real-life data acquisition, as meteorological conditions can only be obtained with many and the quality of the validation mostly depends on the effort made to test the system in the widest variety possible of scenarios.
2. Eval_VV_8 – Effort Needed for Test: This criterion aims to evaluate the effort required (persons-hour) to specify the test cases in the virtual environment, generate tests datasets, including the recordings and all associated data needed to perform the tests, and finally execute the test cases

and evaluate the behaviour of the system. Obtained measures will be compared to the effort needed currently to carry out the recordings on the field. Considering also that obtaining recordings in all the different light and meteorological conditions is practically impossible. (CAF, IKER)

- When it comes to the effort needed, the simulated scenarios have a great impact on the reduction of the time needed to acquire the frames. Moreover, as stated before, the difference between the measured efforts cannot be exactly calculated but estimated due to the impossibility of recording specific meteorological conditions. Taking this into account, an estimation has been conducted. The person-hours needed to manually obtain a dataset with 10000 frames is estimated to be 50h (taking into account track availability, travel time to the often-international tracks, recording time, data extraction and subsequent selection according to what was recorded), and this is reduced to 2 person-hours with the help of the simulated scenarios. In conclusion, the process has been optimized by a factor of 25.

3.9.6 Qualitative Results

The qualitative assessment is applied at the overall use case level. The subjects' profile and the statistical analysis results are as follows:

Participants Profile: QAM is applied to 11 subjects (11 Males) aged in the range of 24-34. The education level is relatively high as the subject pool is composed of 1 Post-Doc or higher degree and 10 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "R&D engineers, system engineers, software engineers, researchers, etc." having experience in the fields of "AI/ML, visualization, cyber-physical security, embedded systems, autonomous vehicle etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-100. The results show that all constructs are correlated with each other.

Table 3-100. UC9 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.517	1									
MO	0.746	0.665	1								
CO	0.549	0.661	0.895	1							
ROI	0.887	0.687	0.918	0.833	1						
PE	0.833	0.700	0.886	0.746	0.949	1					
PT	0.737	0.694	0.927	0.847	0.934	0.949	1				
PR	0.600	0.370	0.681	0.672	0.706	0.559	0.678	1			
SI	0.828	0.686	0.875	0.772	0.890	0.881	0.900	0.663	1		
ATU	0.747	0.666	0.861	0.835	0.892	0.834	0.882	0.705	0.832	1	
BI	0.865	0.707	0.922	0.770	0.968	0.968	0.939	0.707	0.875	0.860	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-101, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-101. UC9 Reliability Analysis

Cronbach-alpha values	
PU	0.312
PEOU	0.622
MO	0.651
CO	0.681
ROI	0.699
PE	0.731
PT	0.659
PR	0.676
SI	0.466
ATU	0.553
BI	0.735

Regression Analysis: Regression analysis is applied to estimating the relationships among QAM constructs, as seen in Table 3-102. For this demonstrator, there exists a right proportional relation between all construct pairs influencing each other in the same direction.

Table 3-102. UC9 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.24xPU + 5	0.556	0.008	3.357
H2	CO-PU	Right	CO = 0.161xPU + 5.45	0.302	0.08	1.971
H3	PEoU-PU	Right	PEoU = 0.19xPU + 5.17	0.268	0.103	1.814
H4	PU-ATU	Right	PU= 2.77xATU - 12.31	0.559	0.008	3.375
H5	PEoU-ATU	Right	PEoU = 0.904xATU - 0.16	0.443	0.025	2.675
H6	ATU-BI	Right	ATU = 0.81xBI + 1.28	0.739	0.001	5.054
H7	ROI-BI	Right	ROI = 1.11xBI - 0.25	0.938	~0	11.639
H8	PE-BI	Right	PE = 1.039xBI - 0.16	0.937	~0	11.589
H9	SI-BI	Right	SI = 0.68xBI + 2.09	0.766	~0	5.433
H10	PT-BI	Right	PT = 1.02xBI + 0.04	0.881	~0	8.175
H11	PR-BI	Right	PR = 0.708xBI + 2.67	0.881	0.015	2.996

3.9.7 Observed Limitations, Lessons Learnt and Best Practices

As far as the use case test bench was done, there were many improvements that we could take advantage of, but not everything was an improvement, also limitations appeared. Having a simulation-based validation and verification system was a big advantage in terms of flexibility of scenario and environmental changes. There was also created an easy way to evaluate the object detection models and create reports.

But not everything was improved. We have seen that the validation scenarios must be improved and the gap between the simulation and the real environment must be minimised. We have also seen that an online simulation system could help with real-time analysis of the model.

These difficulties showed that there is a need to keep working on simulation V&V methods and that's why the VALU3S project has opened new areas of investigation to fill the gaps that we have seen in the project.

The mean and standard deviation values of the QAM constructs in UC9 (Table 3-103) present a relatively high acceptance of the proposed V&V techniques. This outcome is expected as there is a strong level of consciousness about the dependency on the V&V of signalisation systems and safety monitoring in train operations.

Table 3-103 Mean and standard deviation of experts' responses to UC9

UC9	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	6,26	5,73	5,36	5,02	5,05	5,30	5,22	5,07	5,32	5,02	5,34
Std Dev	0,48	1,31	1,52	1,63	1,46	1,56	1,54	1,67	1,46	1,77	1,67

3.10 Use Case 10 – Safety Function Out-of-Context (UC10)

A railway interlocking system is a hierarchy of sub-systems distributed geographically. As shown in Figure 3-43, a Computer interlocking system in a multi-tier control system manages the wayside objects (e.g., signals, point machine). The orders and status of wayside objects travel through the hierarchy. The focus of this demonstrator is a conceptual safety function that can apply to any level of this hierarchical system. To simplify the study and the evaluation of this demonstrator the functionality of the concept is considered as an object controller.

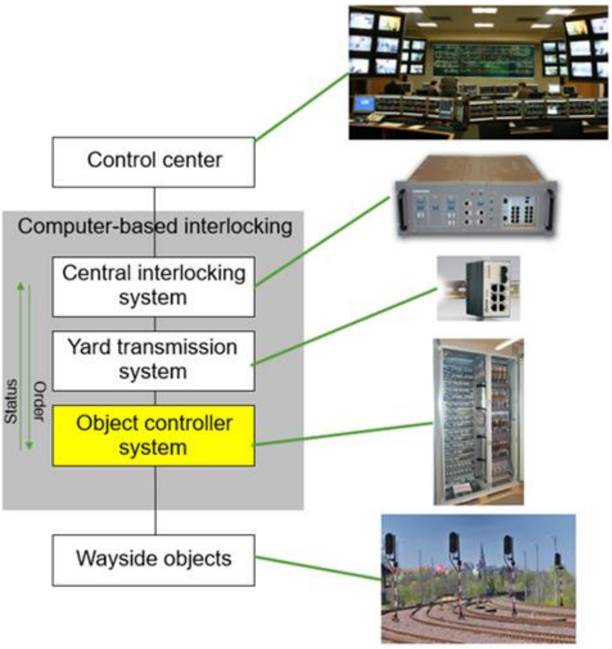


Figure 3-43 Computer-based interlocking system

The object controller is a SIL4 product, and it consists of a *communication interface* where safety orders are received, and a *safety controller* acts on those orders in a timely and safe manner. Additionally, the operation status of the system is reported in conformance with the safety requirements and the maximum Tolerable Hazard Rate (THR) according to SIL4 EN 50126 [21].

3.10.1 V&V Challenges

The demonstrator is a BLDC motor controller with simple functionality to be feasible in the context of the project yet reach enough to show the purpose. The system aims to illustrate new paradigms to achieve the minimum set of functions to conform to railway standards. The conformance is expected to be optimized and achieved using model checking, and mutation testing methodologies. Using such methods indeed has challenges.

The abstract model of the use case in the form of a network of timed automata is built to be verified using the UPPAAL model checker. Even at the abstract level, the model showed that it generates an unfeasible state space to verify safety properties. ISEP came up with the Uppex tool to manage the families of models with different properties and details for verification [23]. Indeed, finetuning and detailing multiple models and arguing the overall soundness of the safety properties of the model from individual families of the model is challenging. The maintenance of consistency between the abstract models and the source code implementation has been done manually. This is facilitated by Uppex, which uses a set of intermediate configuration tables in MS Excel that can be updated by developers to adapt the abstract models and verify the associated properties. AIT with the help of LLSG contributes to this demonstrator with their expertise in *mutation testing*. Using mutation testing, it can automatically generate a set of test cases based on an abstract description of the behaviour. The quality of this set of test cases is mainly measured in terms of the coverage of the system. The verification of these generated test cases is performed either by dedicated testing developers or by software products that generate lower-level tests programmatically.

3.10.2 Contributors

Partners contributing to the UC10: ALSTOM (BT), ISEP, AIT, LLSG

3.10.3 Contributors' Roles & Evaluation Scenario

To evaluate the demonstrator, the following evaluation scenarios have been defined (see also Table 3-104)**:**

1. *VALU3S_WP1_Railway_1 – Inject, detect.* In this use case, we tried to achieve the highest level of safety due to the nature of the distributed control system and higher traffic demands. One way of increasing the confidence of the system to achieve the THR is to use fault injection to evaluate the effect of faults. In this scenario, ALSTOM and AIT with the help of LLSG modelled the control finite state machine (FSM) and introduced faults that should be injected in different parts of the control loop.
2. *VALU3S_WP1_Railway_2 – Normal operation.* In this evaluation scenario, we validate the expected safety functionality of the use case.
3. *VALU3S_WP1_Railway_3 – Systematic and random failures verification.* Safety products, and especially this demonstrator in which we tried to achieve the highest level of safety in terms of qualitative and quantitative approaches, and its mitigation against systematic and random failures. Systematic failures are deterministic and intrinsically occur through the lifecycle of the product and can be described in terms of qualitative approaches. Both EN 50128 [20] and EN 50129 [19] have recommended techniques and approaches in software and hardware to increase the level of safety based on our THR against systematic approaches. On the other hand, random errors can be presented in terms of qualitative approaches. IEC 61508 [22] and EN 50129 present mitigation techniques against random hardware faults.
   ISEP assisted with the design (hardware, software) and verification considering systematic faults using model-checking techniques. AIT, with the help of LLSG, generated test cases using model-based testing techniques to increase confidence in the system against the systematic faults concerning state transition.

Table 3-104 Overview of contribution to evaluation scenarios by UC10 partners

Evaluation Scenario	ALSTOM	AIT&LLSG	ISEP
VALU3S_WP1_Railway_1	X	X	
VALU3S_WP1_Railway_2	X	X	X
VALU3S_WP1_Railway_3	X	X	X

3.10.4 Demonstration

The platform to study and explore the new V&V methods with the collaboration of the interested partners is a SIL4 BLDC motor controller. In railway signalling systems the motor controller (e.g., used in point machines) receives safety function orders via a communication interface from the computer interlocking system (CIS) and acts upon these orders in a safe and timely manner. The motor controller based on a deterministic state machine defines its correct behaviour and failures. After the correct operation, the controller returns a message with the status. In case of a failure detection or potential safety hazard, the controller immediately acts to prevent the hazard and informs the CIS through an alarm message.

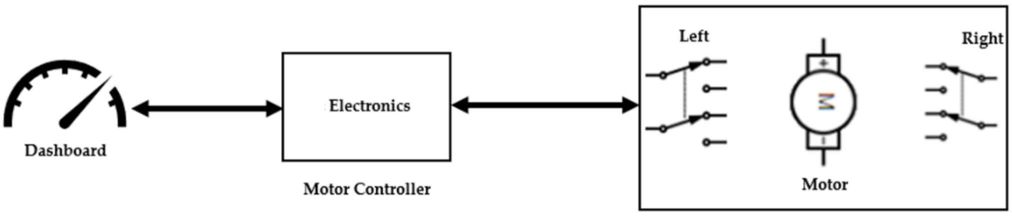


Figure 3-44 DC motor controller overview

Figure 3-44 shows the demonstrator overview. The motor controller performs the simple operation of moving right, left, and stopping and the motor movement is monitored by the limit switches on each side. The motor controller's safety is assured via hardware and software. The dashboard acts as the computer interlocking system to issue commands (e.g., start, stop, left, right) and monitor the operational status of the motor controller. Also, the dashboard logs and displays alarm messages from the motor controller's behaviour.

The safety architecture of the demonstrator is based on the specification of the motor controller as the point machine and is presented in composite principle acting as a 2oo2 (2 out of 2) system. After the hazard analysis, the interested partners in this use case develop their tools in the following steps stated in Table 3-105.

Table 3-105 Use case 10 demonstration

Item #	Demonstration name	Description/Purpose	Type	Responsible
1	Safety verification and validation for the signalling railway application	We investigated the practicality and the requirements of achieving the highest level of safety with a minimal set of state-of-the-art functional safety COTS to be in conformance with EN 50129 using inductive and deductive reasoning and analysis.	Lead Demonstrator Hardware demonstrator	ALSTOM
2	Implementing BLDC motor controller (ALSTOM)	We developed and implemented the use case in the following steps: • Developing GUI software • Developing embedded software • Developing FPGA Firmware • Developing and designing Motor Controller electronics	Lead Demonstrator Hardware demonstrator	ALSTOM

Item #	Demonstration name	Description/Purpose	Type	Responsible
3	Model checking with UPPAAL	Abstract model and requirements in UPPAAL of the DC motor controller based on composite principle were created. These were coupled with the Uppex tool, where we finetuned the model to verify the safety properties based on the preliminary hazard analysis. Also, we started to model a more deterministic version of the motor controller.	Complementary Demonstrator. software demo on request	ISEP/ALSTOM
4	MoMuT - Model based testing	We generated test cases based on a model-based description of the core behaviour of the motor controller, using mutations of this model.	Complementary Demonstrator. Video -Video showing MoMuT application for the use case with test sequence generation.	AIT

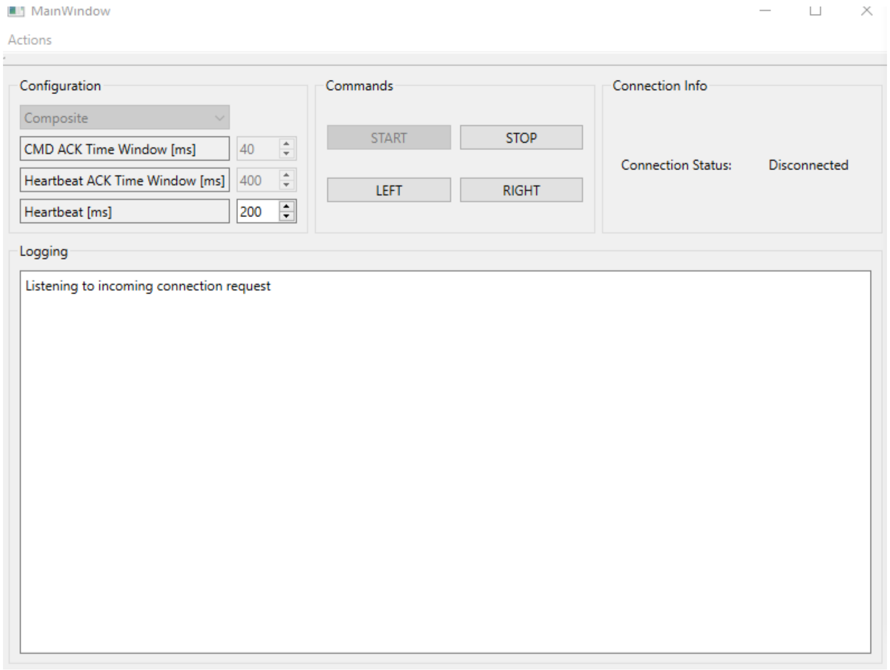
The rest of this section provides further details over some parts of these 4 steps of this demonstrator.

Analysis of a minimal set of state-of-the-art COTS for SIL4 applications (ALSTOM)

In our study, we aimed to investigate the feasibility of attaining the highest level of safety conforming to EN 50129 while using the fewest possible components, by utilizing functional safety Commercial off-the-shelf (COTS). Innovative functional safety components, developed for the automotive industry, have created new opportunities for enhancing safety in other domains such as railways. This Safety Element out of Context (SEooC) components are certified in accordance with ISO 26262 standards and can greatly improve safety applications while reducing the number of required components. By utilizing such components, we were able to enhance the heat signature, reduce the size and cost, and increase confidence in our conceptual safety function.

Implementing BLDC motor controller (ALSTOM)

The GUI, as shown in Figure 3-45, is developed in C++ and the QT framework. The GUI, via three separate UDP sockets, connects to the platform. The platform runs on two different operating systems. Linux handles the UDP sockets via the gateway application. Another application running on Linux is the safety instance that exchanges data with the gateway via IPC. The other operating system runs the same safety instance.



Composite

CMD ACK Time Window [ms] 40

Heartbeat ACK Time Window [ms] 400

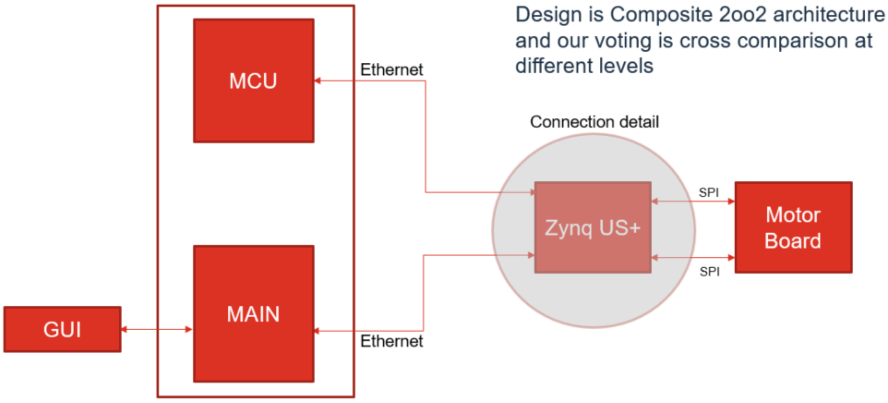
Heartbeat [ms] 200

Connection Status: Disconnected

Listening to incoming connection request

Figure 3-45 UC-10 GUI

As shown in Figure 3-46, the safety controller communicates via separate Ethernets to the Zynq US+ FPGA. The FPGA translates the Ethernet messages to SPI protocol to be able to communicate with the motor board (Figure 3-47). The presence of the FPGA in this use case is only for testing purposes and it doesn’t affect the safety.



Design is Composite 2oo2 architecture and our voting is cross comparison at different levels

Figure 3-46 UC-10 connection overview

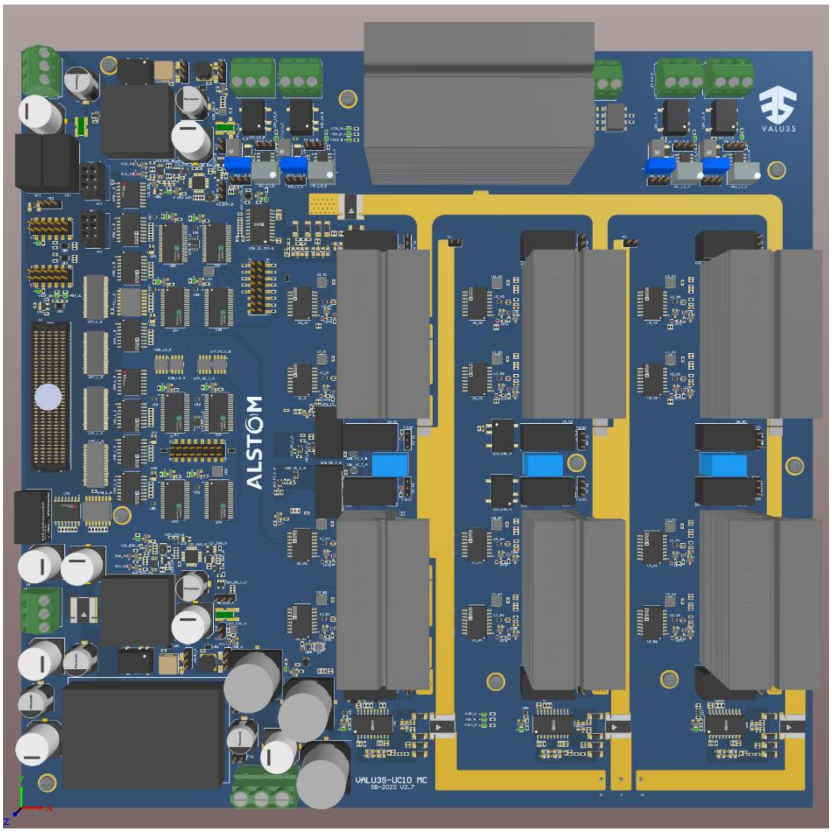


Figure 3-47 UC-10 Motor Board

Model checking with UPPAAL (ISEP, ALSTOM)

We describe below how the verification with UPPAAL model-checking has been performed, and how we overcome some challenges applying ISEP's proposed methods to model-check families of real-time specifications with our new Uppex tool. Further details can be found in a companion scientific publication [23].

ALSTOM and ISEP stated by compiling a set of safety requirements for the controller's software to be verified using model checking. However, when trying to model the expected behaviour using a set of timed automata, including enough information to verify all main requirements, we concluded that it generated a state space too large to be feasible when model-checking. For example, the requirement "the controller component should take less than 100ms to send a given command to the circuit" should not need to consider all combinations of states involving the sending of messages to the dashboard. Similarly, the requirement *"if the controller component receives an error message it should go to a fallback state and the dashboard should be informed within 100ms"* should not need to consider the mechanisms to interact with the circuit.

This led to a family of formal models with different parameters and levels of detail, each targeting different requirements. Our approach tackles 3 challenges: C1: maintain the model, to keep it up to date with the system under development; C2: manage variability, as too many models with commonalities are needed; and C3: improve the collaboration between developers (ALSTOM) and modellers of the formal specifications (ISEP).

Our approach uses a high-level representation of the configurations of the family of formal models for real-time systems. This representation consists of Microsoft Excel spreadsheets with parameters and requirements to be used in the formal models, read by our prototype tool Uppex which automatically generates and verifies the full family of models and requirements. These spreadsheets include, for example, the time bounds of certain components, the size of buffers, and the initial values of certain variables. Furthermore, these values vary according to the set of active features; for example, by activating a feature named SelfTesting, a variable named TSelfTest is set to 200, otherwise, it is set to 0. A special table compiles a set of configurations, each listing its active features. For example, a given configuration could activate SelfTesting, deactivate unrelated monitoring features, and activate its associated requirements.

We start by describing more details over the use-case formal specification as described in [23], which may still evolve to match recent re-designs by ALSTOM. We then describe what information is described in the spreadsheets, and conclude by describing what reports are produced by our Uppex tool.

Formal specification of the controller using Timed Automata

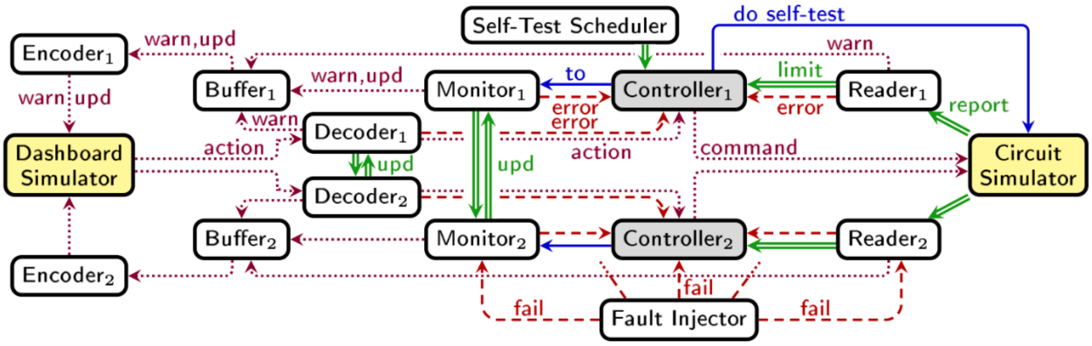


Figure 3-48 Topology of the network of communicating timed-automata of UC10

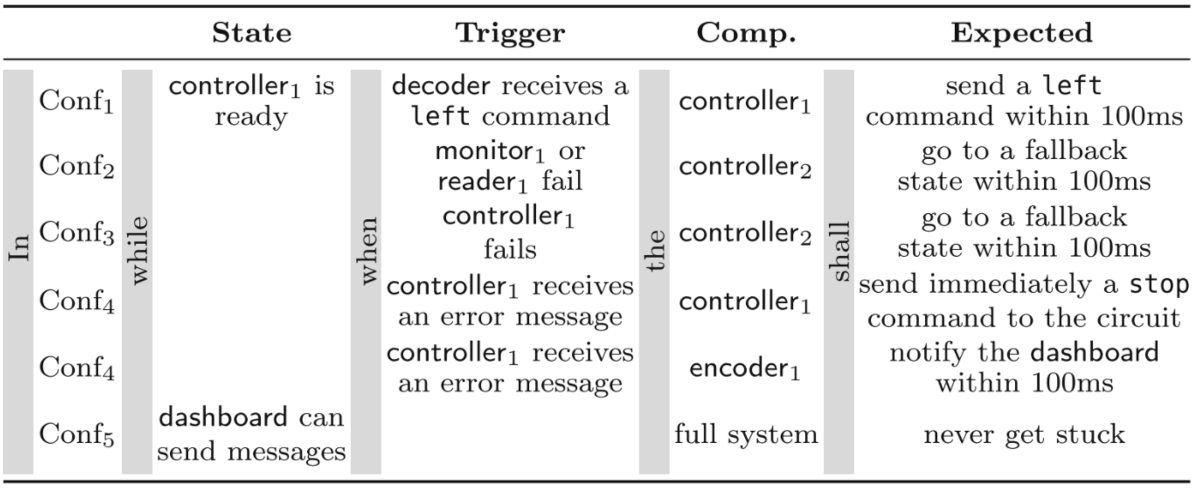
The overall diagram depicting our abstract model of this demonstrator is depicted in Figure 3-48, where each node corresponds to a software task specified by a timed automaton in UPPAAL, and arrows depict interactions. This topology was built iteratively by both developers and formal modellers, during the development of the system, and it is still under development.

The dashboard, circuit, and fault-injector components are parameterised by a scenario, i.e., a sequence of actions with timestamps. The dashboard sends commands to the encoders, the circuit sends reports to the readers describing if there are errors and if the motor reached a limit, and the fault-injector sends messages that cause some components to go to a faulty state with no behaviour. Furthermore, the circuit reports errors for a predefined time window during the self-test phase, and the controllers validate that an error is indeed reported.

Our UPPAAL models were further annotated with special blocks, e.g., starting with *"// @Name"*, which act as hooks that Uppex uses to inject and update the values that configure the model. XML blocks from *"<Name>"* until *"</Name>"* also act as hooks for annotations, which we use to inject and update the properties being verified in the <*queries*> block. We call these @-annotations and XML-annotations, respectively.

Requirements of UC10 in spreadsheet tables

We compiled around 25 requirements, similar to the ones depicted in Figure 3-49, based on a previous hazard analysis performed by ALSTOM. Each of these requirements was manually converted to a logical expression for the Uppaal model checker and included in this table (written as an Excel spreadsheet).



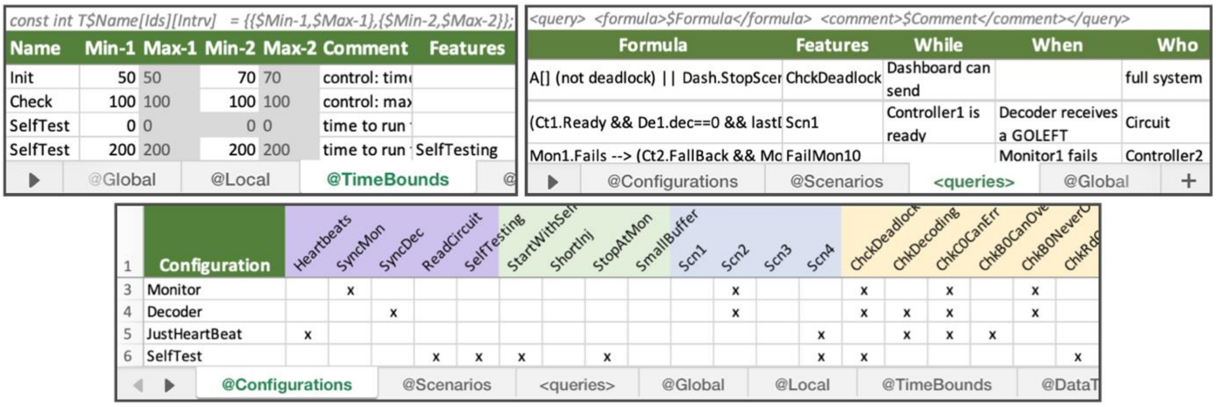
			State		Trigger		Comp.		Expected
In	Conf_1	while	controller$_1$ is ready	when	decoder receives a left command	the	controller$_1$	shall	send a left command within 100ms
	Conf_2				monitor$_1$ or reader$_1$ fail		controller$_2$		go to a fallback state within 100ms
	Conf_3				controller$_1$ fails		controller$_2$		go to a fallback state within 100ms
	Conf_4				controller$_1$ receives an error message		controller$_1$		send immediately a stop command to the circuit
	Conf_4				controller$_1$ receives an error message		encoder$_1$		notify the dashboard within 100ms
	Conf_5		dashboard can send messages				full system		never get stuck

Figure 3-49 Some functional and non-functional requirements for UC10

For example, the 3rd requirement states: *"In Conf3, when controller1 fails the controller2 shall go to a fallback state within 100ms."* Configurations specify the parameters of the model when validating the requirement. This covers both general parameters of the system, such as the time to decode messages and the frequency of operation of monitors, and the scenario consisting of the messages sent by the dashboard, by the circuit, and by the fault-injector. In our example, Conf3 defines a scenario where the dashboard sends a start and a left command after 20ms and 100ms, respectively, and the fault-injector causes controller1 to fail after 120ms.

Parameters and configurations of UC10 in spreadsheet tables

The possible values of the parameters are included in Excel spreadsheets, such as the ones depicted in Figure 3-50. These can be of 3 types, based on their name. The list of configurations is specified in an Excel sheet named @Configurations, such as the one at the bottom of Figure 3-50. Values that should be inserted in the UPPAAL @-annotations are in Excel sheets named @Name, e.g., @Timebounds in the top-left of Figure 3-50. Finally, XML-annotations are included in Excel sheets named <Name>, e.g., <queries> in the top-right of Figure 3-50 containing the requirements.



const int T$Name[Ids][Intrv]   = {{$Min-1,$Max-1},{$Min-2,$Max-2}};

Name	Min-1	Max-1	Min-2	Max-2	Comment	Features
Init	50	50	70	70	control: tim	
Check	100	100	100	100	control: ma	
SelfTest	0	0	0	0	time to run	
SelfTest	200	200	200	200	time to run	SelfTesting

▶ @Global @Local @TimeBounds @

<query> <formula>$Formula</formula> <comment>$Comment</comment></query>

Formula	Features	While	When	Who
A[] (not deadlock) || Dash.StopScer	ChckDeadlock	Dashboard can send		full system
(Ct1.Ready && De1.dec==0 && lastI	Scn1	Controller1 is ready	Decoder receives a GOLEFT	Circuit
Mon1.Fails --> (Ct2.FallBack && Mo	FailMon10		Monitor1 fails	Controller2

▶ @Configurations @Scenarios <queries> @Global +

1	Configuration	Heartbeats	SyncMon	SyncDec	ReadCircuit	SelfTesting	StartWithSel	ShortInj	StopAtMon	SmallBuffer	Scn1	Scn2	Scn3	Scn4	ChckDeadlock	ChkDecoding	ChkC0CanErr	ChkB0CanOve	ChkB0NeverO	ChkRd
3	Monitor		x									x			x		x		x	
4	Decoder			x								x			x	x	x		x	
5	JustHeartBeat	x												x		x	x	x		
6	SelfTest				x	x	x		x					x	x					x

◀ ▶ @Configurations @Scenarios <queries> @Global @Local @TimeBounds @DataT

Figure 3-50 Special Excel tables: @-annotation, XML-annotation, and configurations

The @Configurations influence which rows of the annotations are selected when producing specific instances of the UPPAAL specification and requirements; e.g., the SelfTest configuration in row 6 triggers the SelfTesting feature (column) to be active, which in turn will trigger the parameters of the last row of the @TimeBounds-annotation table to be used.

Bringing all together with Uppex

Uppex is a tool that applies the configurations in the Excel sheets and composes different instances of Uppaal models automatically. Furthermore, it calls Uppaal to verify the requirements and produces a report, when Uppaal is installed. Figure 3-51 depicts this workflow. When asked to verify all configurations in a spreadsheet, Uppex produces a report similar to the one in Figure 3-52, explaining which requirement was marked as valid, failed, or timeout, and in which configuration.

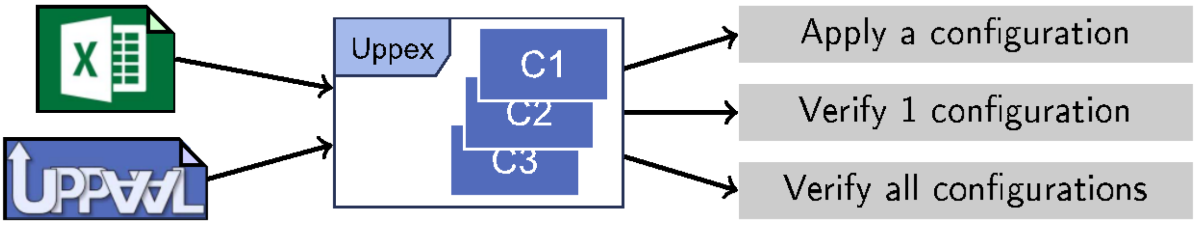
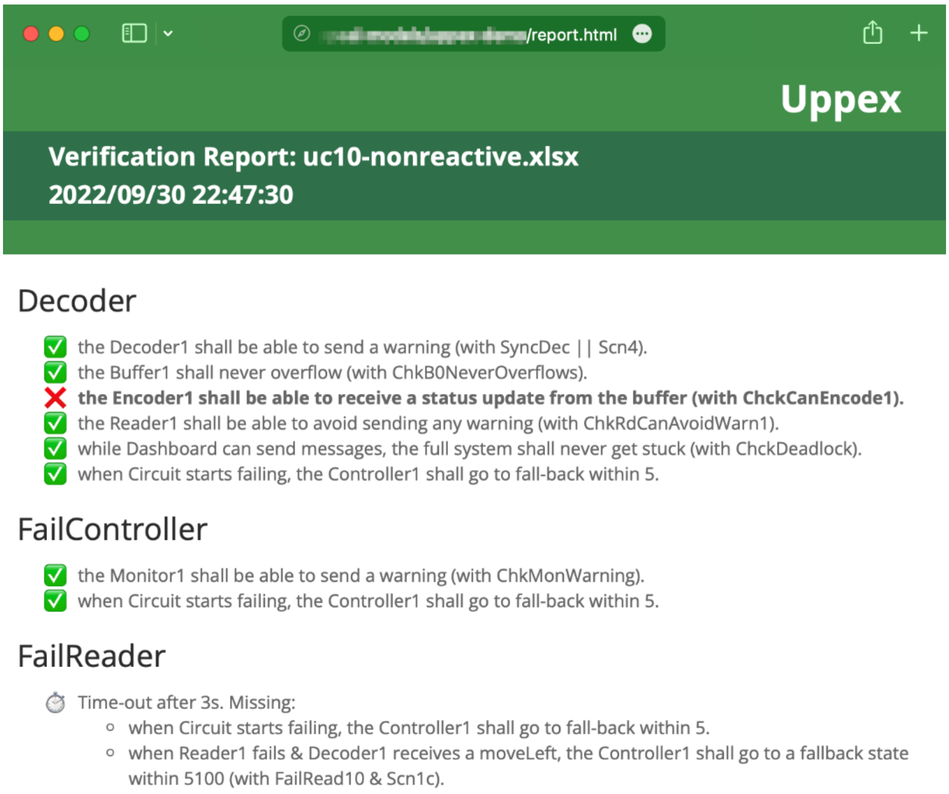


Figure 3-51 Uppex workflow: updating and verifying models based on configuration tables



[redacted]/report.html

Uppex

Verification Report: uc10-nonreactive.xlsx
2022/09/30 22:47:30

Decoder

- ✓ the Decoder1 shall be able to send a warning (with SyncDec || Scn4).
- ✓ the Buffer1 shall never overflow (with ChkB0NeverOverflows).
- ✗ **the Encoder1 shall be able to receive a status update from the buffer (with ChckCanEncode1).**
- ✓ the Reader1 shall be able to avoid sending any warning (with ChkRdCanAvoidWarn1).
- ✓ while Dashboard can send messages, the full system shall never get stuck (with ChckDeadlock).
- ✓ when Circuit starts failing, the Controller1 shall go to fall-back within 5.

FailController

- ✓ the Monitor1 shall be able to send a warning (with ChkMonWarning).
- ✓ when Circuit starts failing, the Controller1 shall go to fall-back within 5.

FailReader

- ⏱ Time-out after 3s. Missing:
  - when Circuit starts failing, the Controller1 shall go to fall-back within 5.
  - when Reader1 fails & Decoder1 receives a moveLeft, the Controller1 shall go to a fallback state within 5100 (with FailRead10 & Scn1c).

Figure 3-52 Screenshot of a verification report produced by Uppex after calling UPPAAL

<u>MoMuT - Model-based testing (AIT, LLSG, ALSTOM)</u>

In the second year of the project, the model-based test case generator tool MoMuT [38] was integrated into the use case application. For this purpose, a UML test model of the target control FSM was developed. The MoMuT tool generates test cases using a mutant-based method to insert faults in the test model, which are defined by mutation operators. All test cases which identify the faults are identified and sorted (kill mutants). As a result, a set of best-matching test cases are collected that guarantee a high level of test coverage. These test cases are transferred back to the use case application and will be used in the test process. The MoMuT tool is parametrized to process UML test models.

Special focus LLSG with Enterprise Architect is on the following items:

- User interface redesign with a focus on usability (design and implementation)
- Performance (implementation)
- Optimization of interface for communication with BE (definition)
- Continued work to optimize MoMuT
- Regular work in use cases UC10 and UC13 to adapt the functionality of tool MoMut development, creating test cases with the mutation strategy. In the case of MoMuT::UML this is a UML model of the input/output behaviour of the SUT to be tested. In UC1 the activities are focused on Security and Cyber-security by utilising the tool ThreatGet.

In this case, a test model of the motor controller was partly implemented with the Enterprise Architect tool, which is a support tool for MoMuT and provides a graphical model designer to define the test model in UML syntax. Figure 3-53 shows the class diagram of the motor controller UML test model. In the centre, there is the *TurnoutController*, which sends status data to the *InterlockingSystem* and which controls the *TurnoutMotor* with position data from the *TurnoutPositionSensor*.

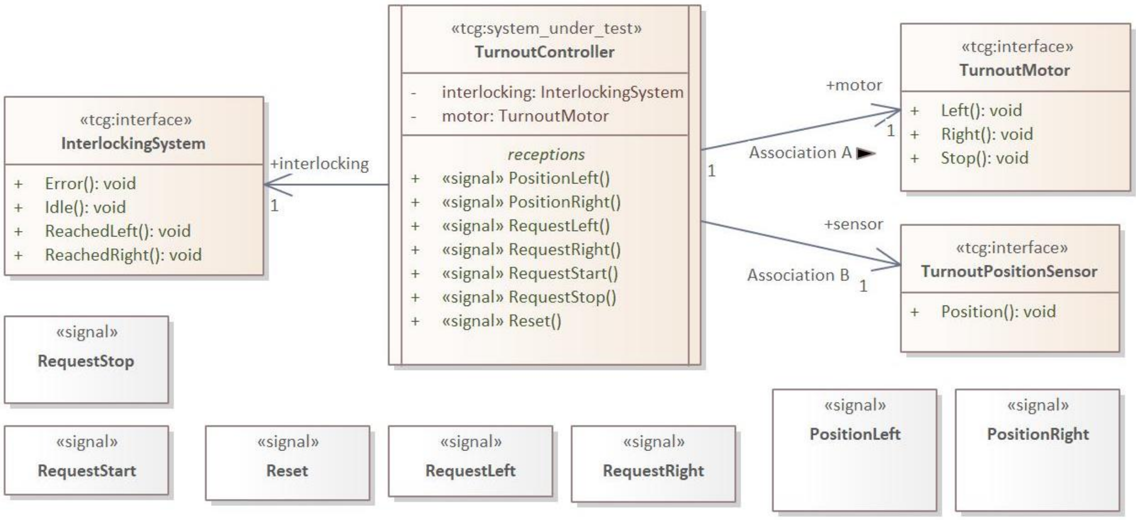


Figure 3-53 UML class diagram of the motor controller UML test model (UC10)

Figure 3-54 shows the partially implemented state machine of the motor controller (*TurnoutController*). Then, as partly documented in D5.4, the state machine consists of the states: Stand By, Ready for Command, Moving Left, Moving Right and Error.

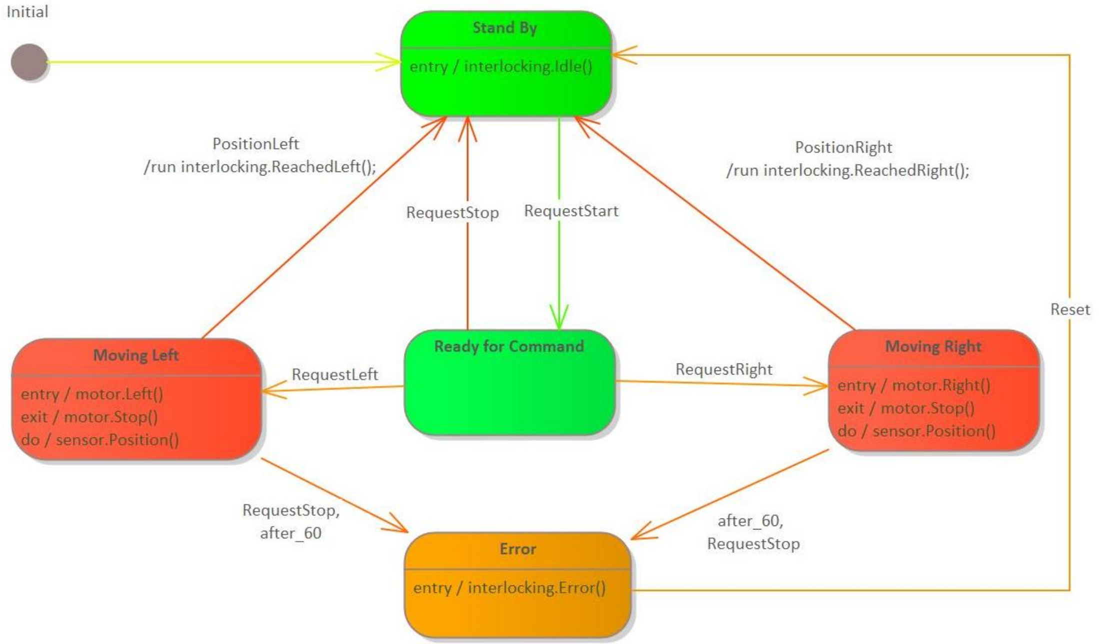
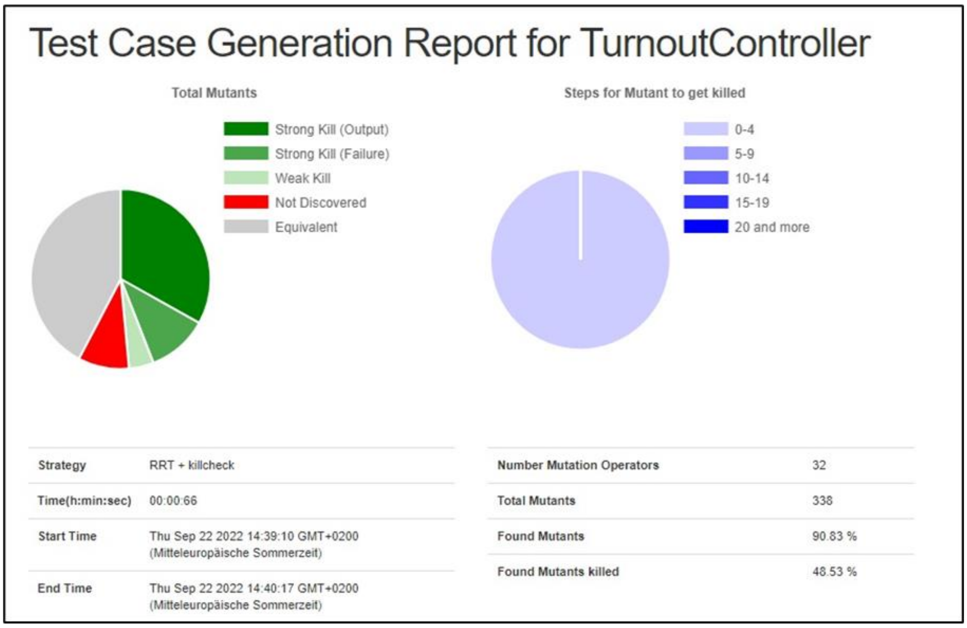


Figure 3-54 UML state machine diagram of the motor controller UML test model (UC10)

This UML test model is executed by the MoMuT tool, which is processed externally on a processing server hosted in Vienna. The model mutation process, the test case evaluation and the test case filtering are extensively processing jobs and need a powerful processing engine. Additional to the generated set of test cases an informative test case generation report is presented as shown in Figure 3-55.



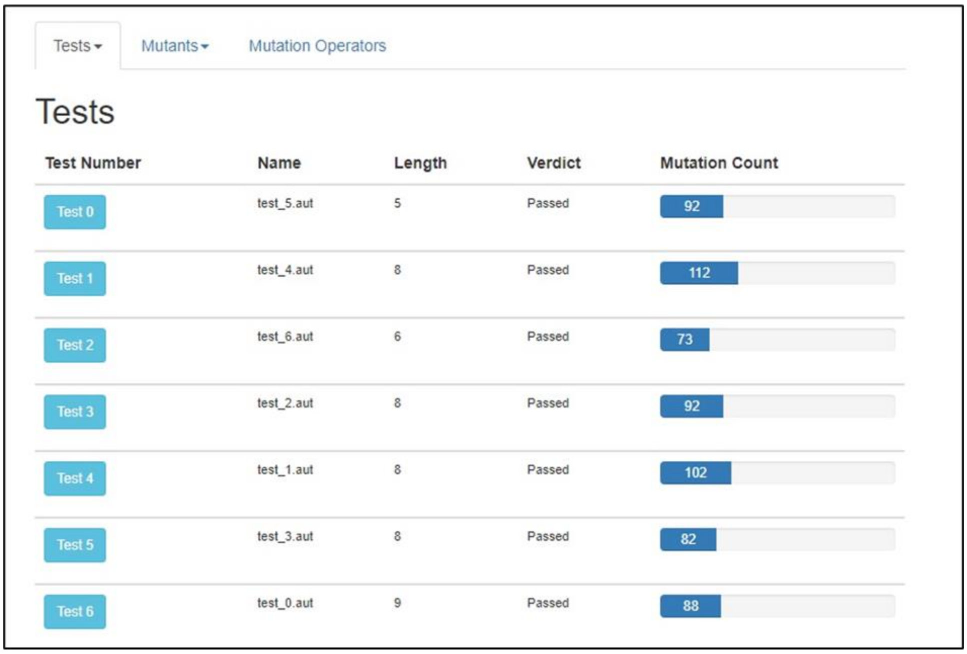
Test Case Generation Report for TurnoutController

Strategy	RRT + killcheck
Time(h:min:sec)	00:00:66
Start Time	Thu Sep 22 2022 14:39:10 GMT+0200 (Mitteleuropäische Sommerzeit)
End Time	Thu Sep 22 2022 14:40:17 GMT+0200 (Mitteleuropäische Sommerzeit)

Number Mutation Operators	32
Total Mutants	338
Found Mutants	90.83 %
Found Mutants killed	48.53 %

Figure 3-55 MoMuT test case generation report, Part A (UC10)

The test case generation report states that there are seven relevant test cases identified. The test cases are listed in the generation report and give a statistical overview of how many mutants are killed by each test case as shown in Figure 3-56.



Tests

Test Number	Name	Length	Verdict	Mutation Count
Test 0	test_5.aut	5	Passed	92
Test 1	test_4.aut	8	Passed	112
Test 2	test_6.aut	6	Passed	73
Test 3	test_2.aut	8	Passed	92
Test 4	test_1.aut	8	Passed	102
Test 5	test_3.aut	8	Passed	82
Test 6	test_0.aut	9	Passed	88

Figure 3-56 MoMuT test case generation report, part B (UC10)

One improvement in VALU3S is the graphical presentation of the test case process flow using a UML sequence diagram. MoMuT delivers the use case description in a textual definition format and

Enterprise Architect supports the graphical representation of the test case sequence flow, see Figure 3-57.

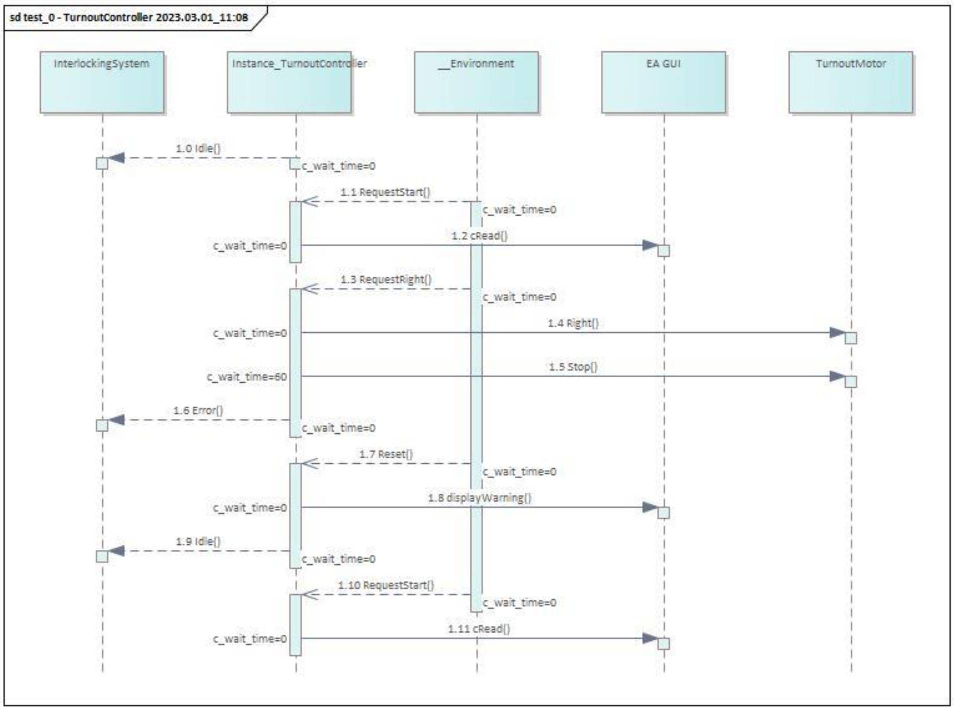
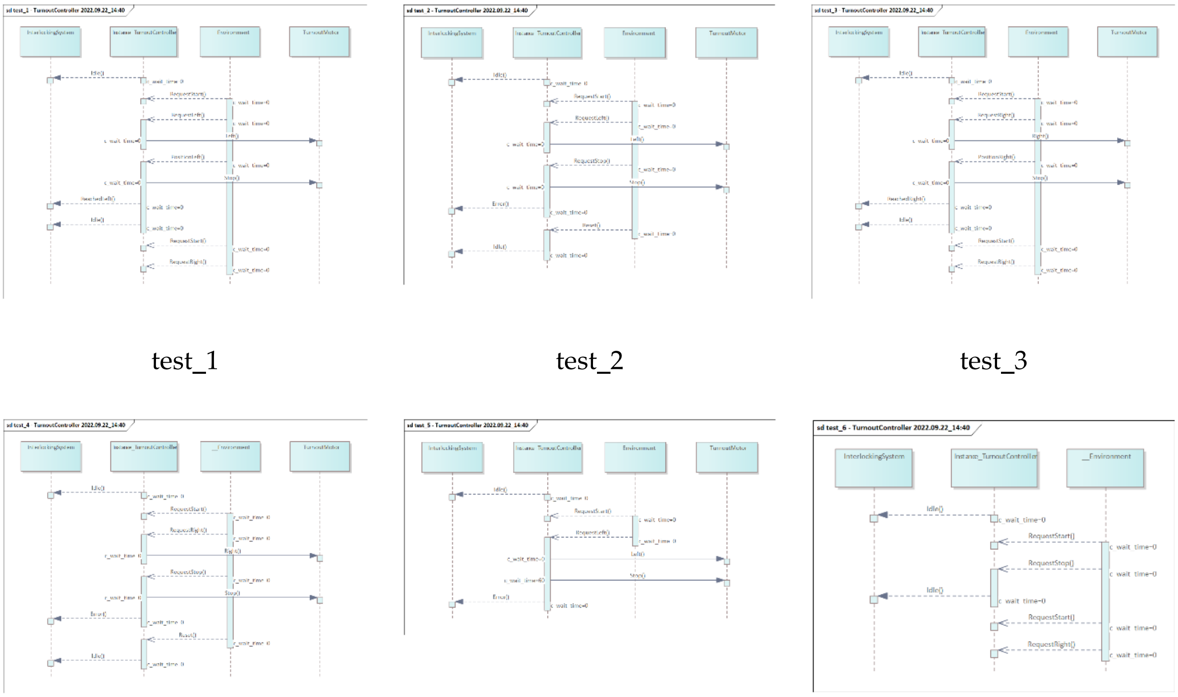


Figure 3-57 MoMuT test case sequence flow for test_0 (UC10)

On top of the diagram, the central _*Environment* of the test suite is placed beside the class elements of the model. The dashed lines show process flows, initiated by methods, and the solid lines show process flows, imitated by commands.

In sum, MoMuT identified seven relevant test cases, with which all defined faults (mutation operators) will be covered. The additional six dedicated test case sequence diagrams of test_1 to test_6 are shown in the following collage Figure 3-58.

In the last year of the project, the MoMuT tool and Enterprise Architect implantation in the use case will continue, especially performing detailed adjustments of the integration and an improvement of using the test cases in the use case application.



test_1 test_2 test_3

test_4 test_5 test_6

Figure 3-58 MoMuT Overview of identified test cases test_1 to test_6 (UC10)

3.10.5 Quantitative Results

Evaluation criteria for SCP:

Eval_SCP_10 – Software Fault Tolerance Robustness. With this criterion, we try to measure the efficiency of our fault injection approach to increase our confidence in the system's safety. Recall that we combine two formal methods: model-based test case generation with MoMut and model-checking with UPPAAL/Uppex, producing strong formal guarantees over the functional safety of this system, and contributing towards the certification for SIL 4 applications. For this criterion we use the first method: we automatically inject faults in a manually written behavioural model of our system and produce test cases to be used in the real implementation that cover these faults.

- *Evaluation results*: In this criterion, we measure the number of faults injected, and covered by the resulting test cases. We counted a total of 146 behavioural faults that are injected and verified in the real implementation via testing.
- *Baseline*: Based on the internal expertise of Alstom with similar projects, we estimated that a typical developer would manually write test cases that would cover around 20 behavioural faults.
- *Conclusions*: We estimate an increase in the number of faults covered by a factor of 7.

Evaluation criteria of the V&V process:

Eval_VV_10 – Reduced Cost and Time for Work on Certification Process and Functional Safety. With these criteria, we try to measure the V&V methodologies in this use case in terms of usage, effort, and efficiency. More concretely, we measure both the effort required to create tests that cover behaviour models using MoMut, and the effort to formally verify properties using UPPAAL/Uppex. This effort is measured in person-days, by keeping an estimate of how many people were involved and multiplying this value with the average accumulated time spent on these tasks. Furthermore, we consider the effort-per-result. I.e., we divide this effort in person-days by the number of results: this is the number of faults covered with MoMut, and the number of properties and variations of the formal model with UPPAAL/Uppex. We call this final number the "*effort-ratio*" of each of our two formal methods. Note that, unlike the previous criterion, more-is-worse, i.e., a larger effort-ratio reflects a larger time and cost per result, which is not desirable.

- *Evaluation results*: We measured the following effort-rations for each of our 2 approaches.
    - Using MoMut we estimated that 4 people worked for around 5 working weeks (100 person-days), covering 146 injected faults, resulting in an effort-ratio of 0.68.
    - Using UPPAAL/Uppex we estimated that 3 people worked for around 5 working days (75 person-days), analysing around 10 properties for 10 variations of the formal model, resulting in an effort-ratio of 0.75.
- *Baseline*: Based on the internal expertise of Alstom with similar projects, we estimated that:
    - one person would take around 8 working weeks (40 person-days) to cover 20 behavioural models with manual tests, and
    - a second person would take also around 8 working weeks (40 person-days) to formally verify 8 temporal properties of two variations of the system.

  These results, respectively, in effort-ratios of 2 (i.e., 40/20) and 2.5 (i.2., 40/(8*2)).
- *Conclusions*: We estimate a reduction of the effort-ratio by 65% with MoMut, and by 70% with UPPAAL/Uppex. I.e., 1-0.68/2 and 1-0.75/2.5.

The evaluation results described above are summarised in Figure 3-59 comparing against baselines the number of behavioural faults (for Eval_CSP_10, where *more is better*) and the amount of effort-ratio (for Eval_VV_10, where *less is better*).

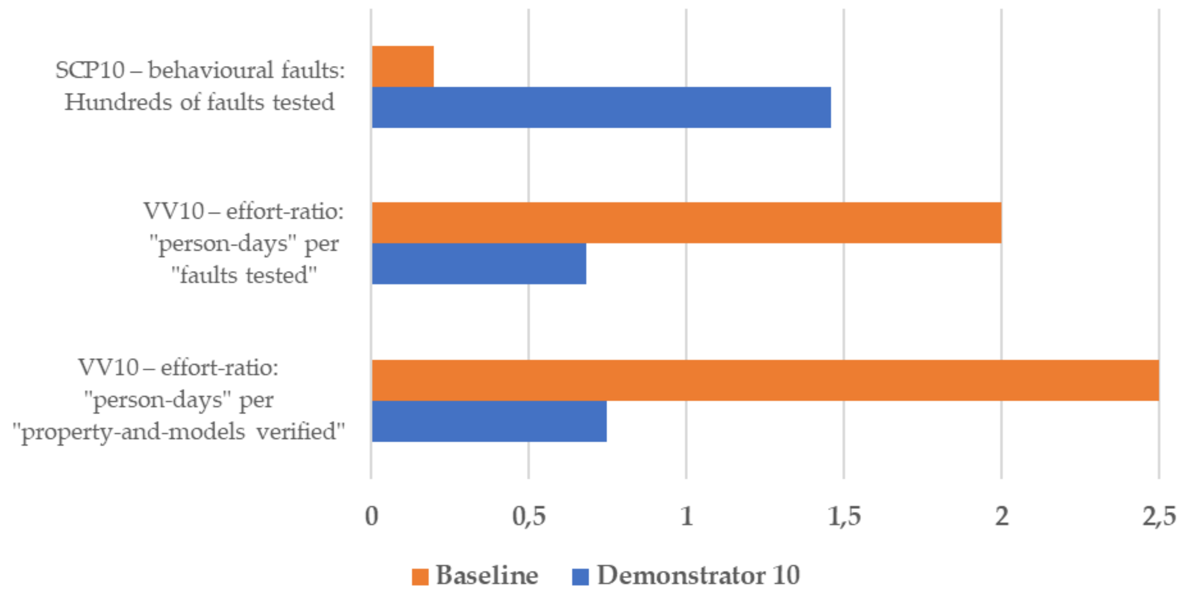


Figure 3-59 Evaluation results of Demonstrator 10; more faults covered is better (SCP10), and less effort is better (VV10)

3.10.6 Qualitative Results

The qualitative assessment is applied at the overall use case level. The subjects' profile and the statistical analysis results are as follows:

Participants Profile: QAM is applied to 10 subjects (9 males, 1 unknown) aged in the range of 25-53. The education level is relatively high as the subject pool is composed of 1 Post-Doc or higher-degree and 1 PhD researcher and 8 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "software/hardware engineers, researchers, etc." having experience in the fields of "railway domain, formal verification, real-time systems, cyber-physical security etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-106. The results show that all constructs are correlated with each other except the BI-CO pair.

Table 3-106. UC10 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.865	1									
MO	0.813	0.533	1								
CO	0.664	0.846	0.461	1							
ROI	0.884	0.767	0.754	0.712	1						
PE	0.566	0.308	0.715	0.268	0.720	1					
PT	0.783	0.929	0.441	0.900	0.849	0.391	1				
PR	0.754	0.503	0.736	0.538	0.923	0.760	0.644	1			
SI	0.845	0.948	0.515	0.770	0.676	0.114	0.823	0.433	1		
ATU	0.932	0.836	0.711	0.768	0.931	0.547	0.862	0.824	0.807	1	
BI	0.633	0.335	0.589	-0.020	0.564	0.753	0.268	0.540	0.230	0.496	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-107, the questions asked to subjects are sufficiently reliable as understood from subject responses, except ROI, PR and BI.

Table 3-107. UC10 Reliability Analysis

Cronbach-alpha values	
PU	0.712
PEOU	0.712
MO	0.172
CO	0.536
ROI	-0.262
PE	0.629
PT	0.419
PR	-0.346
SI	0.107
ATU	0.327
BI	-0.555

Regression Analysis: Regression analysis is applied to estimating the relationships among QAM constructs, as seen in Table 3-108. For this demonstrator, there exists an inversely proportional relation between BI and {ATU, ROI, PE, SI, PT, PR} whereas the other pairs influence each other in the same direction.

Table 3-108. UC10 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 1.64xPU - -4.36	0.661	0.004	3.952
H2	CO-PU	Right	CO = 0.783xPU + 1.42	0.441	0.036	2.515
H3	PEoU-PU	Right	PEoU = 1.22xPU - 1.2	0.748	0.001	4.877
H4	PU-ATU	Right	PU = 0.64xATU + 2.27	0.869	0	7.301
H5	PEoU-ATU	Right	PEoU = 0.811xATU + 1.35	0.699	0.003	4.312
H6	ATU-BI	Inverse	ATU = -0.62xBI + 2.34	0.246	0.145	1.617
H7	ROI-BI	Inverse	ROI = -0.9xBI + 0.85	0.318	0.09	1.929
H8	PE-BI	Inverse	PE = -1.304xBI - 1.71	0.567	0.012	3.235
H9	SI-BI	Inverse	SI = -0.2xBI + 4.81	0.053	0.522	0.67
H10	PT-BI	Inverse	PT = -0.33xBI + 4.22	0.072	0.454	0.787
H11	PR-BI	Inverse	PR = -0.969xBI + 0.26	0.072	0.107	1.814

3.10.7 Observed Limitations, Lessons Learnt and Best Practices

Through our investigation into the feasibility of achieving the highest level of safety according to the EN 50129 standard, we have discovered that it is possible to attain this goal by utilizing a minimal set of functional safety commercial-off-the-shelf (COTS) components. Our analysis has shown that certified components, compliance with ISO 26262, can improve the heat signature, reduce the size and cost, and most importantly, enhance confidence in the safety function in the railway domain. We have observed that incorporating Safety Element out of Context (SEooC) components in this railway use case can further improve the safety case and hence reduce the certification time. Overall, our findings suggest that by using certified SEooC COTS components, it is possible to (1) achieve the highest level of safety conforming to the EN 50129 standard also (2) improve cost-efficiency and (3) reduce certification time.

Using Uppex and UPPAAL to model-check the software used in this demo, we produced a relatively detailed behavioural model. This model is parameterised to experiment with different configurations. Some of the challenges that we encountered include: (1) how to understand if the model matches the system under development; and (2) how to parameterise and decide what to abstract when addressing different properties to be verified. We learned that the development of the formal behavioural model in parallel to the development of the implementation was beneficial for both formal and practical developers, as key decisions were made at development time, further strengthening the connection between the model and the system. We also gained useful insights by discovering what parts of the system could be abstracted or simplified to verify different safety requirements. Furthermore, we observed that a large number of non-deterministic actions could occur, leading to many uncertainties (and state explosions when model-checking). This leads us to start producing a new version of this software, currently under development. This new version avoids most of this non-determinism by design, producing more predictable behaviour and more trustworthy software.

Using test-case generation with MoMuT we were able to produce a small set of optimised tests, capable of covering a large number of behavioural faults. This generation is based on a simplified behavioural model, containing fewer details than the one used by Uppex/UPPAAL, and focused on the core controller task. Some of the remaining challenges that we encountered include: (1) how to increase the trust in this behavioural model used to generate the test scenarios, and (2) how to integrate the generated test scenarios into the software system. Regarding (1), the behavioural model was produced by MoMuT experts while interacting with domain experts (Alstom), providing some guarantees over its correctness; and regarding (2) the integration effort is expected to be relatively large, and avoided due to time limitations and because this demonstrator will not be used in production.

The qualitative assessment results, as indicated in Table 3-109, present a high user acceptance (greater than 5,10/7.00) of all QAM constructs. One of the main reasons for this positive attitude can be the promising results of the quantitative evaluation as the proposed model checker and the underlying solution stack significantly increase the fault detection performance. Although the interfaces are still being developed the respondents seem convinced with the demonstration outputs. Potential end users also think that the proposed innovation may reduce the total workforce needed. Potential further improvement that will enable the automatic integration of test case results and better interaction of the users with the proposed solution stack may increase the level of acceptance.

Table 3-109 Mean and standard deviation of experts' responses to UC10

UC10	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,43	5,43	5,98	5,13	5,60	5,82	5,10	5,80	5,29	5,75	5,88
Std Dev	1,13	0,80	0,56	0,96	0,60	0,55	0,79	0,53	1,09	0,77	0,96

3.11 Use Case 11 - Automated Robot Inspection Cell for Quality Control of Automotive Body-in-White (UC11)

UC11 aims to provide a better fault-tolerant production system to achieve better quality control for automotive body-in-white, see Figure 3-60. Controlling the existence of 2500-3000 body parts is planned to be executed fully automatically by a cartesian robot and a camera-based sensor system [1].

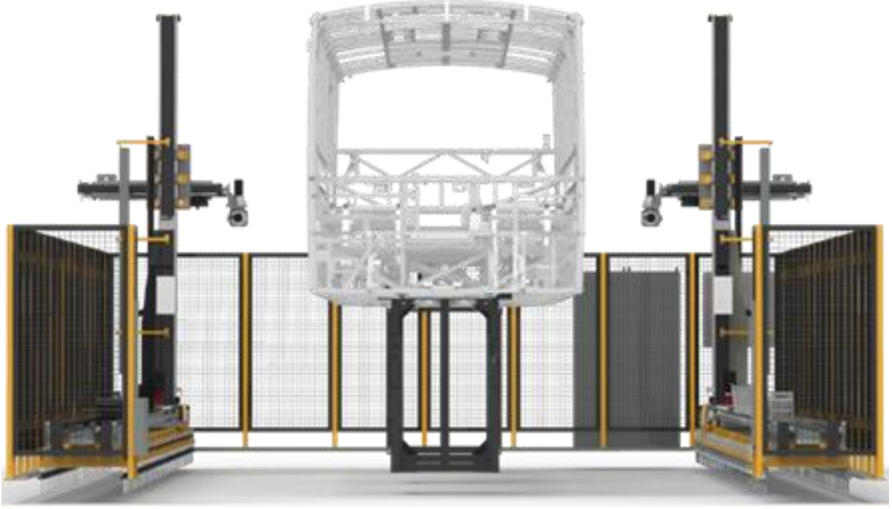


Figure 3-60 Robot inspection cell for quality control in UC11

To ensure that VALU3S technology applies to the robot inspection cell for quality control in this use case, we cover an automated fault and attack injection mechanism, specified for controlling the entire industrial automated line. In this case, there is a need to increase the autonomy of the system by ensuring the safety of the system. Autonomous trajectory generation methods, optimized according to time and safety constraints, are developed in UC11. In addition, the robotic system is evolving to be able to perceive the current state of the environment in real time and implements a dynamic motion plan by considering the state of the environment. The safety of the system has been verified both in the current system that is operating in OTOKAR's Sakarya premises and the system enhanced with the intended improvements covering the security risks as well. The safety requirements cover the safety of the robot and its apparatus as well as static objects in the workspace. As humans are not involved in the automatic mode, human safety will not be studied, but in the manual mode, human safety will be encountered by maintaining a secure communication channel between the operator and the robotic system, see Figure 3-61 for system topology. The use case will be evaluated considering in particular that VALU3S facilitates security and safety by [1]:

- Demonstrating results from simulations and the role of VALU3S in decision-making.
- Exercising and evaluating the existing tools and methods and identifying their strengths and weakness aligned with potential opportunities and threats (SWOT).
- Assessing the full inspection processes in terms of task completion rate, duration and safety metrics; time required to detect and overcome faults and attacks; reducing the number of production faults within a unit time interval or process.
- Manipulation of the data that is collected from system components and stored in the automation system.
- Observation of the inspection process flow data to detect anomalies in production phases; and regarding software-level attacks. Anomaly detection at component and system levels by utilizing ML and/or deep learning-based techniques.
- Qualitative assessment of results by applying the questionnaire to experts having experience in related fields

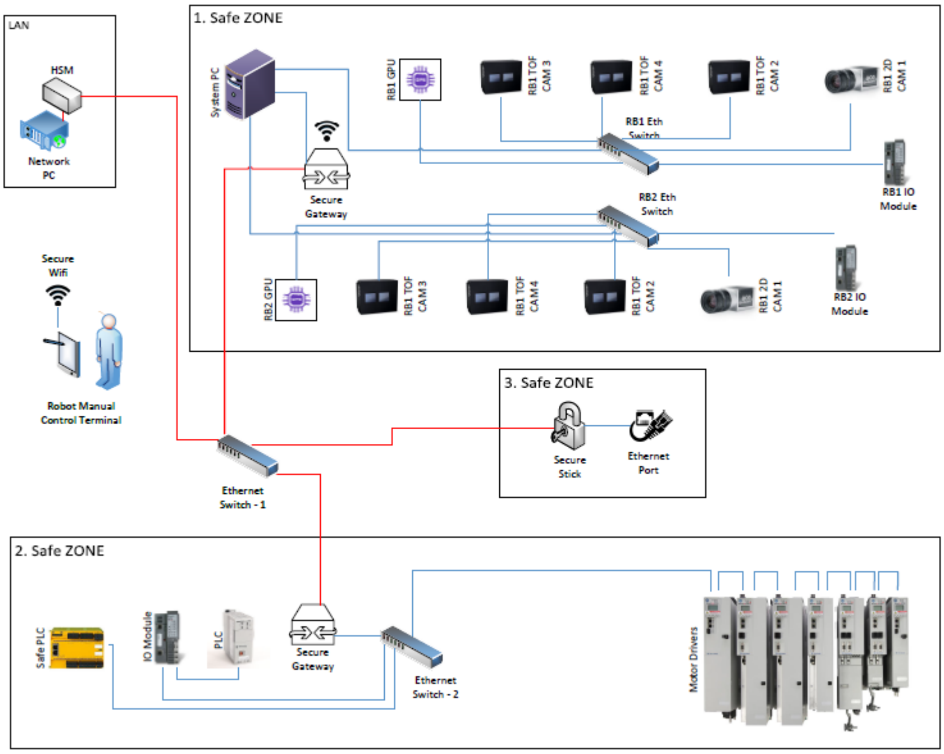


Figure 3-61 Detailed topology of the system investigated in UC11

3.11.1 V&V Challenges

V&V challenges associated with UC11 deal with both security and safety risks. Body-in white system of UC11 works with PLCs, gateways and robotic arms in the real world where both internal and external operators may access the subsystems and components either physically or virtually. To prevent any potential accidents or inefficiencies, a simulation-based quality inspection application is also developed. The experiments are planned to meet the real-life challenges of the actual system which is being currently used in Otokar's premises. Additionally, an experimental setup is installed in IFARLAB (@ESOGU) to test the multistakeholder and multi-institutional collaboration.

There can be many challenges in complex systems, like the targeted robotic solution in Otokar. In VALU3S, the contributors mainly focused on the following V&V challenges to improve the existing solution in Otokar and make it accessible from other virtual end nodes over trustworthy working environments and more optimised and safer in daily quality inspection processes:

- **Manipulation of data and the IoT backend**: Manipulation/corruption of sensor data stream at camera and safety sensors are a critical challenge as it may cause big delays in production processes and also quality problems in automotive body-in-white quality. For instance, manipulation of camera sensor data may affect the quality of process results. Additionally, manipulation of safety data results in safety breaches (robot-human, robot-product, robot-workspace objects collision) or manipulation/corruption of the PLC data stream may increase the risks related to the system failure, collision risks, production faults, or unpredictable system behaviour.
- **Safety Trajectory Optimization**: The robotic inspection system of UC11 can create serious safety problems by hitting humans and surrounding objects due to faulty or missing software and the inability to create appropriate trajectories. Verification of the safety of robot trajectory that automatically covers the robot and its apparatus as well as static objects in the workspace requires an effort due to multiple scenarios like robot-human, robot-product, and robot-workspace objects collision checking. In addition, it is also a challenge to verify that the system is safe in the presence of security attacks on the system.
- **Anomaly detection at the component and system level**: The use of novel AI/ML-based techniques is needed to be adapted to the specific data associated with the UC11 operations to improve the anomaly detection accuracy. The main challenge may be the observation of the inspection process flow data to detect anomalies in production phases and component parameter data by utilizing ML and/or deep learning-based techniques. Such challenges may cause component and system failures.
- **Cyber security-related challenges**: It is indispensable that cyber-physical resilience checks and vulnerability analysis against cyber-attacks are needed. Manipulation/corruption of data flow due to MiTM, DoS and ARP Poisoning Attacks may cause serious system failures, data breaches, unauthorised access, and even physical threats like a collision.

3.11.2 Contributors

Partners contributing to the UC: OTOKAR, ERARGE, ESOGU, IMTGD, TECHY

3.11.3 Contributors' Roles & Evaluation Scenario

List of evaluation scenarios defined for this use case [2] (see also Table 3-110):

1. VALU3S_WP1_Industrial_1: Manipulation of sensor data. Manipulation/corruption of sensor data stream at camera and safety sensors. Partners' roles are as follows: OTOKAR: Testbed

preparation; ERARGE: Improvement and implementation of HSM (Hardware Security Module) and SG (Secure Gateway) tools; IMTGD: Data analysis and integration of the simulation tool.

2. VALU3S_WP1_Industrial_2: Server and PLC communication. Partners' roles are as follows: OTOKAR: Manipulation/corruption of PLC data stream, testbed preparation and implementation of the tool; ERARGE: improvement and implementation of the HSM (PRIGM) tool.
3. VALU3S_WP1_Industrial_3: Safety Trajectory Optimization. Creating robot trajectory points automatically covers the safety of the robot and its apparatus as well as static objects in the workspace. Partners' roles are as follows: OTOKAR: Formalisation of requirements. Implementation and improvement of the OTOKAR Simulation Tool. Runtime validation test time improvement will be reported; ESOGU: implements the verification for the safety of the system by utilizing model-checking and runtime verification via developed MARver tool; IMTGD: - While the chassis body control system is running, a fault injection will be made into the robot source files, and after this situation, it will be checked whether the system makes incorrect movements and situations.
4. VALU3S_WP1_Industrial_4: Anomaly detection at component and system level. Observation of the inspection process data flow to detect anomalies in production phases and component parameter data by utilizing ML and/or deep learning-based techniques. Partners' roles are as follows: OTOKAR: testbed preparation; ERARGE: implementation and integration of PRIGM; IMTGD; Performs fault injection studies in the simulation environment; TECHY: Data collecting, analysis and preparation for ML/AI algorithms.
5. VALU3S_WP1_Industrial_20: Server Ethernet Network Security. Manipulation/corruption of data flow due to MiTM, DoS and ARP Poisoning Attacks. OTOKAR: studies are carried out using Penetration Tests and tools. ESOGU: studies are carried out for run time verification of security under unexpected network traffic flow.

Table 3-110 Overview of contribution to evaluation scenarios by UC11 partners

Evaluation Scenario	OTOKAR	ERARGE	ESOGU	IMTGD	TECHY
VALU3S_WP1_ Industrial_1	X	X		X	
VALU3S_WP1_ Industrial_2	X	X			
VALU3S_WP1_Industrial_3	X		X	X	
VALU3S_WP1_ Industrial_4	X	X		X	X
VALU3S_WP1_ Industrial_20	X	X	X		

UC11 aims to provide an improved fault-tolerant production line to achieve better quality control for automotive body-in-white. The existing quality check processes are still very time-consuming, ineffective, and lacking advanced safety concepts. Additionally, quality check in the existing manufacturing environment is not very responsive and adaptive to online sensing. It works in the Stop & Go mode to provide safety. Figure 3-62 illustrates the V&V tools that are being developed and planned to be used/demonstrated in this use case as well as the V&V methods associated with the tools.

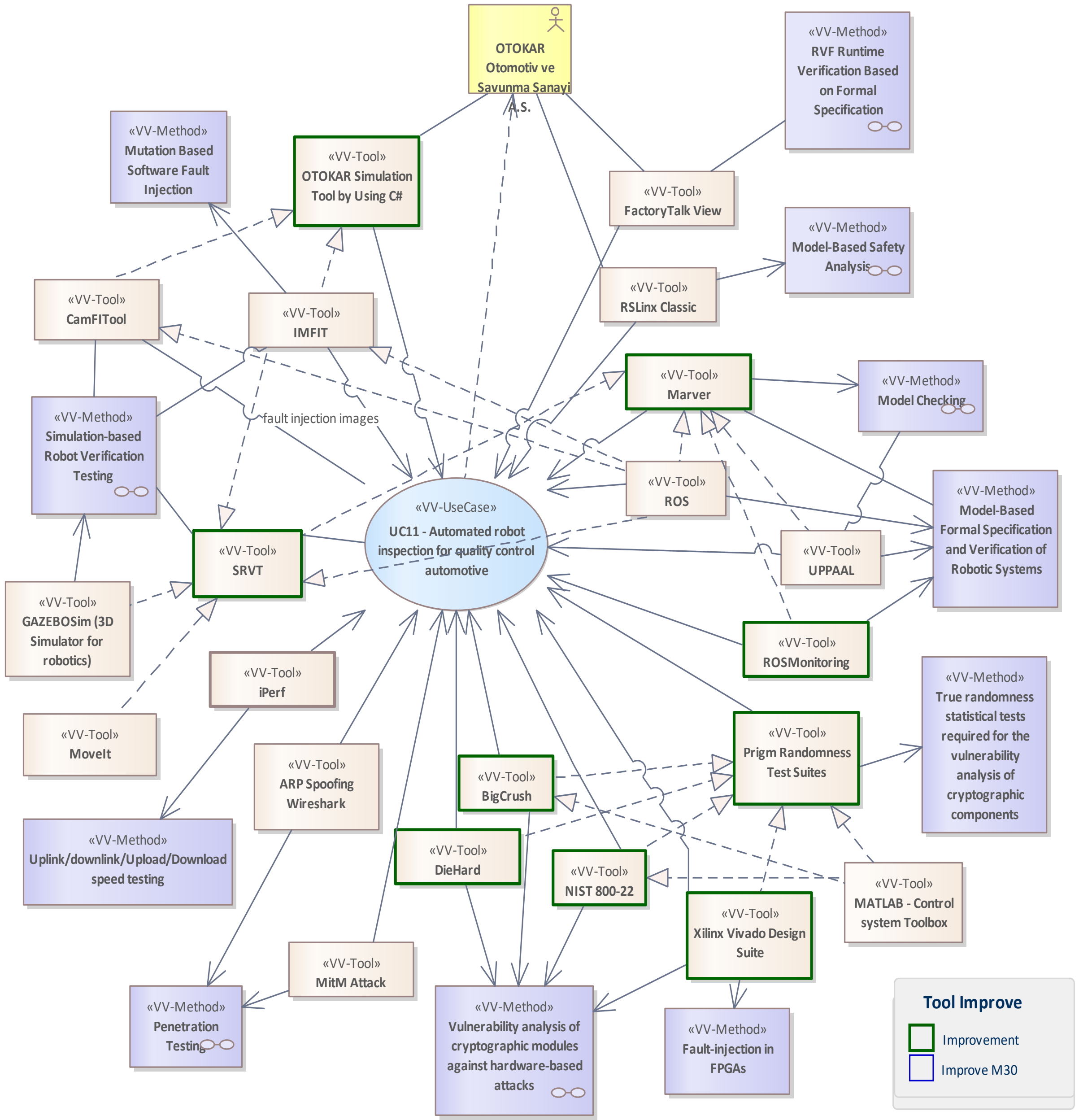

OTOKAR Otomotiv ve Savunma Sanayi A.S.
«VV-Method» RVF Runtime Verification Based on Formal Specification
«VV-Method» Mutation Based Software Fault Injection
«VV-Tool» OTOKAR Simulation Tool by Using C#
«VV-Tool» FactoryTalk View
«VV-Method» Model-Based Safety Analysis
«VV-Tool» RSLinx Classic
«VV-Tool» CamFITool
«VV-Tool» IMFIT
«VV-Tool» Marver
«VV-Method» Model Checking
fault injection images
«VV-Method» Simulation-based Robot Verification Testing
«VV-Tool» ROS
«VV-Method» Model-Based Formal Specification and Verification of Robotic Systems
«VV-UseCase» UC11 - Automated robot inspection for quality control automotive
«VV-Tool» SRVT
«VV-Tool» UPPAAL
«VV-Tool» GAZEBOSim (3D Simulator for robotics)
«VV-Tool» ROSMonitoring
«VV-Tool» iPerf
«VV-Method» True randomness statistical tests required for the vulnerability analysis of cryptographic components
«VV-Tool» MoveIt
«VV-Tool» Prigm Randomness Test Suites
«VV-Tool» ARP Spoofing Wireshark
«VV-Tool» BigCrush
«VV-Method» Uplink/downlink/Upload/Download speed testing
«VV-Tool» DieHard
«VV-Tool» NIST 800-22
«VV-Tool» MATLAB - Control system Toolbox
«VV-Tool» Xilinx Vivado Design Suite
«VV-Tool» MitM Attack
«VV-Method» Penetration Testing
«VV-Method» Vulnerability analysis of cryptographic modules against hardware-based attacks
«VV-Method» Fault-injection in FPGAs
Tool Improve
Improvement
Improve M30

Figure 3-62 Tools and Methods for UC11 – Automated Robot Inspection for Quality Control Automotive

The list of evaluation criteria for the V&V process is as follows [8]:

1. Eval_VV_1 - Time of test execution: In the penetration test, the cases with and without HSM and SG in the system topology are compared. (OTOKAR)
2. Eval_VV_2 - Coverage of Test Creation: During the first stage, the model of the system was created. The use of the model pattern approach, which facilitated the modelling process, was tried. With the model checking method, the model is checked whether it meets the system requirements. In the second stage, the system software developed in accordance with the model is verified at runtime by the runtime verification method. Thus, two staged verification method extends the coverage of verification. (TECHY)

3. Eval_VV_3 - Number of Test Cases: Detection and elimination of gaps in the system with Penetration Test. (OTOKAR)
4. Eval_VV_4 - Effort for Test Creation: Verification is done in two stages: Model checking and runtime verification. The configuration for runtime verification is done by utilizing the models created in the first stage. Thus, runtime verification, which monitors the system and performs the testing process on behalf of the human, becomes easier. In addition, the developed modules such as the online distance tracker and online human tracker facilitate the testing process. (ESOGU)
5. Eval_VV_8 - Testing Effort: Workforce needed for testing of the system (new system or redesigned version) during development and before deployment to the field. (TECHY)
6. Eval_VV_11 - Randomness and Security Assessment Process Performance: Randomness tests defined by NIST are available as open source. Randomness and Security Assessment Process Performance: Assessing the randomness and security should be time- and effort-efficient as the cyber-physical systems to be validated and verified are complex systems and need to be restarted as soon as possible for their actual work. Moreover, the employment of less personnel effort is also crucial to improve labour efficiency. Thus, the randomness and security assessment tests should be performed as fast and efficiently as possible. (ERARGE)

List of evaluation criteria of the SCP process [8]:

1. Eval_SCP_1 - Error Coverage: The operator using the system accesses the system with the basic user's local Windows account. Then, unauthorised attempts are realised to increase the authorisation rights. After system access is provided, technical rights escalation tests are performed. (OTOKAR)
2. Eval_SCP_3 - Number of Malicious Attacks and Faults Detected: The condition that the system software works in accordance with the model and meets the requirements at runtime is evaluated by the runtime verification method. Meanwhile, the system traffic load increases. During the runtime validation, the effect of anomalies in this system on the behaviour implementation according to the expected pattern is observed. The error and attack relationship is verified. (OTOKAR, ESOGU)
3. Eval_SCP_4 - Metrics to Evaluate AI/ML Algorithm: It is recorded every minute regarding environmental conditions during operation, and every 10 minutes under non-operational conditions. Hardware parameters: load voltage, load current, temperature, motor vibration data, sensor parameters, camera parameters, and robot arm parameters are recorded 10 times per minute. Each record is printed as one line to the CSV text file. Each line starts with a timestamp and ends with a tag. At least 100 000 rows of data are collected. All columns are normalized to [0,1]. Missing data cells are complete. Error status labels, 0: error-free state, 1: sensor error, 2: motor error, 3: Software error, 4: PLC error, 5: planning error, 6: model error. Input is divided into three subgroups: training (70%), validation (10%), testing (20%). The k-means algorithm is applied for clustering. (TECHY)
4. Eval_SCP_5 - Potential Impact of Incidents and Attacks: After the environmental and internal data are recorded, labelled appropriately for possible error conditions and trained with different ML methods, error conditions are randomly injected into the system during testing and the developed method to predict these errors and classify them within an acceptable limit

is tested. Error injection can be done with the IMFIT tool or with different FI tools. Algorithms and approaches that give successful results in the simulation environment are also evaluated in the real environment. (TECHY)

5. Eval_SCP_7- Number of prevented accidents: To see both sides of the robots, human identification is made in the images taken from the cameras to be placed, the proximity of the robot will be determined and human entry into these orbits will be detected. Depending on the entered field, the robot slows down or is stopped. (OTOKAR, ESOGU)
6. Eval_SCP_10 - Authentication Accuracy and Time Applied to Human Users and Components: Attempts to access the Secure Authentication Device HSM and/or SecureStick configuration interface, referred to as the SecureStick. To model this, the webauthn.io web service using the FIDO U2F standard was used. A user with SecureStick enters this site and registers to the webauthn.io service by plugging his device into the PC via the USB interface, then logs in from another PC via the same service within 24 hours and the login is successful. (ERARGE) An authentication accuracy test has been done, but time has not yet been performed.
7. Eval_SCP_11 - Randomness and Cryptographic Algorithm Strength: Randomness tests defined by NIST are available as open source. 1-million-bit random number set obtained from HSM will be subjected to 4 different basic randomness tests. Results will also be evaluated according to NIST's results rubrics. The randomness tests can be realised on a live system where the HSM generates bunches of random numbers continuously and the online assessment takes place to verify the true random number generator at the backend hardware (ERARGE)
8. Eval_SCP_10 - Software fault tolerance robustness: In the chassis control system operating scenario, by running the robot source code (ROS) with the mutated code, it is observed that the robot system overcomes the errors and continues its normal operation in cases where the faulty code is applied. (ESOGU, IMTGD) Docker image versions of SRVT are created to test the mutated codes created on IM-FIT in SRVT. In this way, fault injection tests continue to be performed on more than one working SRVT container at the same time. (IMTGD)
9. Eval_SCP_13 - Accuracy of Simulated Sensor Output: While the chassis body control system is running, fault injection is applied to the robot source files, and after this situation, it is checked whether the system makes incorrect movements and situations. The intended situation is that the robot does not behave dangerously when running faulty source codes. (IMTGD) Tests on the SRVT system, which had a fault injected into its cameras using CamFITool, are being expanded and work on IM-FIT SRVT integration continues. While SRVT is running, various errors are injected into both ROS packages and Python codes and how the system is affected by these injections is examined. (IMTGD)

3.11.4 Demonstration:

Demonstration for UC11 was planned by individual partners. Therefore, multiple demonstrable items cover defined challenges and scenarios, and partially VALU3S dimensions (see Table 3-111).

Table 3-111 Overview of demonstration prepared by UC11 partners.

Item #	Demonstration name	Description/Purpose	Format	Responsible
1	OTOKAR Simulation Tool decides the optimum, safe robot trajectory to make the part existence check of the vehicle. Then simulates the image processing steps, which normally occur in run time.	To prevent software-related errors and accidents by creating a digital twin of robots in the field and predict the effects of possible errors, which can occur in run-time, on the system.	Lead Demo Video demo/PowerPoint presentation	OTOKAR
2	Performing safety trajectory planning tests with a combination of SRVT and IM-FIT. (Eval Scenario 3)	Performing verification and validation tests of the UC11 simulation environment running on SRVT.	Complementary Video demo/PowerPoint presentation	IMTGD
3	Demonstration of CamFITool manipulation of sensor data and anomaly detection modules. (Eval Scenario 1 & 4)	Detection of anomalies in real and manipulated pictures taken from the UC11 environment with a CamFITool interface plugin using models trained using the CNN algorithm.	Complementary Video demo /PowerPoint presentation	IMTGD
4	Demonstration of an Integrated Verification for Safety and Security of Industrial Robot Inspection System with the developed tool MARVer	Implement an application of the developed tool MARVer in the simulated and real TRL5 test environment of UC11 to demonstrate the safety verification concerning collision under security issues. Note that the MARVer could also performs verification of safety and security independently.	Complementary Video demo/ PowerPoint Presentation	ESOGU

Item #	Demonstration name	Description/Purpose	Format	Responsible
5	Performing live true randomness tests for the cryptographic backend	Online and live true randomness test is required to verify and/or validate the secret generation or key generation scheme that is used to encrypt critical data within the quality inspection system, e.g., snapshots captured from the automotive body-in-white components.	Complementary Video demo/power point presentation/live demo	ERARGE

3.11.5 Quantitative Results

Demonstrator-1: OTOKAR Simulation Tool

We model the real environment in the virtual environment with the Otokar simulation tool. Thanks to this modelling, we test all experimental data in the real environment and test whether there are any collisions or anomalies. We have already noticed any accidents that may occur in the system, we are working to prevent them, and we increase the safety of the system. On the other hand, we have the chance to observe how accurate the fault injection tool and safety trajectory optimization studies are.

Camera, robot, engine speeds and collision situations were modelled in the Otokar simulation tool. Robot speeds obtained from PLC data were entered into the simulation. In case of any work, visual colouring studies were carried out. About 25 minutes when we test robot speeds in real time; When we disable real time, we test the vehicle in 10-12 minutes. One advantage of this is that we get faster results when the images are injected with the fault injection tool. At the end of the simulation scan, it generates a report. We compare the percentages of seen parts in the real environment and the virtual environment.

Demonstrator-2: Performing safety trajectory planning tests with a combination of SRVT and IM-FIT

According to the data in the given test results, BiEST and RRT algorithms performed successfully in terms of key performance indicators. It was observed that the percentage of task completion **increased by 27.9%** by removing points that may be unnecessary for planners from the task lists. Also, quest **completion times can be reduced from 20 minutes**. Based on these data, it was concluded that the dynamic planning system applied for the ROKOS system works more effectively at fewer location points. **Nearly 900 hours of testing were performed** on the **seven motion planning algorithms** of the OMPL planner and the most suitable planning algorithm was found. At the end of the study, it was determined that the best planning algorithm was the **BiEST algorithm**. It was observed that the task planning completion times using the BiEST algorithm reduced the task completion time of the ROKOS system from 25 minutes. The average times obtained on the ROKOS robot arm were found to be approximately 19 minutes. By removing reset points from the ROKOS task list and using the most appropriate planning algorithm, the ROKOS system **achieved a 20% task completion time gain**. In this way, time was saved in the bus production line and production efficiency was increased.

Verification and validation processes were carried out on the ROKOS system with mutation-based tests performed with IM-FIT. In the tests performed, a total of 4954 mutant codes were obtained in the mutation process for the source codes of the ROKOS system. The ROKOS system was executed by replacing the mutant codes obtained by IM-FIT with the source codes. As a result of the execution, 4560 of 4954 mutants were in the "killed" status and 394 of them were in the "survived" status. The mutation score of the tests performed was evaluated as 92.0468%. Verification and validation processes were carried out for the ROKOS system with mutation-based tests performed with IM-FIT.

Demonstrator-3: Demonstration of CamFITool manipulation of sensor data and anomaly detection modules

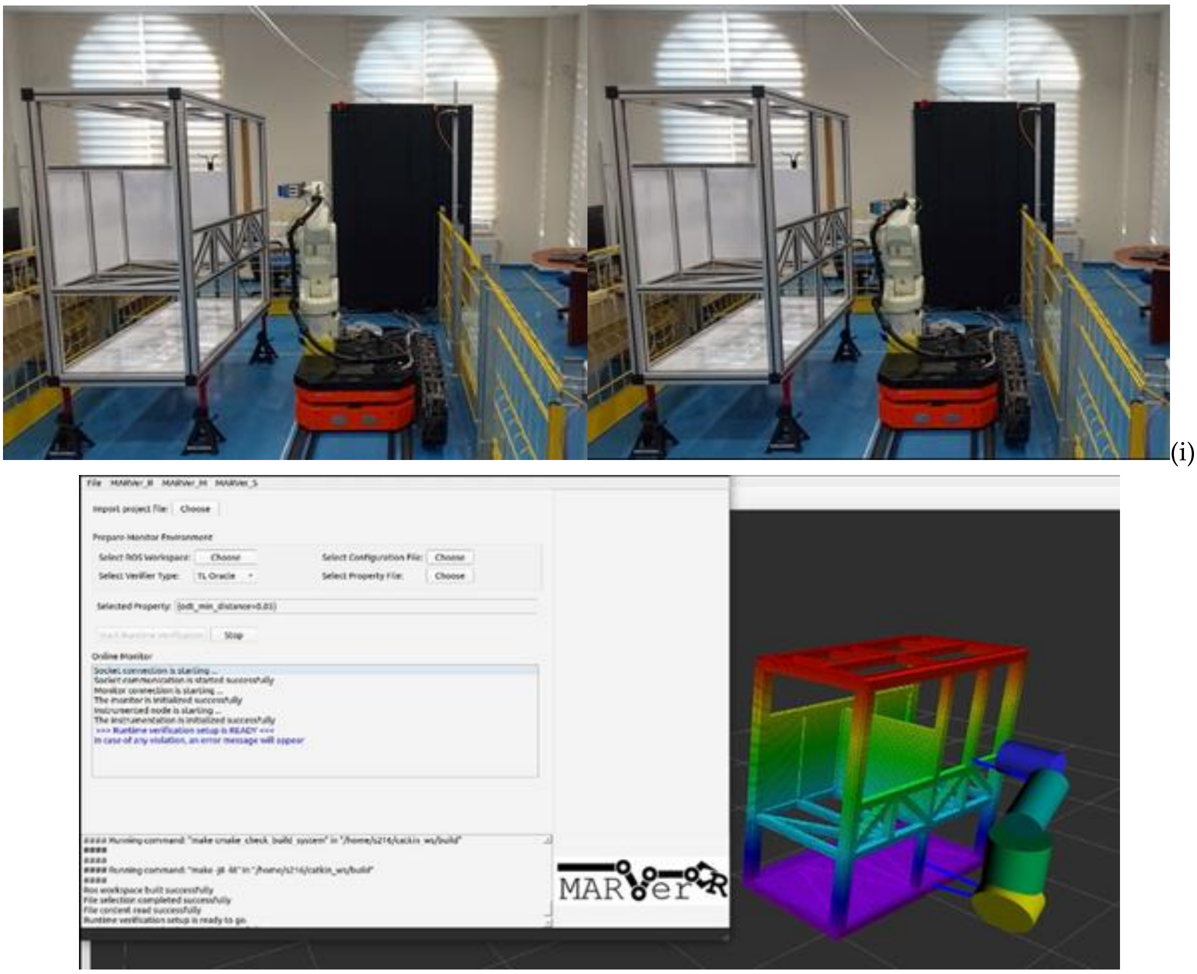
Fault injection tests using CamFITool were performed using **49 different test variations** with different fault rates and different image amounts to **293 real ROKOS environment images**. **9 different fault types** were injected, namely Salt&Pepper, Gaussian, Poisson, Open/Close, Dilation/Erosion, Gradient and Motionblur. The tests were also carried out in two different test categories single fault-focused and multiple fault-focused. In the tests carried out, the types of faults that do not affect the quality control system, affect the quality control system or affect it are determined. The **system robustness value was calculated to be 95.39%.**

CamFITool detects faults in the images using the CNN algorithm. It was first tested using a faulty image library of **4200 images with an accuracy rate of 93.18%**. Afterwards, a binary classification model was created, considering that it is necessary to determine whether the picture is faulty before detecting the type of faults in the images. For the training of the binary model, **two different test libraries** consisting of a new fault list and normal images were used, and an **accuracy rate of 87.5% was obtained**. In the developed interface, the binary model first detects whether the image to be queried is faulty, and then it is estimated which fault it has in the multiclass model. The multiclass model, which was updated with the new system, **achieved an accuracy rate of 80.0%**, while the accuracy value **after the training was measured as 0.9316.**

Demonstrator-4: Demonstration of an Integrated Verification for Safety and Security of Industrial Robot Inspection System with the developed tool MARVer

The demonstration of the results of the developed V&V tools is implemented on a TRL5 experimental setup for quality inspection of the automotive body-in-white platform (mini-ROKOS). One of the most important safety problems in the robotic inspection system is the collision that will occur if the robot trajectory is obtained incorrectly to cause a collision or if an unexpected part is found in the robot trajectory.

Figure 3-63 shows some snapshots from the experiment. As seen in Figure 3-63 (top), there is an unexpected part across the robot trajectory. While the robot moves along the trajectory, it stops at a safe distance to the unexpected part as seen in Figure 3-63 (bottom). At the same time, the MARVer tool monitors the distances between the robot and the parts in the environment. And, it reports that the safety software of the system is verified.


(i)

File MARVer_R MARVer_M MARVer_S
Import project File: Choose
Prepare Monitor Environment
Select ROS Workspace: Choose
Select Configuration File: Choose
Select Verifier Type: TL Oracle
Select Property File: Choose
Start Runtime Verification
Stop
Online Monitor
MARVer

Selected Property: (odt_min_distance>0.01)

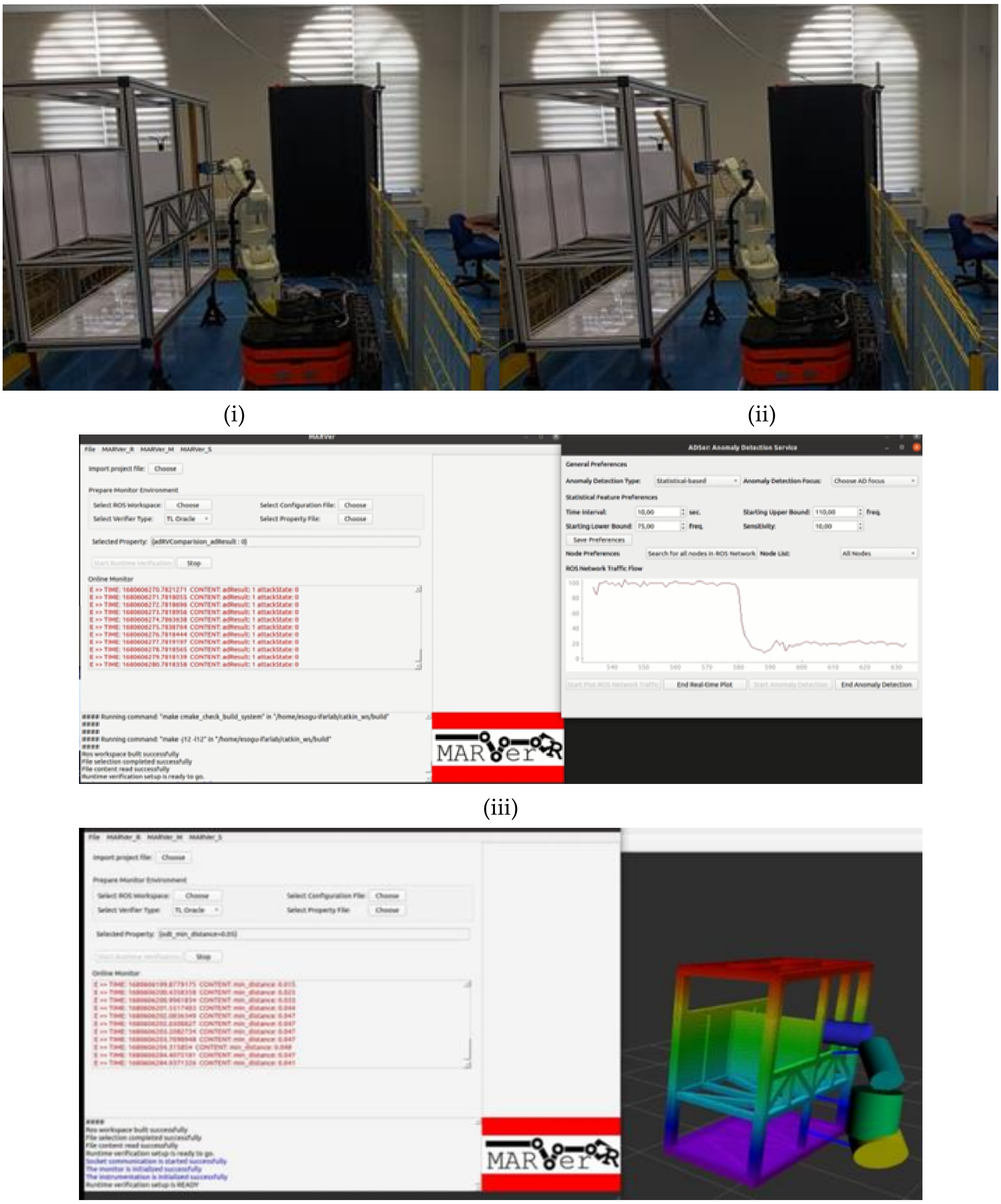
```
Socket connection is starting ...
Socket communication is started successfully
Monitor connection is starting ...
The monitor is initialized successfully
Instrumented node is starting ...
The instrumentation is initialized successfully
 >>> Runtime verification setup is READY <<<
In case of any violation, an error message will appear
```

```
#### Running command: "make cmake_check_build_system" in "/home/s216/catkin_ws/build"
####
####
#### Running command: "make -j8 -l8" in "/home/s216/catkin_ws/build"
####
Ros workspace built successfully
File selection completed successfully
File content read successfully
Runtime verification setup is ready to go.
```

Figure 3-63 System test with an unexpected part in the robot trajectory

In the second stage, the system performs the same inspection task again, however, this time the system is exposed to a security attack. The MARVer tool determines the attacks and reports a warning message as seen in Figure 3-64 (iii). Also, another service of the MARVer determines that the minimum distances fall below the safety threshold values as seen in Figure 3-64 (iv). Also, Figure 3-64 (i) and (ii) shows some snapshots from the experiments proving that the robot hits the unexpected part.

Thus, the current system software meets the safety requirements by not colliding with unexpected parts when there is no security attack in the system. However, when the system is exposed to security attacks, the system software needs to be improved not to cause safety issues.



(i)

(ii)

MARver
File MARver_R MARver_M MARver_S
Import project file: Choose
Prepare Monitor Environment
Select ROS Workspace: Choose
Select Configuration File: Choose
Select Verifier Type: TL Oracle
Select Property File: Choose
Start Runtime Verification
Stop
Online Monitor
ADtect Anomaly Detection Service
General Preferences
Statistical Feature Preferences
Save Preferences
ROS Network Traffic Flow
100
80
60
40
20
0
540
550
560
570
580
590
600
610
620
630
Start Plot ROS Network Traffic
End Real-time Plot
Start Anomaly Detection
End Anomaly Detection
MARver

Selected Property: {adRVComparision_adResult : 0}

Anomaly Detection Type: Statistical-based
Anomaly Detection Focus: Choose AD focus
Time Interval: 10,00 sec.
Starting Upper Bound: 110,00 freq.
Starting Lower Bound: 75,00 freq.
Sensitivity: 10,00
Node Preferences: Search for all nodes in ROS Network
Node List: All Nodes

```
E >> TIME: 1680606270.7821271 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606271.7818055 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606272.7818696 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606273.7818956 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606274.7863638 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606275.7838764 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606276.7818444 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606277.7819197 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606278.7818565 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606279.7818139 CONTENT: adResult: 1 attackState: 0
E >> TIME: 1680606280.7818358 CONTENT: adResult: 1 attackState: 0
```

```
#### Running command: "make cmake_check_build_system" in "/home/esogu-ifarlab/catkin_ws/build"
####
####
#### Running command: "make -j12 -l12" in "/home/esogu-ifarlab/catkin_ws/build"
####
Ros workspace built successfully
File selection completed successfully
File content read successfully
Runtime verification setup is ready to go.
```

(iii)

File MARver_R MARver_M MARver_S
Import project file: Choose
Prepare Monitor Environment
Select ROS Workspace: Choose
Select Configuration File: Choose
Select Verifier Type: TL Oracle
Select Property File: Choose
Start Runtime Verification
Stop
Online Monitor
MARver

Selected Property: {odt_min_distance<0.05}

```
E >> TIME: 1680606199.8779175 CONTENT: min_distance: 0.015
E >> TIME: 1680606200.4358358 CONTENT: min_distance: 0.025
E >> TIME: 1680606200.9561834 CONTENT: min_distance: 0.033
E >> TIME: 1680606201.5517483 CONTENT: min_distance: 0.044
E >> TIME: 1680606202.0836349 CONTENT: min_distance: 0.047
E >> TIME: 1680606202.6308827 CONTENT: min_distance: 0.047
E >> TIME: 1680606203.2082734 CONTENT: min_distance: 0.047
E >> TIME: 1680606203.7698948 CONTENT: min_distance: 0.047
E >> TIME: 1680606204.313854 CONTENT: min_distance: 0.048
E >> TIME: 1680606284.4075181 CONTENT: min_distance: 0.047
E >> TIME: 1680606284.9371326 CONTENT: min_distance: 0.041
```

```
####
Ros workspace built successfully
File selection completed successfully
File content read successfully
Runtime verification setup is ready to go.
Socket communication is started successfully
The monitor is initialized successfully
The instrumentation is initialized successfully
Runtime verification setup is READY
```

(iv)

Figure 3-64. System test when there is an unexpected part in the trajectory and the system is exposed to security attacks.

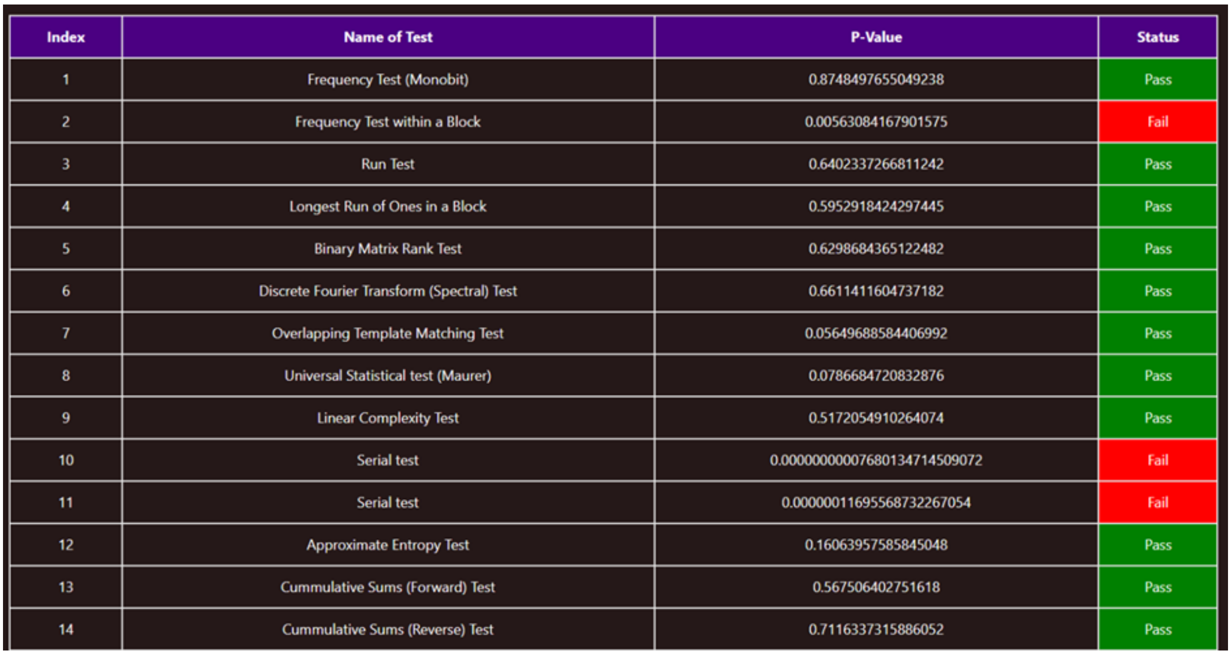
The MARver tool is used for 7 test cases under UC11. There is an average improvement of 75% for effort for test creation. The minimum improvement is 57% while the maximum improvement for effort for test creation is 91%. Malicious security attack detection accuracy is greater than 80%.

<u>Demonstrator-5: Performing live true randomness tests for the cryptographic backend</u>

The quantitative assessment in Demonstrator-5 is 4-fold:

1. **NIST SP 800-22 Randomness Analysis**: NIST SP 800-22 is a standard published by the United States National Institute of Standards and Technology (NIST). This standard is a test guide for

testing and evaluating random number generators. This standard consists of many different tests and these tests are used to measure the quality of a random number generator. Tests include serial tests, frequency tests, long string tests, correlation tests, and longest repeat tests. NIST SP 800-22 is used in many different fields. For example, it is used in the fields of cryptography and security to ensure that random number generators are working correctly. It is also used for testing and validating mathematical and statistical models. Test results are measured by the p-value, which is a measure of system randomness. The P-values are visible in the results snapshot below (Figure 3-65). The p-value takes a value between 0 and 1 and is usually tested against 0.05 or 0.01. A P value of approximately 0 indicates that the probability of the observed statistical pattern occurring by chance is very low. On the other hand, when the p-value is close to approximately 1, the observed pattern is likely to occur by chance. In this case, the observed pattern appears to have occurred by chance. As for the relation of the system with randomness, when the randomness is higher, the probability of the observed statistical patterns occurring by chance also increases. Similarly, in the case of low randomness, the observed statistical patterns are less likely to occur by chance.



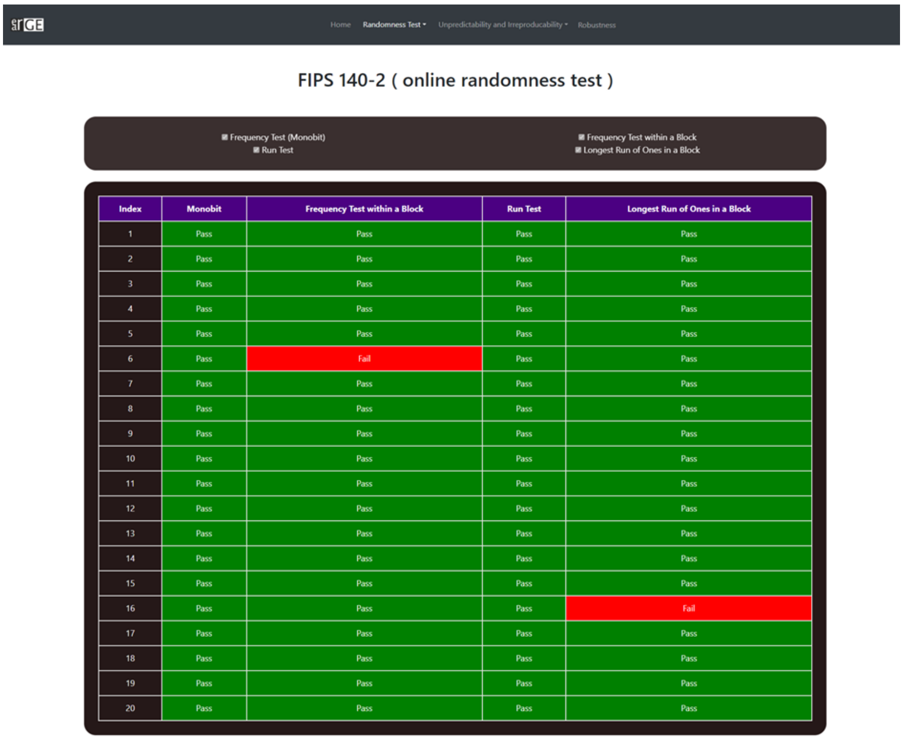
Index	Name of Test	P-Value	Status
1	Frequency Test (Monobit)	0.8748497655049238	Pass
2	Frequency Test within a Block	0.00563084167901575	Fail
3	Run Test	0.6402337266811242	Pass
4	Longest Run of Ones in a Block	0.5952918424297445	Pass
5	Binary Matrix Rank Test	0.6298684365122482	Pass
6	Discrete Fourier Transform (Spectral) Test	0.6611411604737182	Pass
7	Overlapping Template Matching Test	0.05649688584406992	Pass
8	Universal Statistical test (Maurer)	0.0786684720832876	Pass
9	Linear Complexity Test	0.5172054910264074	Pass
10	Serial test	0.00000000007680134714509072	Fail
11	Serial test	0.00000011695568732267054	Fail
12	Approximate Entropy Test	0.16063957585845048	Pass
13	Cummulative Sums (Forward) Test	0.567506402751618	Pass
14	Cummulative Sums (Reverse) Test	0.7116337315886052	Pass

Figure 3-65 Statistical test's P-value results according to NIST SP 800-22

2. **FIPS 140-2 Randomness Analysis:** FIPS 140-2 (Federal Information Processing Standard 140-2) is a cryptographic module certification standard published by the United States National Institute of Standards and Technology (NIST). This standard is used to evaluate and certify the security of cryptographic modules used by federal government agencies and their suppliers. FIPS 140-2 covers many different features of cryptographic modules. These features include key management, encryption, authentication, random number generation, and other cryptographic operations. FIPS 140-2 ensures that a certified cryptographic module is secure and meets a specified level of security requirements. To become certified for any of the 4 levels of FIPS 140-2, a cryptographic module must pass a specific set of tests that include the random number generation process. These tests may include:

- Random number generation: These tests aim to verify that the module generates random numbers and that these numbers are unpredictable. These tests include Frequency Test, Serial Test, Long String Test, Poker Test and other tests.
- Entropy source: The entropy sources used for the generation of the random numbers of the module must be sufficiently unpredictable and random. FIPS 140-2 does a specific set of tests to test the module's entropy source used.
- Production quality: The quality of the hardware and software used to generate random numbers directly affects the random number generation performance of the module. FIPS 140-2 tests the manufacturing quality of the module to ensure that the random number generation process works correctly.

The test results shown in Figure 3-66 are based on the test passing if the p-value, which is formed as a result of testing the numbers flowing live from the random number generator, is greater than 0.01. In this way, four basic randomness tests are performed in systems that need live random number generation.



FIPS 140-2 (online randomness test)

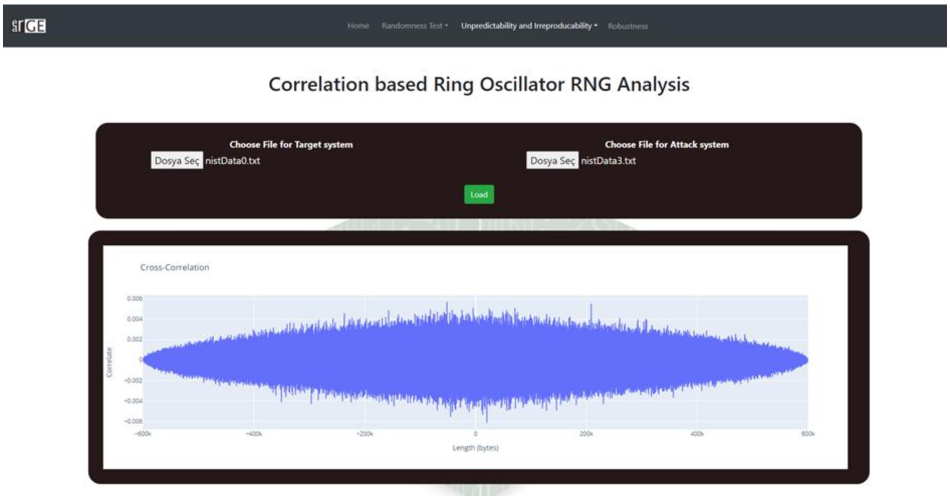
☑ Frequency Test (Monobit)
☑ Run Test
☑ Frequency Test within a Block
☑ Longest Run of Ones in a Block

Index	Monobit	Frequency Test within a Block	Run Test	Longest Run of Ones in a Block
1	Pass	Pass	Pass	Pass
2	Pass	Pass	Pass	Pass
3	Pass	Pass	Pass	Pass
4	Pass	Pass	Pass	Pass
5	Pass	Pass	Pass	Pass
6	Pass	Fail	Pass	Pass
7	Pass	Pass	Pass	Pass
8	Pass	Pass	Pass	Pass
9	Pass	Pass	Pass	Pass
10	Pass	Pass	Pass	Pass
11	Pass	Pass	Pass	Pass
12	Pass	Pass	Pass	Pass
13	Pass	Pass	Pass	Pass
14	Pass	Pass	Pass	Pass
15	Pass	Pass	Pass	Pass
16	Pass	Pass	Pass	Fail
17	Pass	Pass	Pass	Pass
18	Pass	Pass	Pass	Pass
19	Pass	Pass	Pass	Pass
20	Pass	Pass	Pass	Pass

Figure 3-66 Randomness test's P-value results according to FIPS 140-2

3. **Correlation-Based Cryptanalysis:** Correlation-Based Cryptanalysis (CBCA) is a crypto-analysis technique that poses a threat to the security of systems such as ring oscillators used for random number generation. Ring oscillators are a type of oscillator used to generate random numbers in digital systems. The output of these oscillators is a signal that rotates between a set of inverters or NAND gates that are fed back to each other by performing logic operations such as a set of NAND gates or inverters. This return is used to generate a random number. By analysing the output of these oscillators, CBCA tries to predict the internal structure of the

oscillator and the random number generation algorithm. In this analysis, the correlation level between the output signal of the oscillator and a predicted signal is calculated. This level of correlation can help predict the internal structure of the oscillator and the random number generation algorithm. CBCA can threaten the security of systems used for random number generation, such as ring oscillators. Therefore, measures can be taken to protect against CBCA, such as using more complex oscillator structures or using stronger random number generation algorithms. Also, using crypto-analysis techniques such as CBCA, the security of systems used for random number generation can be tested and improved. As presented in Figure 3-67, there is a tab designed for this purpose in the Randomness web tool. It can be used for correlation analysis of results obtained from two different ring structures of a ring oscillator. As a result, two random number datasets are subjected to correlation analysis and numerical results are displayed on a graph. A maximum correlation of 0.1 for two sets of random numbers is acceptable. If a correlation is measured above this, the correlation of the two sets of numbers is compromised and could result in a security vulnerability (Figure 3-68). If it is less than 0.1, the two sets of numbers can be judged to be unrelated (Figure 3-69).



Correlation based Ring Oscillator RNG Analysis

Choose File for Target system: Dosya Seç nistData0.txt

Choose File for Attack system: Dosya Seç nistData3.txt

Figure 3-67 Tab for Correlation-based analysis of a Ring Oscillator

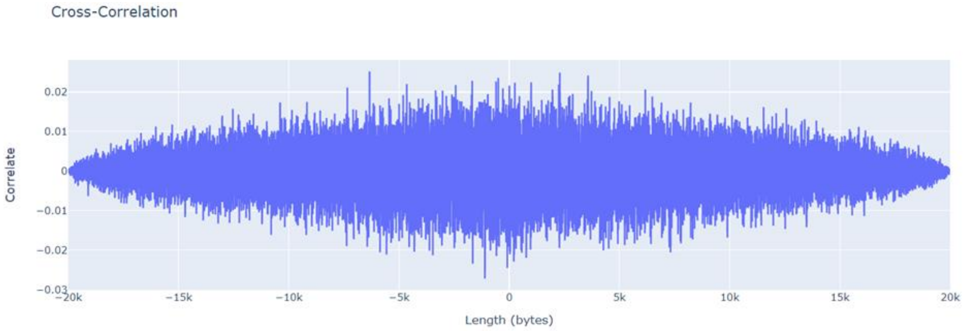


Figure 3-68 Correlation graph between two sources of a Ring Oscillator's sub-rings' outputs. Correlation is very low.

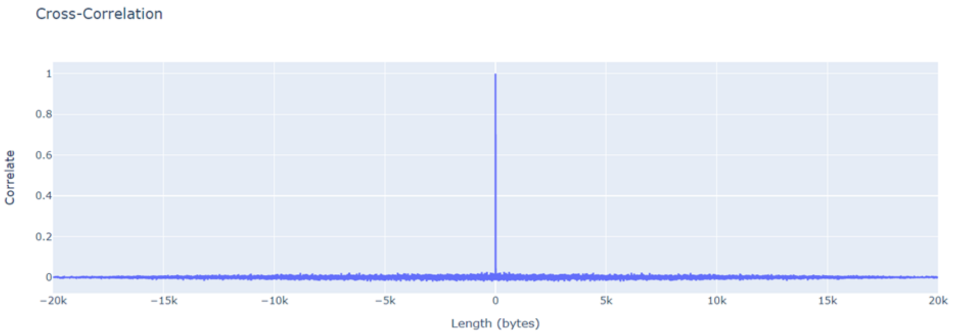


Figure 3-69 Correlation graph between two sources of a Ring Oscillator's sub-rings' outputs, Correlation is very high

4. **Synchronization Based Cryptanalysis:** The Jacobian matrix of a chaotic mathematical equation shows how much the velocity of each variable in the system varies with the velocity of the other variables. The Jacobian matrix provides information about the stability of each point in the system.

Each term is calculated as the derivative of the corresponding function, the relevant variable. For example, the term d(dx/dt) / dx is calculated as the derivative of the velocity of the variable x concerning the variable x. The Jacobian matrix provides information about the stability of the system and is especially important in the analysis of chaotic systems. CLE (Conditional Lyapunov Exponent) is a parameter that measures how a small deviation in a chaotic system grows over time. This parameter indicates that the system is chaotic and can be used to generate random numbers. CLE value is a criterion used to characterize the chaotic behaviour of dynamical systems. A negative CLE value means that minor faults in the system fade over time and the system behaviour is stable. A positive CLE value means that minor errors in the system grow over time and the system behaviour is chaotic.

All the quantitative analysis results indicate a positive impact on the industry. As seen in Figure 3-70, the industry stakeholders, experts in Otokar and Koç Holding, have a very positive opinions about the achievements. The results are well-adopted at a level higher than 80% in all evaluation criteria. These results show that the observations at the project end meet the expectations at the start. As supported by the qualitative assessment results, the proposed solution stack improves not only the overall efficiency of the automotive body-in-white quality inspection process but also its resilience against cyber-physical attacks, human error and safety issues.

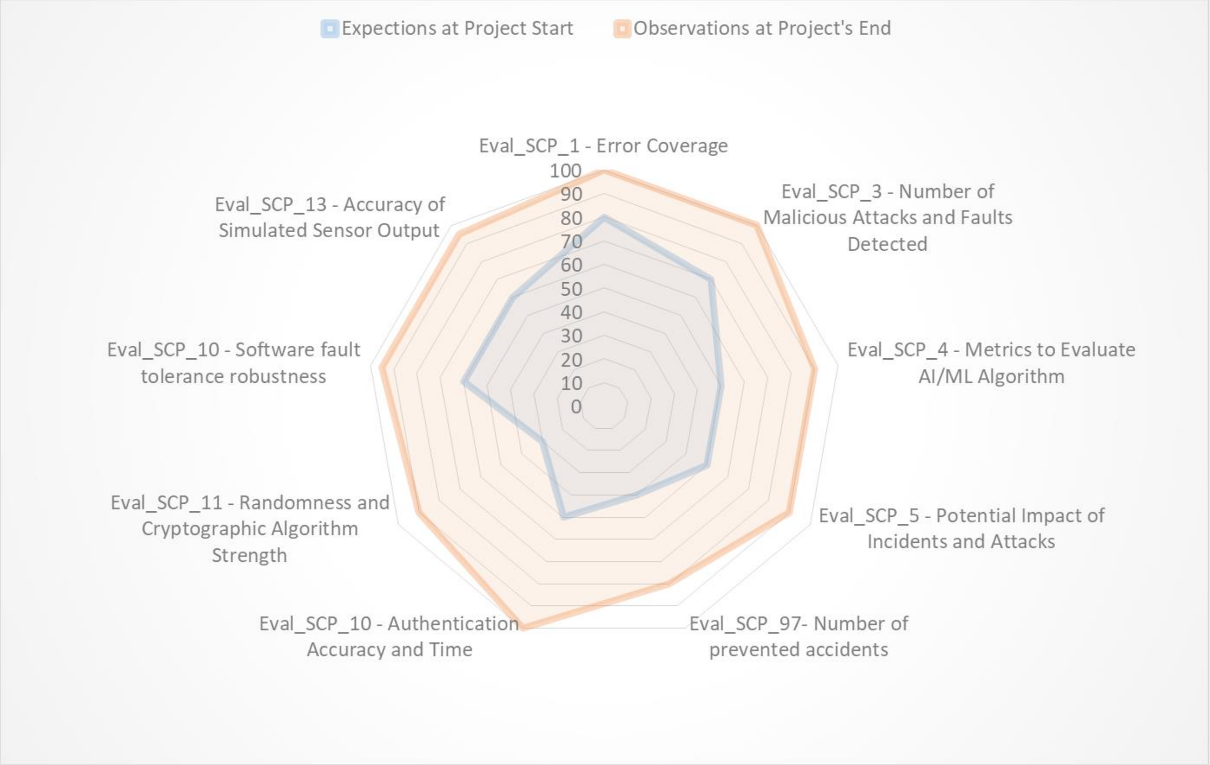


Figure 3-70 Coverage (%) of results adopted by the industry (Otokar and Koç Holding)

3.11.6 Qualitative Results

The qualitative assessment is applied at the overall use case level. The subjects' profile and the statistical analysis results are as follows:

Participants Profile: QAM is applied to 29 subjects (33 Males, 7 females, 1 non-binary) aged in the range of 24-48. The education level is relatively high as the subject pool is composed of 3 Post-Doc or higher-degree and 2 PhD researchers and 24 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "academicians, R&D Engineers, software/hardware engineers, embedded systems engineers, directors, managers, etc." having experience in the fields of "robotics, cyber-physical security, AI/ML, anomaly detection, industrial processes, automated manufacturing etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-112. The results show that all constructs are correlated with each other.

Table 3-112. UC11 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.699	1									
MO	0.635	0.879	1								
CO	0.592	0.905	0.715	1							
ROI	0.504	0.727	0.634	0.709	1						
PE	0.407	0.627	0.620	0.676	0.815	1					

PT	0.472	0.696	0.586	0.758	0.878	0.823	1				
PR	0.253	0.466	0.378	0.520	0.730	0.673	0.754	1			
SI	0.418	0.738	0.690	0.758	0.776	0.785	0.832	0.625	1		
ATU	0.292	0.566	0.402	0.645	0.731	0.746	0.861	0.680	0.785	1	
BI	0.470	0.663	0.593	0.686	0.828	0.728	0.920	0.658	0.801	0.823	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-113, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-113. UC11 Reliability Analysis

Cronbach-alpha values	
PU	0.723
PEOU	0.623
MO	0.509
CO	0.518
ROI	0.598
PE	0.301
PT	0.493
PR	0.210
SI	0.482
ATU	0.126
BI	0.568

Regression Analysis: Regression analysis is applied to estimating the relationships among QAM constructs, as seen in Table 3-114. For this demonstrator, there exists a right proportional relation between all construct pairs influencing each other in the same direction.

Table 3-114. UC11 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.40xPU + 3.31	0.403	~0	4.272
H2	CO-PU	Right	CO = 0.32xPU + 3.84	0.351	0.001	3.818
H3	PEoU-PU	Right	PEoU = 0.44xPU + 3.23	0.488	~0	5.074
H4	PU-ATU	Right	PU = 0.52xATU + 2.15	0.086	0.124	1.589
H5	PEoU-ATU	Right	PEoU = 0.63xATU + 1.83	0.321	0.001	3.569
H6	ATU-BI	Right	ATU = 0.73xBI + 1.80	0.678	~0	7.536
H7	ROI-BI	Right	ROI = 0.92xBI + 0.75	0.686	~0	7.673
H8	PE-BI	Right	PE = 0.87xBI + 0.72	0.530	~0	5.523
H9	SI-BI	Right	SI = 0.82xBI + 1.20	0.641	~0	6.941
H10	PT-BI	Right	PT = 1xBI + 0.13	0.846	~0	12.166
H11	PR-BI	Right	PR = 0.67xBI + 2	0.433	~0	4.544

3.11.7 Observed Limitations, Lessons Learnt and Best Practices

The following findings have been recognised to proliferate the impact of the demonstration outputs:

- The architecture of the MARVer tool offers a service-based framework for the verification of robotic systems. Currently, online minimum distance tracking, online people tracking, online motion tracking, security attacker, and simulation services have been implemented.
- With the addition of new services in the future, the scope of system verification can be expanded.
- By using the services simultaneously in a single test scenario, multiple verification tasks can be implemented simultaneously by saving time and effort spent on testing.
- However, the simultaneous execution of multiple verification tasks creates a computational load on the system. Therefore, powerful computers may be needed.
- Modelling the engines and making them work in real time was the most challenging part on the simulator side. All models were made into objects and worked in their own time concepts, and the data was published. Ensuring synchronous operation with ROS logic was the most challenging part (Observed Limitation).
- Making the communication between modules master-slave (Lesson Learnt).
- It allows the object to work independently of the system and does not have a state to stop or prevent the system, it just broadcasts. When the system needs it, it listens to that broadcast and takes action. (Best Practices)
- Mutation-based testing in Python has some limitations that can affect its effectiveness. One of the primary limitations is the execution time, as Python is an interpreted language and can be slower compared to compiled languages, resulting in longer execution times for mutation testing. Another limitation is the difficulty in generating meaningful mutations, which can be challenging to achieve. Additionally, mutation testing may not be suitable for programs that rely heavily on external resources like databases or network connections. Overall, while mutation-based testing can be a useful technique, it should be used in conjunction with other testing methods to ensure comprehensive test coverage (Observed Limitations).
- Learned how mutation-based tests can be performed for Python-based ROS systems (Lesson Learnt).
- Optimization and automation of industrial robotic systems using dynamic trajectory planning algorithms (Best Practices).
- End-to-end and holistic cyber-physical security has been realised by integrating HSM and secure IoT gateways and node/person authentication within the robotic system. The vulnerability analysis scheme based on analysing the reliability, robustness and unpredictability of true random number generators has played a crucial role in the design and implementation of the cryptographic hardware.

On the other hand, the qualitative assessment results support the adoption and potential uptake of the proposed set of innovations in UC11. Since the demonstrations are verified both in the laboratory (ESOGU's IFARLAB) and Otokar's factory environment, the level of technology acceptance is relatively high. As presented in Table 3-115, the user responses to QAM constructs are in the range of [5,06, 5,94]. Since the project outputs have already been deployed and started using in Otokar's daily operations, the observed ATU is high. Both for safety and security needs, the demonstrated project outputs present a satisfying level. UC11 is a very good example of the application of research outputs to industry and presents an exemplary case for further uptake in the automotive industry domain. One of the main reasons behind this positive attitude and high intention to use the project outputs is that the quantitative assessment results have presented very concrete outputs, such as a reduction in the total time of automotive body-in-white inspection, an increase in fault and cyber anomaly detection rate and improved usability and transferability of the project results to the other sister automotive companies of Otokar (e.g., Tofaş, Ford Otosan, Iveco, etc.).

Table 3-115 Mean and standard deviation of experts' responses to UC11

UC11	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,41	5,46	5,22	5,91	5,08	5,37	5,25	5,06	5,13	5,94	5,40
Std Dev	0,80	1,26	1,25	1,47	1,12	1,04	1,13	1,22	1,21	1,40	1,24

UC11 has resulted in 5 exploitable tools and devices already implemented in 10 toolchains and validated in 2 environments (ESOGU IFARLAB and Otokar Sakarya premises. The exploitable tools which are listed below can be commercialised in short or midterm after the project ends (i.e., up to 3 years):

1. Tailored Mutation-based Fault Injection Tool (IM-FIT)
2. Camera Fault Injection Tool (CamFITool)
3. Simulation-based Robot Verification Tool (SRVT)
4. Model-Aided Runtime Verification for Robotic Systems (MARVer)
5. Prigm Randomness Test Suites and the Prigm Hardware Security Modules and cyber-physical security platform (Figure 3-71)

Figure 3-71 PRIGM Hardware Security Module - a device equipped with the vulnerability analysis methods applied as a part of the V&V of cryptographic hardware

3.12 Use Case 13 – Industrial Drives for Motion Control (UC13)

Industrial drives for motion control systems are typically built with PLCs (Programmable Logic Controller) and power inverters for controlling electrical motors (see Figure 3-72). Products in this field cover a large variety of variable frequency inverters for electric motors supporting many different application scenarios such as factory automation and robotics. Modern motion control systems have strict requirements on precise and timely control. These requirements are accompanied by mandatory compliance with safety standards and security standards. The use case is built on a digital twin of such a system.

Figure 3-72 Industrial Drives for Motion Control in UC13

3.12.1 V&V Challenges

There is a variety of challenges such systems are confronted with:

- Industrial Drives for Motion Control systems have a tight integration of functionality, are safety-related, require high reliability, and need thorough verification. For such complex systems, it is hard to find a complete set of test cases that covers all relevant aspects – an approach that analyses the system and generates interesting test cases would be a plus.
  ➔ This topic was addressed by UML modelling with Enterprise Architect and MoMuT in cooperation with AIT and LLSG.
- Another challenge is the verification of analogue signals for motor models. Their theoretically infinite state-space together with non-linear behaviour makes it hardly possible to verify every scenario. A method for verifying signals such as motor revolutions per minute, phase voltages, and currents in a way that is easy to specify and execute is of interest and would be beneficial.
  ➔ The Real-Time Analog Monitoring Tool addresses this issue – the integration and application of this tool to the use case are supported by the partner AIT.
- A replacement of processor cores, which might for example be caused by supply chain issues, represents large verification efforts including hardware and software. The hardware must be adapted to the new processor and verification activities have to be taken. An early evaluation

of the interplay between the platform and the new processor supported by a digital twin would give valuable information before potential deep verification activities.

➔ This issue is addressed by the migration of the industrial drives for motion control digital twin to a RISC-V-based implementation and related functional tests.

3.12.2 Contributor

Partners contributing to the UC: SIEMENS, AIT, FRAUENHOFER IIS, LLSG

3.12.3 Contributors' Roles and Evaluation Scenario

Overview of individual partners' contributions within the evaluation scenarios can be found in Table 3-116.

Table 3-116 Overview of contribution to evaluation scenarios by UC13 partners

Evaluation Scenario	SIEMENS	AIT	FHG IIS	LLSG
VALU3S_WP1_Industrial_5	X	X		X
VALU3S_WP1_Industrial_6	X	X		X
VALU3S_WP1_Industrial_8	X	X		X
VALU3S_WP1_Industrial_9	X	X		X
VALU3S_WP1_Industrial_19	X		X	

- SIEMENS implemented a prototype of the digital twin of the demonstration vehicle with a RISC-V-based QEMU model and supports the partners for the application of methods and tools in the use case.
- AIT contributes by working with LLSG to generate test cases with MoMuT from the model that LLSG created in UML. Generated test sequences aim to provide optimized tests for the detection of mutations based on a test model. Furthermore, AIT works on signal monitoring techniques for the digital twin motion control system.
- LLSG contributes by creating UML models for UC13 (state machines and VVML flow diagrams) with Enterprise Architect (EA) that are associated with the abstract parts of the use case system. LLSG has also been working with AIT on the MoMuT extension to EA, visualizing differences in models during test and verification, identification of the valuable model structure for test cases, and analysing the potential of using model-based mutations in the planned verification process.
- FRAUENHOFER IIS contributes by supporting system-level modelling and simulation (QEMU, SystemC/TLM) for the digital twin, supporting the creation of a distributable motor model (SystemC-AMS), and the application of suitable RISC-V implementations / simulators.

Tools

There are several tools (see Figure 3-73) that support the above-mentioned methods such as RTAMT, Enterprise Architect and MoMuT for the methods *Behaviour-Driven Model Development*, *Test-Driven*

Model Review, Model-Based Mutation Testing, Model-Based Robustness Testing, and *Test Oracle Observation at Runtime.*

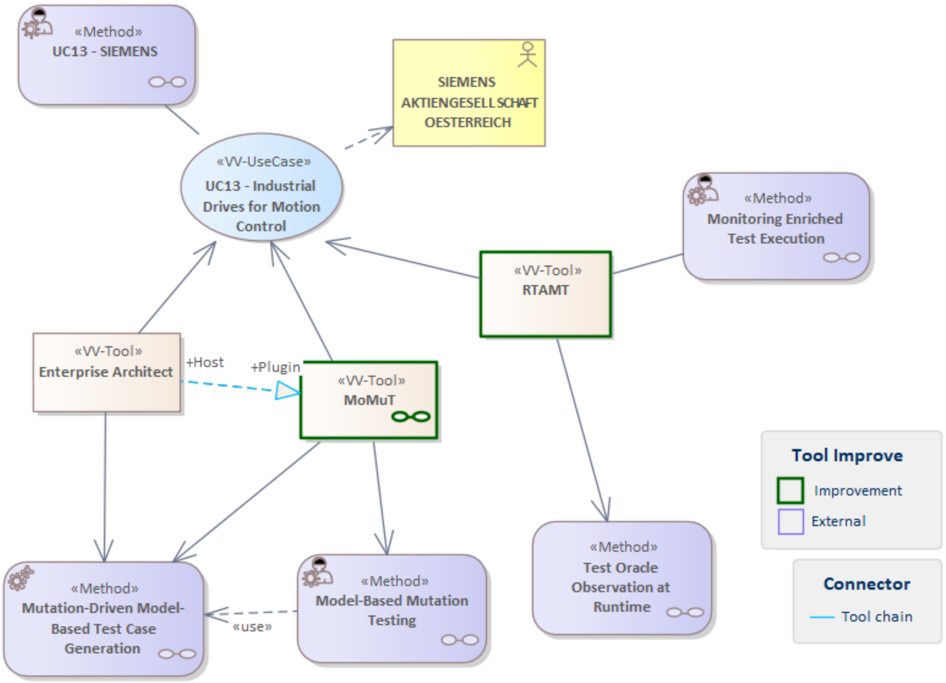

«Method»
UC13 - SIEMENS
SIEMENS AKTIENGESELLSCHAFT OESTERREICH
«VV-UseCase»
UC13 - Industrial Drives for Motion Control
«Method»
Monitoring Enriched Test Execution
«VV-Tool»
RTAMT
«VV-Tool»
Enterprise Architect
+Host
+Plugin
«VV-Tool»
MoMuT
Tool Improve
Improvement
External
Connector
Tool chain
«Method»
Mutation-Driven Model-Based Test Case Generation
«use»
«Method»
Model-Based Mutation Testing
«Method»
Test Oracle Observation at Runtime

Figure 3-73 Tools and Method Overview for UC13

Additionally, *Processor Integration Verification within a System-Level Digital Twin of Legacy Systems* verification applies several standard tools such as SystemC, AMESim, QEMU, and FreeRTOS.

3.12.4 Demonstration

Table 3-117 presents the list of demonstrators connected to this use case.

Table 3-117 Overview of demonstration prepared by UC13 partners.

Item #	Demonstration Name	Description/Purpose	Type	Responsible
1	Real-Time Analogue Signal Monitoring (RTAMT) for a Digital Twin for Motion Control	Addresses the extended verification and validation capabilities for the continuous domain of the industrial drives for motion control digital twin in UC13. Correct and faulty simulation data are checked with the RTAMT tool, which shows requirement violations graphically with fault explanation.	Lead Demonstrator. Video demo showing the application of signal monitoring in combination with the UC13 digital twin.	AIT, SIEMENS

Item #	Demonstration Name	Description/Purpose	Type	Responsible
2	Model-Based Mutation Test (MoMuT) Modeling with Enterprise Architect (EA) for motor control	Show the application of MoMuT for the development of motion control for industrial drives system.	Complementary Demonstrator. Video showing MoMuT application for the use case with test sequence generation.	AIT, LLSG, SIEMENS
3	Demo Processor Integration verification enabled by a digital twin.	Show results of the ported system from legacy to RISC-V-based digital twin.	Complementary Demonstrator. Slideshow for the ported digital twin for processor integration verification using various tools and simulators such as QEMU, SystemC, FreeRTOS and AMESim. The digital twin builds the base for the Signal Monitoring Demo (#1).	FRAUNHOFER IIS, SIEMENS
4	The VVML workflow for UC13 and related methods/tools are presented.	Provide an overview of the applied methods/tools in UC13.	Complementary Demonstrator. PowerPoint/Poster	SIEMENS, AIT, FRAUNHOFER IIS, LLSG

3.12.5 Quantitative Results

List of evaluation criteria of the V&V process

1. Eval_VV_2 – Coverage of test set (AIT, FRAUENHOFER IIS, LLSG, SIEMENS)
   - Measured quantity/artefacts:
     Measuring how much of software/hardware test coverage items (e.g., lines of code, branches, faults, and attacks depending on selected test design technique) has been covered by a test set (set of test cases, also known as test suite). Increased coverage means increased trust in the analyzed system.
     - Model-Based Mutation Testing (with documented test sequence diagram)
       The number of test coverage items covered by the executed test cases. This metric focuses on new tests derived from the application (execution) of Model-Based Testing techniques.
       Number of (reachable) states covered = 100%
       This is the number of state transitions (MoMuT is applied on UML state machines) covered by executed test cases. MoMuT can generate test cases that will cover 100% of the reachable states.

List of evaluation criteria for SCP:

1. Eval_SCP_2 – Number of safety/security requirement violations (AIT, FRAUNHOFER IIS, LLSG, SIEMENS)
    - Measured quantity/artefacts:

      Measuring the number of violated SCP attributes/requirements/properties that have been checked by runtime monitors, software testing, and/or formal verification is useful for comparing the effect of changes to requirements engineering, development, and verification processes. It is important to remember that the violation of security requirements can negatively impact a system's ability to uphold its safety requirements. For example, a security violation in an autonomous vehicle could result in it not recognizing the vehicle in front of it which may cause a crash and thus result in the violation of related safety requirements.
        - Fault Localization for Specification-based real-time monitoring (RTAMT)

          Support for coverage of safety/security requirement violations = 46%

          Injected faults for simulation data and requirements violations were detected with RTAMT for the selected test cases. Signal Monitoring (RTAMT) was initially identified to support 9 out of 21 safety/security requirements for V&V activities in UC13 (43%). The tool allows its application, especially for signal analysis in the continuous domain (motor speed, phase voltages, currents, position data). Note that there are other ways for implementing verification measures for the analysis of signals (such as an implementation in C++), however, they do not offer an elegant way of defining formal specifications and graphically annotated guidance for specification violations. The number shows the potential application of RTAMT for the detection of safety and security violations (triggered by fault injections during simulation) – With the application of RTMAT in the use case, one additional safety/security violation was discovered – which led to a new requirement for co-simulation stability. Thus, the application of RTAMT with fault-explanation increased support for coverage of safety/security violations from 43% to 46%.

          Additionally, a higher verification quality was reached due to RTAMT's capability for fault-explanation, which supports verification engineers to pinpoint faults.
        - Model-Based Mutation Testing (with documented test sequence diagram)

          MoMuT supports the verification of around 48% of safety and security requirements. Assessed was the verification potential of the MoMuT method for such requirement types via UML modelling in EA, MoMuT and expert estimations. MoMuT can generate test cases for around 48% of safety and security requirements. 100% test coverage is not feasible with this method, because there are requirements (e.g., security requirements) that are not in the application domain of MoMuT.
        - Processor Integration Verification within a System-Level Digital Twin of Legacy Systems

          Support for coverage of safety/security requirement violations = 100%

          Injected faults for simulation data where requirement violations can be detected after core replacement in the digital twin. In this test case, potential safety/security violations can be discovered as well as with the legacy core digital twin.

2. Eval_SCP_10 – Software fault tolerance robustness (AIT, FRAUNHOFER IIS, LLSG, SIEMENS)
   - Measured quantity/artefacts: Number of detected fault types (assignment, algorithm, timing etc.), which is the majority of faults and attacks detected among all fault injections applied. Fault-injection techniques are planned to analyse the fault tolerance robustness. Measurement will be based on defined safety-assignments, security-assignments, functional-assignment faults, and timing faults.
     - Processor Integration Verification within a System-Level Digital Twin of Legacy Systems

       Fault tolerance robustness = denied or wrong services/number of injected faults = 0

       After integrating a new core in the digital twin, the injected faults result in the same fault tolerance number as with the legacy core. The digital twin has just minimal functional deviations after core replacement, thus, the injected faults result in the same fault tolerance robustness number.
     - Model-Based Mutation Testing (with documented test sequence diagram)

       The number of possible faults is stated by "weak kills". MoMuT generates a large set of faults (mutants), which can help to detect unexpected behaviour and decide implementation measures for fault-tolerance. Fault-injection in MoMut is still in a very experimental state.

       The number of possible faults is stated by "weak kills" (i.e., system stops operation) After fixing bugs after a manual iteration with MoMuT the number of weak kills should decrease, which helps to improve fault-tolerance robustness.

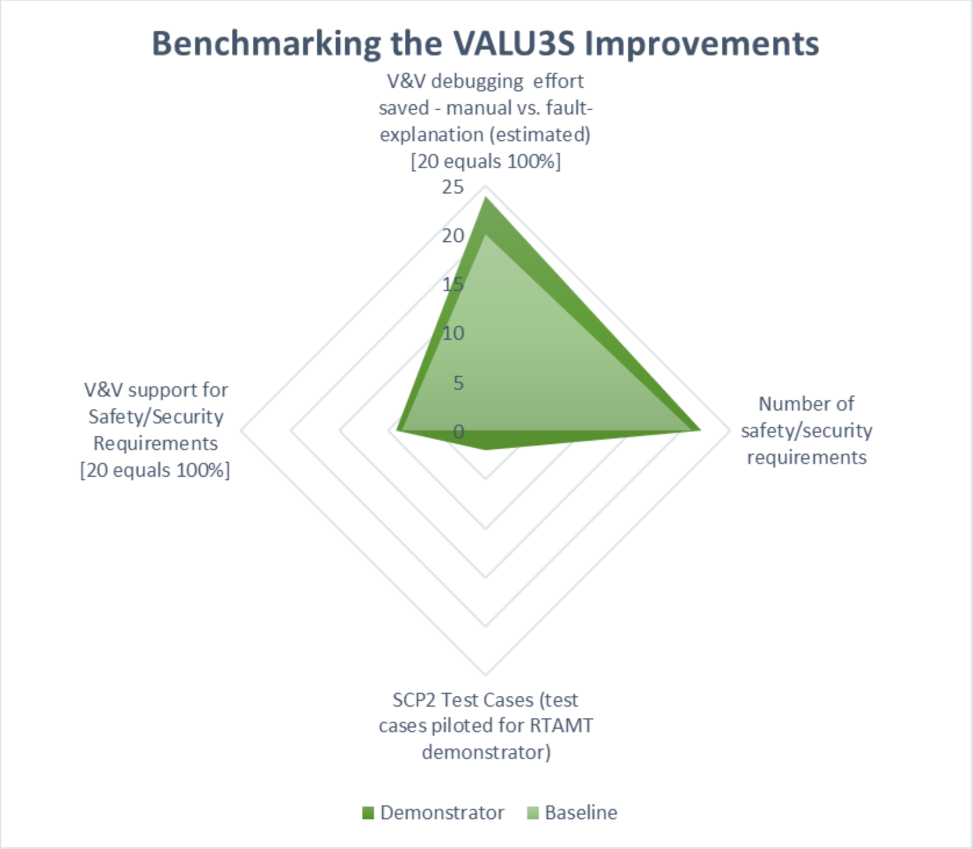


Figure 3-74: Radar plot for improvements with the lead demonstrator in UC13

Improvements with the demonstrator regarding real-time analogue signal monitoring are depicted in Figure 3-74. V&V support for safety/security requirements was increased from 43% to 46% - the tools' application can lead to new requirements. The support for coverage of safety/security requirement violations was increased by about 5%. The application of signal monitors can also increase the overall verification quality by revealing design flaws in the Design-under-Test (model and simulation setup). Furthermore, signal monitoring also enables support for system optimization (tighter/looser specification for signals) due to fault-explanation.

Debugging efforts for this application were reduced by 20% (expert estimation). This is reasoned by a faster verification process (support for bug/fault analysis) enabled by the fault-explanation method (visualization of specification violations) and reduced verification environment development efforts due to automatically generated signal monitors for verification. Additionally, there are potential licensing cost savings due to the open-source license model of RTAMT.

3.12.6 Qualitative Results

Demonstrator 1: Digital Twin for Motion Control Signal Analysis with Real-Time Analogue Signal Monitoring (RTAMT)

Participants Profile: QAM is applied to 34 subjects (33 Males, 1 female) aged in the range of 25-51. The education level is relatively high as the subject pool is composed of 1 Post-Doc or higher degree and 1 PhD researcher and 32 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "research engineers, system engineers, software/hardware engineers, designers, experts, Q&A etc." having experience in the fields of "NIDS, computer vision, NLP, embedded systems, cyber-physical security, hardware & software design, automotive, railway, ASIC etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-112. The results show that all constructs are correlated with each other.

Table 3-118. UC13 – Demonstrator 1 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.370	1									
MO	0.434	0.662	1								
CO	0.323	0.603	0.857	1							
ROI	0.450	0.607	0.895	0.952	1						
PE	0.429	0.607	0.871	0.926	0.965	1					
PT	0.382	0.603	0.884	0.945	0.963	0.947	1				
PR	0.423	0.563	0.821	0.844	0.902	0.883	0.962	1			
SI	0.500	0.573	0.839	0.873	0.930	0.923	0.938	0.940	1		
ATU	0.451	0.624	0.826	0.869	0.917	0.901	0.911	0.879	0.936	1	
BI	0.390	0.621	0.871	0.893	0.934	0.968	0.904	0.836	0.863	0.897	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-119, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-119. UC13 – Demonstrator 1 Reliability Analysis

Cronbach-alpha values	
PU	0.413
PEOU	0.612
MO	0.143
CO	0.165
ROI	0.372
PE	0.682
PT	0.535
PR	0.284
SI	0.523
ATU	0.307
BI	0.565

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-120. For this demonstrator, there exists a right proportional relation between all construct pairs influencing each other in the same direction.

Table 3-120. UC13 – Demonstrator 1 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.112xBI + 5.154	0.188	0.010	2.724
H2	CO-PU	Right	CO = 0.095xBI + 5.23	0.105	0.062	1.934
H3	PEoU-PU	Right	PEoU = 0.114xBI + 5.04	0.137	0.031	2.254
H4	PU-ATU	Right	PU = 0.112xBI – 6.87	0.203	0.007	2.856
H5	PEoU-ATU	Right	PEoU = 0.095xBI - 0.299	0.390	~0	4.521
H6	ATU-BI	Right	ATU = 0.975xBI + 0.486	0.804	~0	11.462
H7	ROI-BI	Right	ROI = 0.984xBI + 0.499	0.872	~0	14.771
H8	PE-BI	Right	PE = 1.011xBI + 0.289	0.936	~0	21.724
H9	SI-BI	Right	SI = 0.989xBI + 0.749	0.744	~0	9.646
H10	PT-BI	Right	PT = 1.055xBI + 0.542	0.818	~0	11.997
H11	PR-BI	Right	PR = 0.959xBI + 0.833	0.698	~0	8.604

Demonstrator 2: Model-Based Mutation Test (MoMuT) Modeling with Enterprise Architect (EA) and MoMuT for a motor control cycle

Participants Profile: QAM is applied to 10 subjects (9 males, 1 female) aged in the range of 26-31. The education level is relatively high as the subject pool is composed of 2 PhD researchers and 8 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as “research engineers, system engineers, software engineers, Q&A, etc.” having experience in the fields of “fault injection, cyber security, embedded systems, 3d visualisation, etc.”.

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-121. The results show that the majority of the constructs are correlated with each other except PT, SI and PR.

Table 3-121. UC13 – Demonstrator 2 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.598	1									
MO	0.723	0.879	1								
CO	0.738	0.855	0.912	1							
ROI	0.719	0.636	0.778	0.777	1						
PE	0.827	0.566	0.741	0.819	0.913	1					
PT	0.001	0.396	0.270	0.321	-0.263	-0.129	1				
PR	0.511	-0.143	-0.010	0.064	0.495	0.558	-0.673	1			
SI	0.443	0.776	0.859	0.666	0.658	0.560	0.140	-0.051	1		
ATU	0.660	0.004	0.183	0.199	0.590	0.575	-0.571	0.806	-0.066	1	
BI	0.595	0.054	0.203	0.312	0.694	0.667	-0.682	0.842	-0.009	0.901	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-122, the questions asked to subjects are sufficiently reliable as understood from subject responses, except the questions and answers related to the constructs PT, PR and SI.

Table 3-122. UC13 – Demonstrator 2 Reliability Analysis

Cronbach-alpha values	
PU	0.269
PEOU	0.881
MO	0.540
CO	0.718
ROI	0.617
PE	0.436
PT	-0.178
PR	-2.775
SI	0.059
ATU	0.155
BI	0.409

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-123. For this demonstrator, there exists a right proportional relation between SI-BI and PT-BI construct pairs influencing each other in the same direction. While the other pairs are inversely proportional to each other.

Table 3-123. UC13 – Demonstrator 2 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Inverse	MO = -0.74xPU + 1.97	0.523	0.018	2.960
H2	CO-PU	Inverse	CO = -0.71xPU + 2.31	0.545	0.015	3.093
H3	PEoU-PU	Inverse	PEoU = -0.39xPU + 3.96	0.358	0.068	2.110
H4	PU-ATU	Inverse	PU = -0.92xATU + 0.44	0.436	0.038	2.485
H5	PEoU-ATU	Inverse	PEoU = -0.003xATU + 6.08	0.000	0.992	0.010
H6	ATU-BI	Inverse	ATU = -0.89xBI + 0.77	0.812	0.000	5.872
H7	ROI-BI	Inverse	ROI = -1.16xBI – 0.42	0.482	0.026	2.726
H8	PE-BI	Inverse	PE = -0.93xBI + 0.70	0.445	0.035	2.532
H9	SI-BI	Right	SI = 0.01xBI + 6.24	0.000	0.980	-0.026
H10	PT-BI	Right	PT = 1.55xBI + 14.76	0.466	0.030	-2.640
H11	PR-BI	Inverse	PR = -0.97xBI + 0.29	0.709	0.002	4.419

3.12.7 Observed Limitations, Lessons Learnt and Best Practices

This use case showed that it is very valuable to have fault-explanation capabilities in a verification tool. Unexpected errors in the analysed data occurred, which were not obvious to detect. Violations are reported graphically by denoting areas having faulty signals (fault-explanation), which enables engineers to easily spot a range of faults within the data. The tool tells the verification engineer to look at the right place and immediately confirm where the problem is. Without a fault explanation, the verification engineer debugging the system might easily be misled by the identified voltage spike by suspecting only a single error, which typically would result in fixing the bug and re-running the verification suites. This extra verification spin can potentially be avoided by applying signal verification based on formal specification with fault-explanation, as shown in this small use case example.

A promising method for generating additional test sequences to cover potential implementation faults was the method Model-Based Mutation Testing (with documented test sequence diagram). The state of the tool application for the use case is still experimental. Test sequences could be generated for a partial test model of the industrial drives for the motion control use case, and state coverage could be observed. A limitation is given by the interface between the simulation model and the generated test sequences from the test models. Interfacing an already existing digital twin to test sequences requires a lot of effort – future applications of this method must consider preparing test drivers for a simulation model and interfaces matching the test activities generated by MoMuT. Furthermore, keeping a test model and a simulation model (digital twin) synchronized is a challenging task and currently requires manual work for such a use case setup.

As presented in Table 3-124 and Table 3-125, the overall technology acceptance level is high for both demonstrations. Especially for the second demonstration where the model-based mutation test modelling is applied for the more effective operation of the motor control cycle, the experts' opinions are very high. Note that the experimental state the technology is in, might not have been brought successfully to the full awareness of the questionnaire participants. The main reason behind the positive attitude towards using the proposed solution stack can be multifaceted. However, the forthcoming reasons can be the direct implementation of the solution in industrial settings, the high reputation of the proposer organisations, and the potential cascade effects of the proposed technique related to its transferability to other cyber-physical settings. The responses to QAM factors like PEOU, SI and ATU are relatively lower (<5.00) as compared to other factors in UC13-Demonstration-1. If potential problems originating from user responses (like less awareness or concentration during the questionnaire) are ignored, one of the reasons can be the scepticism about the digital twin notion. It is noteworthy that even many experts in the domain may still not have sufficient technical information about the digital twin, or they may think that the digital twin term has become a buzz word.

Table 3-124 Mean and standard deviation of experts' responses to UC13 - Demonstration 1

UC13-Demonstrator-1	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,68	4,83	5,76	5,40	5,83	5,63	5,09	5,24	4,88	4,86	5,68
Std Dev	0,82	1,20	0,81	0,70	0,93	0,86	1,11	1,32	1,38	1,24	1,20

Table 3-125 Mean and standard deviation of experts' responses to UC13 - Demonstration 2

UC13-Demonstrator-2	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
mean	6,15	5,54	5,66	5,37	5,73	5,93	5,52	6,07	5,85	6,10	6,20
Std Dev	0,52	0,78	0,50	0,53	0,43	0,51	0,31	0,62	0,92	0,72	0,71

3.13 Use Case 14 – CardioWheel (UC14)

As a key element of the driving task and road safety, it is of extreme importance to guarantee that the driver (or teleoperator) of a vehicle is in a cognitive state complacent with the skill, readiness, and responsibility such task demands. Moreover, in some specific contexts such as professional fleets and share-driving vehicles, it is essential to verify the identity of the driver, ensuring correct liability in case of accidents, vehicle misuse, and digitalization of the process.

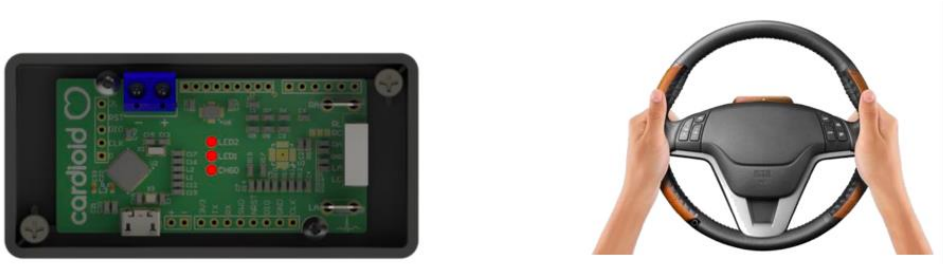


Figure 3-75 CardioWheel embedded system and steering wheel cover.

CardioWheel is an Advanced Driver Assistance System that acquires electrocardiogram (ECG) from the driver's hands to continuously detect drowsiness, cardiac health problems, and perform biometric identity recognition [17], [18]. It (see Figure 3-75) is composed of an analogue front-end, which measures the ECG signal, and an embedded processing unit that performs signal processing and sends information to the CardioGW using Bluetooth Low Energy (BLE). ECG measurement uses two dry electrodes, seamlessly integrated into a steering wheel cover, using conductive materials. This system is protected under the patent WO2013109154A1 [31].

To enable real-time ECG-based driver monitoring, the CardioWheel is inserted into an ecosystem consisting of the device itself, a gateway (CardioGW) that communicates with CardioWheel via BLE to collect the ECG data, and a cloud service where machine learning models are trained before being installed in the CardioGW, see Figure 3-76. The system is based on pre-trained models that were fitted using larger curated datasets, but that are refined using data collected from drivers using the system. GDPR compliance is guaranteed by the design of the system.

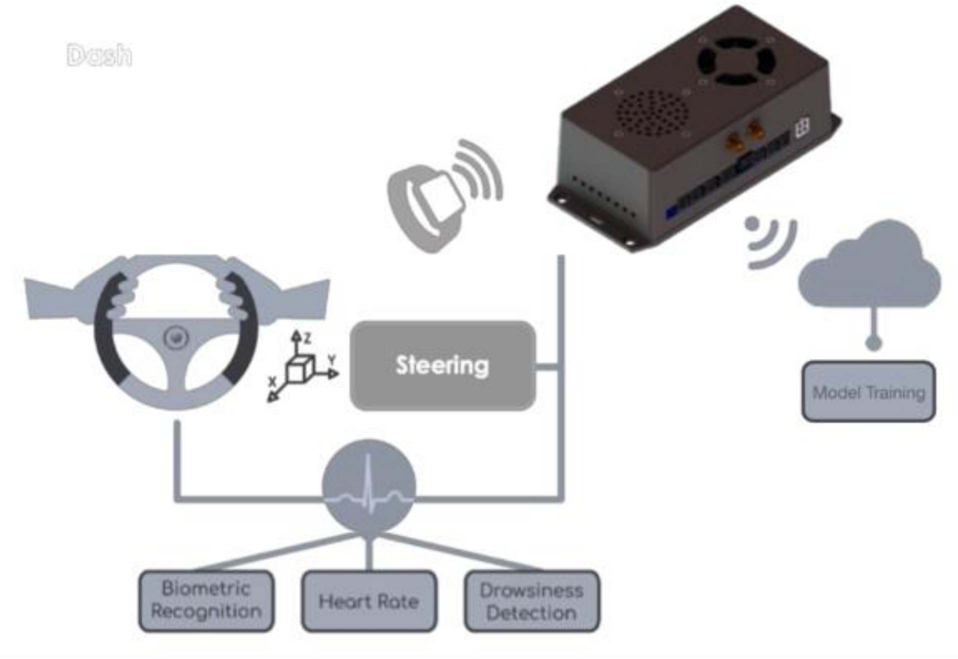


Figure 3-76 Diagram of CardioWheel ecosystem.

3.13.1 V&V Challenges

The CardidoWheel system requires processing capability to acquire ECG signal at 1kHz and perform all signal conditioning and feature extraction tasks in real-time. Furthermore, the biometric nature of data, acquired and processed, requires that the system implements adequate methods to ensure data privacy. And finally, driver-state recognition models used in the system should be accurate and robust against data noise, so that users can trust their predictions.

With this in mind, three main V&V challenges are defined for this use case:

- V&V workflows are needed to ensure a sound firmware architecture capable of handling all required tasks.
- Data Privacy must be guaranteed by robust cryptographic methods, and communication channels must be thoroughly tested.
- Develop an objective metric for drowsiness to produce labelled data on which driver state models are robustly validated.

3.13.2 Contributors

Partners contributing to the UC: CARDIOID, ISEP, COIMBRA, VTI

3.13.3 Contributors’ Roles & Evaluation Scenario

Evaluation scenarios defined for this use case are listed below; the assignments of the contribution of all UC partners can be found in Table 3-126:

- VALU3S_WP1_Automotive_15 - Security and integrity of transmission model Evaluate cases of man-in-the-middle attacks and data sniffing that could compromise driver privacy and/or corrupt model training.
- VALU3S_WP1_Automotive_16 - Verification of embedded real-time properties Identify safety properties related to timing that can be observed and verified upon runtime of the system, evaluating a specification of CardioWheel's software under the MARS domain specification language.
- VALU3S_WP1_Healthcare_10 – Driver state recognition accuracy under uncertainty Determination of driver state recognition models' accuracy under various cognitive states and for different individuals.
- VALU3S_WP1_Crossdomain_1 – Driver state recognition reliability under uncooperative environments
  Evaluation of model outcomes validity when drivers are not cooperative with the system (e.g., do not place both hands on the steering wheel).
- VALU3S_WP1_Crossdomain_2 – Verification of biometric models security Evaluate the cryptographic security of implemented biometric models.

Table 3-126 Overview of contribution to evaluation scenarios by UC14 partners

Evaluation Scenario	CARDIOID	ISEP	COIMBRA	VTI
VALU3S_WP1_Automotive_15	X	X	X	
VALU3S_WP1_Automotive_16	X	X	X	
VALU3S_WP1_Healthcare_10	X	X	X	X
VALU3S_WP1_Crossdomain_1	X	X		
VALU3S_WP1_Crossdomain_2	X			

Individual UC14 partners are contributing to the evaluation scenarios as follows (see also Figure 3-77):

CARDIOID is involved in all evaluation scenarios as the use case provider. Specifically, CARDIOID prepared the Hardware in the Loop (HiL) where VALU3S_WP1_ Automotive_15, VALU3S_WP1_ Automotive_16 and VALU3S_WP1_Crossdomain_1 are implemented, with modules that inject monitoring code and simulated faults, producing reports that are evaluated by the frameworks provided by ISEP and COIMBRA. In VALU3S_WP1_Healthcare_10 CARDIOID defined drowsiness metrics alternative to KSS, in cooperation with VTI, to have a more reliable set of data on which to evaluate its state recognition models. For VALU3S_WP1_Crossdomain_2, encryption techniques for biometric data obfuscation were studied to increase the privacy-related properties of biometric models.

ISEP is contributing to the evaluation scenarios VALU3S_WP1_Automotive_15, VALU3S_WP1_Automotive_16 and VALU3S_WP1_Crossdomain_1 by providing and adapting, their method "Safe Generation and Instrumentation of Runtime Verification Architectures", which uses tools such as MARS and VAITP. This method defines monitors that can assert system states at runtime, with those assertions being based on a set of requirements that the system must fulfil. ISEP is also contributing to VALU3S_WP1_Healthcare_10 by designing machine learning model evaluation methods capable of measuring the state recognition models' robustness against faulty data (e.g., missing or falsely detected heartbeats that may compromise the HRV analysis).

COIMBRA is contributing to VALU3S_WP1_Automotive_15 and VALU3S_WP1_Automotive_16 by providing tools for fault injection, measuring the impacts of those faults in the real-time properties of the system, as well as potential exploitations that such faults can uncover. COIMBRA also developed a "safety net" system for the driver monitoring models. This system estimates the probability that a given model output is a false negative, allowing the system to suppress potential FN outputs and so increase the safety of the system.

VTI has developed a simulation setup to support V&V activities in this use case. Two simulators will be used to collect data from research participants in different driving and lighting conditions as a contribution to VALU3S_WP1_Crossdomain_1. CardioWheel has been integrated into the simulator to collect relevant data for further V&V of the CardioID's systems.is integrating the CardioWheel with two of their driving simulators as a contribution to VALU3S_WP1_Crossdomain_1. This integration also involves the adaption of the simulation scenario to have an additional drowsiness metric based on reaction times. VTI also coordinates data collection sessions that are the basis of evaluating model performance.

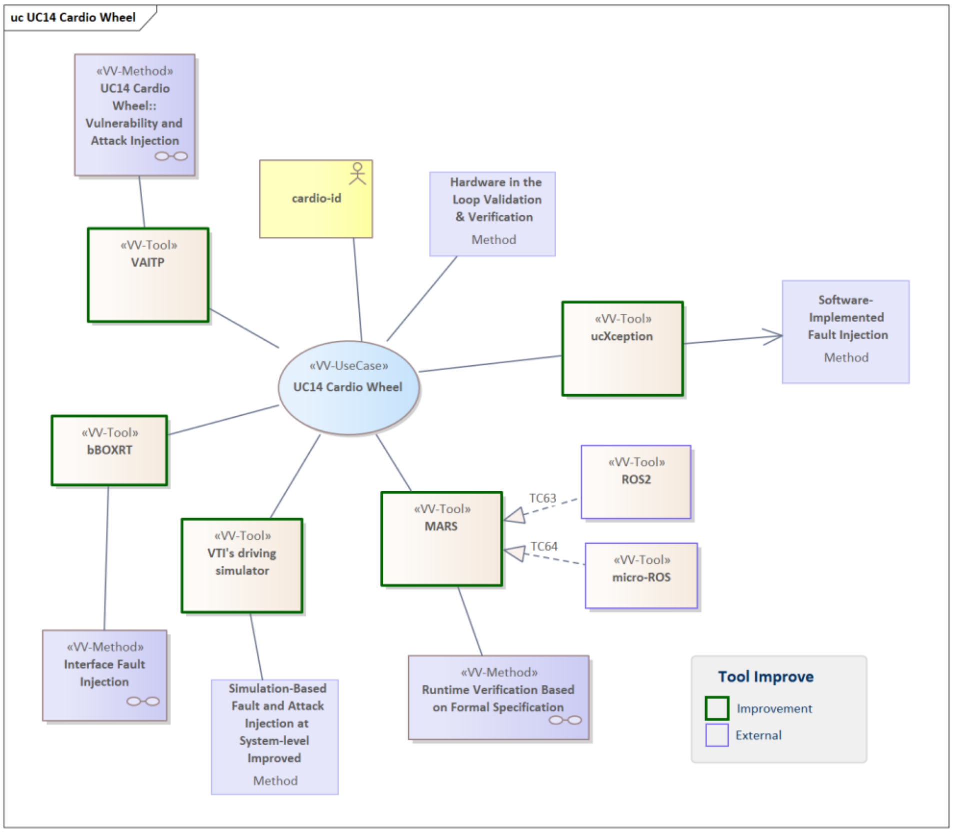

uc UC14 Cardio Wheel
«VV-Method» UC14 Cardio Wheel:: Vulnerability and Attack Injection
cardio-id
Hardware in the Loop Validation & Verification Method
«VV-Tool» VAITP
«VV-Tool» ucXception
Software-Implemented Fault Injection Method
«VV-UseCase» UC14 Cardio Wheel
«VV-Tool» bBOXRT
«VV-Tool» ROS2
TC63
«VV-Tool» MARS
TC64
«VV-Tool» VTI's driving simulator
«VV-Tool» micro-ROS
«VV-Method» Interface Fault Injection
Simulation-Based Fault and Attack Injection at System-level Improved Method
«VV-Method» Runtime Verification Based on Formal Specification
Tool Improve
Improvement
External

Figure 3-77 Methods and Tools overview for UC14.

3.13.4 Demonstration

Lists the proposed demonstrators for UC14 can be seen in Table 3-127.

Table 3-127 Overview of demonstration prepared by UC14 partners.

Item #	Demo Name	Description/Purpose	Type	Responsible
2	Hardware in-the-Loop Validation Station	Show how developed methods and workflows can be combined in an automated station that validates devices at the end of production pipeline. In this demonstrator, Runtime validation based on formal requirements and software implemented fault injection are implemented on an automated station that cuts both time and costs in validation/verification processes, by being faster than human based actions, and by decreasing the expertise needed by operators of this equipment.	Lead Demonstrator Stand where HiL is installed and running, supported by PowerPoint presentation with further details.	CARDIOID/ ISEP/ COIMBRA
3	Instrumented Driving Simulator for drowsiness data generation	Two VTI simulators were equipped with CardioWheels to conduct drowsy driver data collection. On human-factor-based ML systems, data quality and quantity are of the utmost importance to guarantee reliable predictions. This demonstrator reports the data acquisition protocol and simulator setup with a video and a photo collection of the steps involved.	Lead and Complementary demonstrator. Videos/PowerPoint Presentation	CARDIOID/ VTI/ ISEP

<u>HiL test bed demo (CARDIOID, ISEP, COIMBRA)</u>

Show how developed methods and workflows can be combined in an automated station that validates devices at the end of the production pipeline. In this demonstrator, runtime validation based on formal requirements and software-implemented fault injection is implemented on an automated station that cuts both time and costs in validation/verification processes, by being faster than human-based actions, and by decreasing the expertise needed by operators of this equipment.

<u>Instrumented Driving Simulator for drowsiness data generation (CARDIOID, VTI, ISEP)</u>

Two VTI simulators were equipped with the CardioWheel to conduct drowsy driver data collection. On human-factor-based ML systems, data quality and quantity are of the utmost importance to guarantee reliable predictions. This demonstrator reports the data acquisition protocol and simulator setup with a video and a photo collection of the steps involved.

3.13.5 Quantitative Results

The development of this use case was measured through SCP Evaluation and V&V Evaluation criteria. Moreover, to provide a unified format to report the overall achievements on the different evaluation dimensions, the following listings both describe the evaluation procedure for each criterion and the results expressed in terms associated with the artefacts that support them, as well as a five-point evaluation measuring the improvement level. In this regard, all baseline measures are set to 1 (no improvement) and the results are evaluated up to 5 (very good improvement)

SCP Evaluation criteria used in this demonstrator are:

1. Eval_SCP_1 – Error Coverage – by injecting faults and attacks to evaluate CardioWheel's system capacity to cope with incorrect, invalid or untrusted data, and this error coverage are used to evaluate the system responses to those tests.

   Baseline: No fault/injection performed at baseline: 0.

   Current: Structured error space probing, with both faults injected, measuring sufficient fault tolerance for the expected faults in current operational use (inside road vehicles). But the necessity of increased system redundancy for applications in other critical environments (space and similar). Other controlled errors are now systematically studied, such as lack of user contact (Lead-OFF) which is detected 100% of the time for Lead-Off durations larger than 200 ms.

   5 - very good improvement – Formal fault injection methods are employed to properly evaluate the system's fault tolerance and our response to an error space that covers the expected operation mode of the system.
2. Eval_SCP_2 – Number of Safety/Security Requirement Violations - CARDIOID uses this criterion to evaluate results from runtime verification tests. Part of these tests deal with altered data and attempts to inject attacks that could, in an unsafe system, expose personal data from users. Because of that, good evaluation scores on these metrics point to a safe system privacy-wise.

   Baseline: No systematic evaluation of safety/security requirement compliance: 0.

   Current: 25 tests were performed where Man-in-the-Middle attacks tried to recover ECG data being sent over the custom encrypted transport based on TLS 1.3, other tests were performed to define run-time monitors that validate timing-related properties of the CardioWheel. These tests report 0% of requirement violations.

   4 – good improvement – Improved requirement definition and the number and completeness of tests to measure requirement violations. On top of a better automatization of the process, further specification and formalization of requirement definitions can be achieved.
3. Eval_SCP_4 – Metrics to evaluate AI/ML algorithms – These metrics are used to evaluate results from Machine learning validation tests. As a baseline state, CARDIOID prefers the utilization of MMC for the state recognition model performance evaluation, given the class imbalance that these problems usually present, while TAR, FAR, FRR and EER are used to quantify the performance of biometric models. Independently of CARDIOID's baseline preference, the project is used to validate these choices.

   Baseline: Average 80% Accuracy for known drivers, 20% performance reduction for some unknown drivers (models trained using KSS information only).

Current: Over 90% accuracy for known drivers and over 80% for unknown ones. Adversarial training reduced FN rate to less than 5% (models trained using reaction time information).

5 – very good improvement – Defined new objective drowsiness metric that allows more robust ML/AI evaluation. Employed adversarial training method that increases model robustness to faulty data. Employed False Negative monitor that reduces false negative outputs by up to 70%.

4. Eval_SCP_9 – Randomness and cryptographic algorithm strength - the different metrics of this evaluation criteria are used to measure the adequacy of cryptographic algorithms implemented in this system.

   Baseline: No formal testing on cryptographic strength at baseline: 0. Current: 25 tests show 100% of message exchanges were encrypted and authenticated using the custom transport developed for UC14 based on TLS 1.3 and DDS-XRCE, making transmitted data inaccessible by third parties. 25 tests performed show 0% leakage of data to unauthorized 3rd-parties.

   3 – some improvement – validated system's capability to send data through an encrypted communication channel. However, further development is needed to implement a similar cyber-security level on a Bluetooth-based channel instead of Wi-Fi + TLS 1.3.

Furthermore, V&V evaluation criteria are used to measure the impact of improved V&V workflows, namely:

1. Eval_VV_2 – Coverage of test set – The design of workflows related to this use case improves the completeness of the test set. At baseline, tests are manually conducted and focus on isolated elements of the system. As already envisioned both the total number of tests and also the proportion of requirements systematically verified are greatly improved.

   Baseline: 0.

   Current: 10 tests cover 8/14 defined requirements.

   4 – good improvement – The organization and coverage of the test set were greatly improved, with further test cases and requirements needing coverage for a fully completed test set.

2. Eval_VV_8 – Effort needed for test – Person-hours needed for system validation and verification is measured for the baseline test suit and compared to those of improved V&V workflows, it is expected that the systematisation of this procedure has a positive impact on its efficiency.

   Baseline: 15min/device.

   Current: 3min/device with the automated system, which allows parallelization of testing pipelines for multiple devices at once.

   5 – very good improvement - The automation of the validation procedure greatly reduced the time needed per test and released qualified engineers from its supervision, reducing the cost associated with this step and allowing its deployment on production lines with less qualified personnel.

3. Eval_VV_11 - Randomness and Security Assessment Process Performance - Improvement of cyber-security-related tests efficiency is measured using this criterion.

Baseline: No formal cryptographic tests implemented at baseline: 0.

Current: 25 tests show that 100% of the message exchanges in the context of extremely resource-constrained devices were encrypted and authenticated using the custom transport developed for UC14 based on TLS 1.3 and DDS-XRCE, making transmitted data inaccessible by third parties. Initial results from the Tests were performed using resource-constrained devices 25 performed tests performed show 0% leakage of data to unauthorized 3rd-parties with an overhead of around 20% more time when compared to the non-encrypted communication when considering single-value transmissions.

3 – small improvement – the performance impact of the applied cryptographic protocol was analysed in the context of extremely resource-constrained devices, but no improvement on the randomness of the algorithm itself was achieved.

Figure 3-78 shows the distribution of improvement over all selected evaluation criteria, indicating the excellent improvements in Error coverage, ML/AI evaluation and test effort reduction as well as the areas on which special focus will be given following what was learned from the project, namely the application of solid cryptographic methods and improving its evaluation routines.

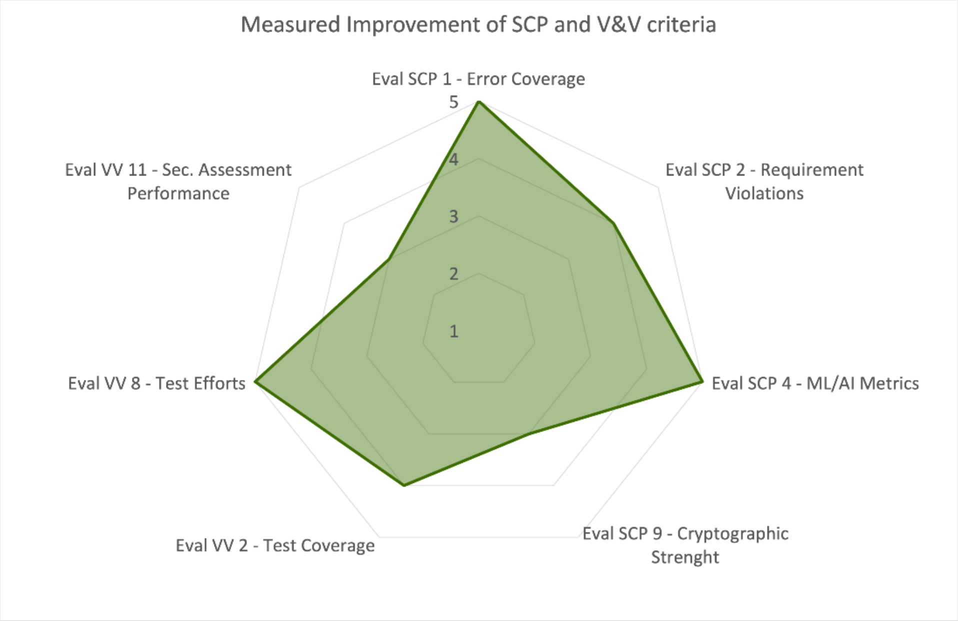


Figure 3-78 Radar plot showing the distribution of improvements regarding evaluation criteria for UC14 (1=no improvement, 5=very good improvement).

3.13.6 Qualitative Results

<u>Demonstrator 1: Hardware-in-the-Loop Test Bed</u>

Participants Profile: QAM is applied to 10 subjects (9 males, 1 female) aged in the range of 24-46. The education level is relatively high as the subject pool is composed of 2 Post-Doc or higher-degree and 1 PhD researcher and 7 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as “CTOs, professors, research engineers, system engineers, software engineers, etc.” having experience in the fields of “health, biometric data collection, fault injection, security, embedded systems, REST APIs, data storage, etc.”

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-128. The results show that all constructs are correlated with each other except the PT-PU and PT-MO pairs.

Table 3-128. UC14 – Demonstrator 1 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.790	1									
MO	0.879	0.722	1								
CO	0.154	0.529	0.264	1							
ROI	0.263	0.676	0.211	0.769	1						
PE	0.381	0.617	0.482	0.060	0.368	1					
PT	-0.158	0.213	-0.002	0.597	0.314	0.299	1				
PR	0.174	0.501	0.427	0.520	0.479	0.522	0.267	1			
SI	0.352	0.601	0.485	0.753	0.394	0.267	0.780	0.354	1		
ATU	0.324	0.711	0.479	0.416	0.506	0.725	0.363	0.568	0.561	1	
BI	0.500	0.662	0.727	0.439	0.389	0.682	0.424	0.513	0.663	0.843	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-129, the questions asked to subjects are sufficiently reliable as understood from subject responses.

Table 3-129. UC14 – Demonstrator 1 Reliability Analysis

Cronbach-alpha values	
PU	0.417
PEOU	0.281
MO	0.423
CO	0.222
ROI	0.530
PE	0.105
PT	0.104
PR	0.559
SI	0.241
ATU	0.467
BI	0.353

Regression Analysis: Regression analysis is applied for estimating the relationships among QAM constructs, as seen in Table 3-130. For this demonstrator, there exists a right proportional relation between all construct pairs influencing each other in the same direction.

Table 3-130. UC14 – Demonstrator 1 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.86xPU + 0.69	0.722	0.001	5.210
H2	CO-PU	Right	CO = 0.14xPU + 4.81	0.024	0.670	0.442
H3	PEoU-PU	Right	PEoU = 1.21xPU -0.94	0.624	0.007	3.641
H4	PU-ATU	Right	PU = 0.31xATU + 4.05	0.105	0.36	0.968
H5	PEoU-ATU	Right	PEoU = 1.03xATU + 0.26	0.505	0.024	2.858
H6	ATU-BI	Right	ATU = 0.87xBI + 1.21	0.711	0.002	4.44
H7	ROI-BI	Right	ROI = 0.39xBI + 3.98	0.151	0.267	1.193
H8	PE-BI	Right	PE = 0.65xBI + 2.57	0.465	0.03	2.635
H9	SI-BI	Right	SI = 0.44xBI + 3.78	0.439	0.037	2.503
H10	PT-BI	Right	PT = 0.39xBI + 4.12	0.180	0.222	1.323
H11	PR-BI	Right	PR = 0.88xBI + 1.40	0.263	0.129	1.69

<u>Demonstrator 2 : CardioWheel enabled VTI driving Simulator for Drowsy data collection</u>

Participants Profile: QAM is applied to 10 subjects (9 males, 1 female) aged in the range of 25-37. The education level is relatively high as the subject pool is composed of 1 Post-Doc or higher-degree and 1 PhD researcher and 8 domain experts who have at least undergraduate degrees in the relevant areas of experience. Subjects are employed as "CTOs, research engineers, system engineers, Q&A, etc." having experience in the fields of "car and vehicle technology, driver status and assistance, embedded systems, REST APIs, data storage etc.".

Correlation Analysis: Correlations among the QAM constructs are presented in Table 3-131. The results show that the majority of constructs are correlated with each other. BI seems not correlated with many constructs.

Table 3-131. UC14 – Demonstrator 2 Correlation Analysis

	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
PU	1										
PEOU	0.791	1									
MO	0.436	0.376	1								
CO	0.600	0.482	0.267	1							
ROI	0.789	0.648	0.219	0.422	1						
PE	0.753	0.903	0.389	0.298	0.796	1					
PT	0.131	0.188	0.331	0.645	-0.134	-0.008	1				
PR	0.015	0.446	0.014	0.232	0.155	0.377	0.535	1			
SI	0.339	0.314	0.031	0.618	0.228	0.191	0.788	0.632	1		
ATU	0.480	0.562	0.102	0.174	0.458	0.607	0.348	0.616	0.570	1	
BI	-0.157	-0.182	0.118	-0.067	-0.199	-0.292	0.430	0.421	0.508	0.273	1

Reliability analysis: To measure the reliability of the questionnaire, Cronbach alpha values are computed for each QAM construct. As shown in Table 3-132, the questions asked to subjects are sufficiently reliable as understood from subject responses., except the questions and answers related to BI.

Table 3-132. UC14 – Demonstrator 2 Reliability Analysis

Cronbach-alpha values	
PU	0.111
PEOU	0.457
MO	0.052
CO	0.092
ROI	0.938
PE	0.075
PT	0.442
PR	0.612
SI	0.109
ATU	0.439
BI	-0.253

Regression Analysis: Regression analysis is applied to estimating the relationships among QAM constructs, as seen in Table 3-133. For this demonstrator, there exists a right proportional relation between all construct pairs influencing each other in the same direction.

Table 3-133. UC14 – Demonstrator 2 Regression Analysis

Hypothesis	Description	Proportional Relation	Regression	R^2	*p*-Value	*t*-Value
H1	MO-PU	Right	MO = 0.44xPU + 3.28	0.190	0.208	1.369
H2	CO-PU	Right	CO = 0.48xPU + 3.17	0.360	0.067	2.120
H3	PEoU-PU	Right	PEoU = 0.99xPU + 0.19	0.625	0.006	3.655
H4	PU-ATU	Right	PU = 0.04xATU + 3.36	0.230	0.161	1.546
H5	PEoU-ATU	Right	PEoU = 0.64xATU + 2.28	0.316	0.091	1.922
H6	ATU-BI	Right	ATU = 0.23xBI + 4.81	0.075	0.445	0.803
H7	ROI-BI	Right	ROI = 0.18xBI + 7.21	0.040	0.581	-0.575
H8	PE-BI	Right	PE = 0.23xBI + 7.49	0.085	0.414	-0.862
H9	SI-BI	Right	SI = 0.35xBI + 4.21	0.258	0.134	1.667
H10	PT-BI	Right	PT = 0.52xBI + 3.44	0.185	0.215	1.348
H11	PR-BI	Right	PR = 0.40xBI + 3.98	0.177	0.225	1.313

3.13.7 Observed Limitations, Lessons Learnt and Best Practices

The VALU3S project was an excellent learning opportuning regarding complex autonomous system validation and verification. Given the complex and multidisciplinary nature of Use Case 14, comprising Hardware, real-time firmware, signal processing and machine learning, system development and feature addition was a strenuous effort that required strict collaboration between several engineers with experience in these fields. Structuring the system's validation by defining requirements, test cases and evaluation scenarios allowed the design of an automated testing station that reduces both time and costs of validation. The contact with V&V experts in the consortium allowed the selection of appropriate tools and formal methods to conduct verification of system design, making it easier to evaluate the impact of new feature additions.

The system's ML classifiers pose an added challenge when it comes to validation and verification. Given that drowsiness is not a discrete classification problem, like telling cats and dogs apart is, defining safety requirements needs to consider the fuzzy nature of these classifiers' outputs. This issue was addressed by framing requirements in terms of maximum levels of false negatives allowed, and further narrowing the subjectivity of the drowsiness state by testing the annotation of drowsy data using reaction times instead of self-reported KSS levels. While these steps help to materialize a set of conditions that the system must meet to be deemed safe, we expect continuous work on improving the safety requirements associated with these ML components of the CardioWheel system.

Finally, the inclusion of formal V&V methods into the validation workflow as obtained in the VALU3S project is to be considered in future product developments, as the project has demonstrated the benefits of a strong and organized V&V workflow based on these methods in ensuring production traceability and high confidence in manufactured products.

As given in Table 3-134 and Table 3-135, both demonstrations are well received by the experts with high acceptance. In all QAM factors, the scores are above 5.00 and the mean values of BI are greater than 6.00/7.00. These evaluations clearly state that the respondents are highly impressed with the UC14 outputs. One of the main reasons is that they can see a tangible output, e.g., a smart electronic wheel, and impressive videos presenting the concrete benefits of the solution. Since this solution improves driver safety, and this can be observed from the realistic simulations, the opinions of the users are generally positive and also promising for future uptake of the demonstrated technologies.

Table 3-134 Mean and standard deviation of experts' responses to UC14 - Demonstration 1

UC14-Demonstrator-1	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,51	5,34	5,61	5,12	5,69	5,60	5,32	5,45	5,44	5,74	6,19
Std Dev	0,92	0,60	0,94	1,04	0,89	0,94	0,97	0,52	1,34	0,87	0,90

Table 3-135 Mean and standard deviation of experts' responses to UC14 - Demonstration 2

UC14-Demonstrator-2	PU	PEOU	MO	CO	ROI	PE	PT	PR	SI	ATU	BI
Mean	5,71	5,61	5,51	5,34	5,90	5,64	5,26	5,56	5,65	5,88	6,18
Std Dev	0,78	0,62	0,76	0,98	0,69	0,77	0,50	0,65	0,89	0,71	0,61

Chapter 4 Impact Assessment

VALU3S offers a wide portfolio of innovations covering six industrial sectors including, i) automotive, ii) agriculture, iii) railway, iv) healthcare, v) aerospace and vi) industrial automation and robotics. These innovations are the natural outcomes of deep scientific research and engineering triggered by the industry to solve the safety, security and reliability problems of cyber-physical automation systems.

Demonstrated the current state of play by VALU3S use cases and demonstrators, this section aims to summarise the PESTLEE analysis results obtained from the pilot activities of the VALU3S project by revisiting the lessons learnt and best practices gathered throughout the developments and demonstrations. PESTLEE analysis is critical to understanding the external threats & opportunities arising because of the macro environment developments. VALU3S consortium has re-elicited the vision, mission and reason of existence for V&V solutions in six industrial domains while identifying the needs of the stakeholders and analysing the technical requirements. Detailed state-of-the-art analyses and market watch updates have been made to re-investigate the strengths, weaknesses, opportunities and threats of the offered innovations in targeted sectors. The provided innovation space has consolidated innovations for more effective automation and control systems, robotics, fault monitoring systems, cyber-physical threat and safety analysis and AI-based cyber-resilience solutions, hardware-supported complex system solutions, vulnerability assessment of tools and services, IoT-enabled multimodal data governance, and enhanced online services for quality control applications.

PESTLEE is a mnemonic which in its expanded form denotes **P** for Political, **E** for Economic, **S** for Social, **T** for Technological, **L** for Legal, **E** for Environmental, and **E** for Ethics. It gives a bird’s eye view of the whole environment from many different angles that one wants to check and keep track of while contemplating a certain idea/plan. In VALU3S the PESTLEE factors are investigated in qualitative assessment up to some extent by collecting the opinions of the experts. A more quantitative PESTLEE analysis can always be done but since such an analysis requires too much effort and time, this is left as a further study. In general, PESTLEE analysis, within the scope of this deliverable, aims to find answers to the following questions by discussing the lessons learnt, best practices, and concluding remarks obtained in twelve use cases of the VALU3S project.

1. **P:** What is the political situation of the country and how can it affect the industry?
2. **E:** What are the prevalent economic factors?
3. **S:** How much importance does culture have in the market and what are its determinants?
4. **T:** What technological innovations are likely to pop up and affect the market structure?
5. **L:** Are there any current legislations that regulate the industry or can there be any change in the legislation for the industry?
6. **E:** What are the environmental concerns for the industry?
7. **E:** What are the ethical concerns that may prevent or postpone the wider acceptance of the offered innovations?

The PESTLEE analysis can be very comprehensive and requires dense work. However, in VALU3S we aim to give an overview of findings mainly in the core fields of innovation which are AI, automation and robotics, and SCP-aware solutions that have comprehensively and centrally addressed in all use cases. The following subsections will summarise the overall impact of VALU3S concerning the PESTLEE factors without going into use-case-specific details.

4.1 Political Factors

VALU3S innovation landscape can have significant political implications both domestically and internationally. Here are some of the key political factors related to the demonstrated V&V innovations:

Employment: The increasing use of AI, robotics and automation in targeted industrial domains can have significant effects on employment rates. As more jobs become automated, there may be a reduction in certain types of jobs, which can lead to social and economic tensions. Such tension should be regulated by policymakers and governments and new employment strategies should be created to overcome this problem. VALU3S solutions increase the level of automation on the one hand but on the other hand, they may create new jobs like maintenance workers, system integrators, logistic and supply chain stakeholders, technicians, and R&D personnel.

Regulations: As offered technologies advance, there will be an increasing need for regulations and laws to ensure the safe, secure and ethical use of the technology. Governments will need to develop policies that balance the potential benefits of i.e., robotics and automation with the need to protect individuals and society from potential risks. VALU3S outputs can present a technical background for the V&V of automated cyber-physical systems which can shed light on how security, safety, ethical and inefficiency problems can be addressed.

International competition: Technologies developed in VALU3S are rapidly developing across the globe, and countries are competing to become leaders in the field. Europe may invest in research and development to gain a competitive advantage and enhance EU security. This requires updates in regulations to incentivise technology developments. VALU3S partners are not only active in their regions but also on an international scale. All partners are members of AENEAS, EPoSS, ARTEMIS and similar initiatives which will be a catalyst for the wider application of VALU3S outputs to industry across Europe.

Privacy: As AI and CPS technologies become more advanced, concerns about privacy and surveillance are likely to arise. Governments will need to balance the potential benefits of these technologies with the need to protect individual privacy rights. VALU3S solution stack comprises the cyber security and privacy requirements by-design. For instance, cyber security and privacy protection have been taken into account by developing hardware- and software-level cyber resilience solutions. The solutions have taken GDPR and national regulations into account even in design phases.

4.2 Economic Factors

The use of AI, automation, robotics, and complex CPSs in six industrial domains can have a significant impact on the economy. Some of the economic factors are as follows:

Increased productivity & Cost savings: Robots and automation technologies can perform tasks more quickly and efficiently than humans, leading to increased productivity and output. The reduction in time and human power in V&V processes leads to significant savings. For instance, offered automated quality control and fault diagnosis techniques developed in VALU3S may prevent design and production line faults. Implementing automated V&V procedures can also reduce labour costs, as fewer human workers may be needed to perform certain tasks.

Improved quality: The offered V&V innovations can help perform tasks with greater precision and accuracy, leading to improved quality of products and services. For instance, the applied system artefact detectors, fault detection mechanisms, and quality inspection algorithms are just a few examples of offered innovations that have a direct impact on improved quality of production and maintenance.

Job displacement & Increased demand for skilled workers: As robotics and automation technologies improve, some jobs may become automated, leading to job displacement for workers in certain industries. This can be perceived as a negative impact by the public. However, in the meantime, this trend will lead to increased professions forcing people to invest in their education and leave the muscle-work to machines. The use of AI and robotics may create a demand for skilled workers who can design, program, and maintain robots VALU3S innovations addressing the needs of workers with disabilities, autonomous train operations, agricultural operations, and teleoperated vehicles are some relevant examples.

Capital investment: Implementing AI-powered automation systems and robotics technologies can require a significant capital investment, which may be a barrier to adoption for some businesses. For instance, evolving from conventional to ADAS-powered vehicles or the adoption of autonomous systems in production lines require significant infrastructure and governance investments. However, as qualitatively observed, the V&V mechanisms may positively influence the return of investment not only in terms of money but also time.

Increased competition: Companies that adopt VALU3S outputs may gain a competitive advantage over those that do not, leading to increased competition in targeted industrial domains. For instance, a faster and more accurate quality inspection system for automotive body-in-white will carry the OEMs to a better position in the market. This can be realised by increased quality, reduced time to market, reduced maintenance costs and increased reputation.

4.3 Social Factors

Social factors of AI, automation, robotics and CPS are complex and multifaceted. As the use of autonomous, collaborative and connected systems become more widespread, it will be important for society to carefully consider the social impacts of this technology and work to address any negative consequences that arise. Although VALU3S is a technology-focused research project social factors associated with offered V&V innovations have been addressed and summarised as follows:

Impact on employment: AI, automation and robotics have the potential to replace human workers in some industries, which can lead to job loss and unemployment. This can have significant social and economic impacts, particularly in communities where certain industries are dominant. The popular media may also provoke regular people about conspiracy theories like robots replacing humans or AI tools like ChatGPT replacing human intelligence.

Ethical concerns: AI, automation and robotics raise ethical questions about the role of machines in society, including questions about autonomy, accountability, and responsibility. For example, if a robot causes harm or makes a mistake, who is responsible?

Changes in social norms: AI and robotics can also change social norms and expectations. For instance, as more robots are used in caregiving roles, society may need to redefine what it means to provide care for others.

Accessibility and inclusion: AI and robotics have the potential to improve accessibility and inclusion for people with disabilities. However, there are concerns that some groups may be excluded from the benefits of such technologies, particularly if they lack access to the technology or the skills to use it. AI bias may become a big problem when biometric authentication and recognition technologies are used.

Privacy and security: Complex CPSs strengthened with automated solutions may also raise concerns about privacy and security. For example, as robots become more common in public spaces, there are questions about how data collected by these devices will be used and protected.

VALU3S use cases have addressed the above topics even during the design phase. Especially SCP requirements have covered safety, security and privacy concerns by-design. Safety functions have been addressed in all use cases that are considered preventive mechanisms against accidents or long-term health problems in some cases. Special attention is given to workers with disabilities in one use case where human-robot collaboration is tried to be optimised. In some cases, innovations are offered for the safety of humans in harsh environments like production lines, factories, aircrafts, etc. On-vehicle innovations supported with ADAS, autonomous operations, traffic surveillance, and human health monitoring are just a few examples of socially-responsible technologies. Personal data protection has been addressed by cyber-physical resilience solutions enabling secure communication and authentication technologies. All offered technologies will have a positive impact on scepticism about AI and robotics as all present significant improvement in performance making people more open to use or benefit from them. The qualitative assessments support related hypotheses indicating public willingness and technology acceptance. The dissemination and training activities implemented in VALU3S will support the increasing interest in project outputs and help wider adoption.

4.4 Technological Factors

Development of automated CPSs in targeted industrial domains is highly dependent on advances in these and other areas of technology, as well as interdisciplinary research and collaboration between engineers, scientists, and other experts. VALU3S use cases have presented a wide portfolio of technologies and disseminated knowledge through a proactive strategy as evident from published papers and organised events and training activities. In line with the VALU3S scope, several technology factors are crucial to the development and advancement of AI, automation, robotics and complex CPSs. Some of these factors include:

Sensors and IoT: Sensors are critical components of any complex CPSs. They enable CPS components like robotic arms or IoT systems to gather data from the devices, systems, humans and environment and make decisions based on that multidimensional data. There are many types of sensors used in VALU3S use cases, including cameras, LIDAR, ultrasound, and infrared sensors. In all cases of VALU3S applications, the IoT backend systems are somehow utilised as impressive applications of sensor-powered CPSs.

Actuators: Actuators are the components that enable robotic systems or connected, cooperative and/or autonomous vehicles and systems to move and perform physical actions. These can include motors, hydraulic and pneumatic systems, and other mechanisms that enable robots to manipulate objects in their environment. VALU3S have presented exemplary use cases where actuators are effectively used. For instance, automated robot inspection cell for quality control, industrial drives for motion control, autonomous train operation control systems, agricultural robots are just a few examples.

Control systems: Control systems are responsible for coordinating the actions of automated systems, including movement and manipulation. These systems use algorithms and software to translate sensor data into actions that actuators can take. All use cases have showcased exemplary control systems. Infusion control MNT, robotic arms used is disassembly processes and production lines, aircraft engine controllers, and motion control solutions have been demonstrated as well-appreciated cases during evaluations.

Artificial intelligence (AI) and machine learning (ML): AI and ML are increasingly important components of automated technologies. They enable CPSs to learn from experience, make decisions, and adapt to changing environments. Anomaly detection, fault tolerance techniques, safety trajectory optimisation techniques, autonomous operations, tele-operated vehicles, vital signs controller by means of drug infusion, cyber incident detection, person authentication, driver behaviour monitoring, dependable systems are some examples where AI/ML algorithms are deeply used.

Communication and Cyber Security: The offered innovations often require communication with other robots, humans, or other devices. Advances in wireless communication and networking technologies are benefited to enable sophisticated and complex interactions between CPSs and other systems. In VALU3S, advanced hardware- and software-based cyber security solutions have been tested. Among these, hardware security modules, secure IoT gateways, AI-based cyber anomaly detection and management systems have been implemented to support end-to-end and holistic security of CPSs.

Power and energy: Automated systems require a significant amount of power to operate, and energy efficiency is an important consideration in their design. Advances in battery technology and power management systems are helping to improve the capabilities of robots and extend their operating times. The power requirements are all addressed by default in all use cases of VALU3S. For instance, industrial drives for motion control systems are typically built with PLCs (Programmable Logic Controller) and efficient power inverters for controlling electrical motors.

4.5 Legal Factors

As automated technologies advance and become more prevalent, several legal factors need to be considered. Some of these legal factors include:

Liability: With the increased use of autonomous systems and robotics, there is a need to determine who is responsible for any damages or accidents that may occur as a result of the CPS actions. Liability laws may need to be updated to address the unique issues that arise with the evolving V&V procedures.

Intellectual Property: The VALU3S technology landscape is often protected by patents and other intellectual property rights. As such, legal frameworks need to be developed to protect the interests of the owners of these rights while also encouraging innovation and competition.

Privacy: The AI/ML-based technologies addressed in VALU3S can be used to gather large amounts of data, including personal information. Laws and regulations need to be developed to protect individual privacy in the context of GDPR, national rules and the practical use of automated systems and robotics in targeted domains of application.

Employment: Robotics and autonomous technologies have the potential to displace human workers, leading to job losses. Legal frameworks may need to be developed to address the potential impact of machine-dominated technologies on employment.

Safety: As automated technologies and robotics become more complex, there is a need to ensure that it is safe for use in various settings. Safety standards may need to be developed to ensure that the CPSs are designed and operated in a way that minimizes risks to human users.

Overall, as data-oriented technologies continue to advance, there will be a need for legal frameworks that address the unique issues that arise with this technology. VALU3S solution stack has special attention to address the GDPR and safety issues in all use cases starting from the early phases, i.e. requirement elicitation. The data collection, quantitative and qualitative assessment methodologies can be seen as best practices for legal authorities to observe how legal factors can be addressed in targeted industrial domains of demonstrations. The applied IP management and exploitation strategy, as evident from the project consortium agreement, is an exemplary work. The qualitative assessments and standards scrutinise also help policy-makers to re-elicit and identify the gaps in the existing legal framework. The outputs can be used to update regulations in emerging fields where human-robot collaboration takes places or where disabled/disadvantaged people work in industrial environments. The legal framework related to surveillance and person authentication can be extended to node (sensor, actuator, device, vehicle, etc.) and system level which is still an open topic in the use of CPSs in complex environments.

4.6 Environmental Factors

It is important to consider the environmental factors when designing, deploying, and operating automated systems to ensure optimal performance and longevity. Europe has a strict strategy to reach zero-carbon economy by 2050, namely Green Deal, which forces industries to evolve to environment-friendly technologies. There are several environmental factors that can impact the operation and performance of automated systems and robotics. In this section, we present some examples.

Uncertainty in environmental conditions where VALU3S outputs can be used: The operation environments are complex and usually chaotic by nature due to the following factors:

- Temperature: CPSs can be affected by extreme temperatures. High temperatures can cause components to overheat and fail, while low temperatures can cause issues with lubrication and impact the accuracy of sensors.
- Humidity: High humidity levels can cause corrosion in components and impact the accuracy of sensors. It can also cause electrical short circuits in some cases.
- Dust and debris: Dust and debris can cause components to clog, impairing the movement and function of automated systems.
- Lighting: Poor lighting conditions can affect the performance of sensors used in robotics, such as cameras and laser scanners.
- Magnetic fields: Strong magnetic fields can disrupt the operation of electronic components and sensors in robots, leading to inaccurate readings and errors.
- Noise: Excessive noise levels can interfere with the operation of sensors and make it difficult for robots to accurately detect and respond to their environment.
- Vibration: Vibration can cause damage to sensitive components and impair the precision and accuracy of robotic movements.

Environmentally sensitive productions: The smart production facilities represent a leap forward from more traditional automation to a fully connected and flexible system. Such automated systems utilise a constant stream of data from connected operations and production systems to learn and adapt to new demands. Legacy Supervisory control and data acquisition (SCADA) devices are being replaced by new connected devices allowing for increased control over the processes and a greener operation.

Pollution & Climate Change: Industrialisation is one of the biggest threats to the environment as conventional factories, fossil-fuel vehicles and energy systems are seen as forthcoming polluters. Climate change, when combined with natural disasters and wars, is also expected to cause serious social problems like the migration of people from one place to another, reduced living environments and quality of life.

VALU3S partners are aware of the Green Deal goals and consider the environmental constraints, standards and regulations during the design phase. Since the proposed V&V methods aim to improve the quality and effectiveness of automated systems, operational efficiency is a natural outcome of the offered innovations. For instance, autonomous vehicles used in agriculture, tele-operated vehicles, and robotic systems used in production lines are all serving the Green ICT concept. The developed control system is supporting the no-light factories and quantitative and precise monitoring of GHG emissions. Effective monitorisation of temperature, humidity, dust, lighting, environmental conditions, production processes, and maintenance services will measurably improve the Green Deal assessment. The offered innovations will support the connectivity, cooperativeness and autonomy of railway, air and road vehicles which will lead OEMs to manufacture carbon-free solutions in the automotive domain.

4.7 Ethical Factors

Automated systems and robotics technology raise several ethical questions, including issues related to autonomy, responsibility, and transparency. Ethics come to the fore with increasing violations of privacy, biases in decision-making, and lack of control over automated systems and robots. And solutions to ethics issues need to be scalable, as intelligent automation becomes ever more widely applied, more deeply embedded in customer solutions, and more responsible for decisions that affect lives—such as medical diagnoses, government benefit payments, and SCP concerns including both daily and industrial settings.

Legal scholars are already busy identifying the issues that will inevitably arise and proposing frameworks and principles for dealing with them consistently [48]. The four elements of PAPA, including Privacy, Accuracy, Property, and Accessibility, are also the core issues of ethics at the technology and data levels. Since the V&V concept is multifaceted and still evolving, the data protection rules, e.g., GDPR, privacy preservation tactics and anytime-anywhere accessibility options should be seriously analysed by considering cyber resilience and trustworthiness.

Four of these principles are certain to remain pillars of “responsible automation.” To avoid causing reckless or heedless damage, solutions will have to be unbiased, transparent, controllable, and protected. A good example in VALU3S is the utilisation of surveillance systems and camera-based observation systems in several use cases. Human tracking systems are prone to ethical concerns not only from the GDPR perspective but also from the ethical perspective. Any AI bias on a specific social clique or low-performing algorithms discriminating people may raise debates. Automated decision-making will be subjected to even greater scrutiny in years to come. For instance, in autonomous driving, automatic assembly processes or infusion control in NMT practices trustworthy AI, security and personal data protection are must-have countermeasures against ethical concerns and privacy leakages. VALU3S has covered such problems, also evident from the qualitative assessment results, presenting a commendable and credible solution stack.

Chapter 5 Conclusion

In this report, updated quantitative and qualitative evaluation, concerning baseline measurements, expert opinions collected through online questionnaires and the description of achievements, observations, lessons learnt, and best practices are presented. The main purpose of the UC demonstrations is to show how the project results improve the quality of developed products and improve the time, cost, and effort spent in the engineering phases of V&V processes. Each demonstrator has adopted a different set of evaluation criteria and selected metrics that could be used for a successful demonstration. The observed measurements are used for the quantitative evaluation of project results and field demonstrations. Additionally, the Qualitative Assessment Model is applied to evaluate the experts' feedback through a questionnaire. The questionnaire responses are statistically analysed by considering the reliability, correlation, and significance of user responses to identify how QAM constructs (PU, PEOU, MO, CO, RO, PE, PT, PR, SI, ATU, and BI) influence each other. The results of the quantitative analyses, as expected, show that the developed tools and toolchains perform well enough and even beyond the state–of–art in many cases. For the qualitative assessment, the expert opinions are usually positive. However, since it is practically not feasible to reach high numbers of experts (first because it is very hard to find experts in niche areas of implementation; second the questionnaire applications take about 40-50 minutes per subject) the statistical significance can be improved in further studies.

In this deliverable, 21 different demonstrators have been described (identified by 13 use cases of the project reported in D1.1 [1], without UC12 since the use case provider terminated its participation in the project but including UC14 recently provided by CARDIOID) covering all main domains. Moreover, an evaluation of the baseline has been introduced and measurements of V&V aspects (both qualitative and quantitative) have been interpreted by applying the PESTLEE criteria to assess their overall impact compared with the final state of demonstrators. The PESTLEE analysis presents a visionary overview of VALU3S outputs and how they can impact the European industry in targeted fields of application.

References

[1] NXP-DE et al., “Deliverable D1.1 – Description of use cases as well as scenarios”. VALU3S Consortium, May 27, 2021.

[2] UTRCI et al., “Deliverable D1.2 – SCP requirements as well as identified test cases”. VALU3S Consortium, May 28, 2021.

[3] FRAUNHOFER IESE et al. “Deliverable D4.4 – Initial Detailed Description of Improved Process Workflows”. VALU3S Consortium, Oct. 29, 2021.

[4] LLSG et al., “Deliverable D4.5 - Initial Implementation of V&V Tools Suitable for the Improved Process Workflows”. VALU3S Consortium, Oct. 29, 2021.

[5] AIT et al., “Deliverable D3.5 - Interim description of methods designed to improve the V&V process”. VALU3S Consortium, Oct. 29, 2021.

[6] AIT et al., “Deliverable D3.6 – Final description of methods designed to improve the V&V process ”. VALU3S Consortium, Nov. 26, 2021.

[7] BUT et al., “Deliverable D5.1 – Initial Demonstration Plan and a List of Evaluation Criteria”. VALU3S Consortium, Dec. 17, 2020.

[8] BUT et al., “Deliverable D5.2 - Final Demonstration Plan and a List of Evaluation Criteria”. VALU3S Consortium, Oct. 31, 2021.

[9] CAMEA et al., “Deliverable D5.3 - Initial Demonstrator Implementation Status Report”. VALU3S Consortium, Apr. 29, 2022.

[10] CAMEA et al., “Deliverable D5.4 - Demonstrator prototypes”. VALU3S Consortium, Oct. 31, 2022.

[11] CAMEA et al., “D5.5 - Final demonstrator implementation status report” VALU3S Consortium, April. 31, 2023.

[12] Definitions of quantitative and qualitative assessment, https://csrc.nist.gov/glossary.

[13] Marangunić, Nikola, and Andrina Granić. "Technology acceptance model: a literature review from 1986 to 2013." Universal access in the information society 14.1 (2015): 81-95.

[14] Williams, Michael D., Nripendra P. Rana, and Yogesh K. Dwivedi. "The unified theory of acceptance and use of technology (UTAUT): a literature review." Journal of enterprise information management (2015).

[15] Giannakopoulou, D., Pressburger, T., Mavridou, A., & Schumann, J. (2020). “Generation of formal requirements from structured natural language”. In International working conference on requirements engineering: Foundation for software quality (pp. 19-35), Springer, 2020.

[16] Luckcuck, M., Farrell, M., Sheridan, O., & Monahan, R. “A Methodology for Developing a Verifiable Aircraft Engine Controller from Formal Requirements”, In IEEE Aerospace Conference, 2022.

[17] A. Lourenço, A. P. Alves, C. Carreiras, R. Duarte och A. L. Fred, ”CardioWheel: ECG biometrics on the steering wheel,” i European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - ECML/PKDD, Porto, Portugal, 2015.

[18] A. Lourenço, C. Carreiras, A. Alves, S. Silveira och J. Cardioso, ”CardioWheel: Physiological Driver Monitoring,” i 6th International Symposioum on Naturalistic Driving Research, The Hague, Netherlands, 2017.

[19] CENELEC (European Committee for Electrotechnical Standardization), “EN 50129 Railway applications. Communication, signalling and processing systems. Safety related electronic systems for signalling”, 2003.

[20] CENELEC (European Committee for Electrotechnical Standardization), “EN 50128 Railway applications –Communications, signalling and processing systems –Software for railway control and protection systems”, 2011.

[21] CENELEC (European Committee for Electrotechnical Standardization), “EN 50126 Railway applications - The specification and demonstration of Reliability, Availability, Maintainability and Safety (RAMS)”, 2000.

[22] IEC (International Electrotechnical Commission), ”IEC 61508 Functional safety of electrical/electronic/ programmable electronic safety-related systems”, 2010.

[23] José Proença, Sina Borrami, Jorge Sanchez de Nova, David Pereira and Giann Nandi "Verification of multiple models of a safety-critical motor controller in railway systems", RSSRail2022.

[24] Aguirre, A., Lozano-Rodero, A., Matey, L. M., Villamañe, M., & Ferrero, B. (2014). A novel approach to diagnosing motor skills. IEEE Transactions on Learning Technologies, 7(4), 304-318.

[25] Aguirre, A., Lozano-Rodero, A., Villamañe, M., Ferrero, B., & Matey, L. M. (2012). OLYMPUS: An Intelligent Interactive Learning Platform for Procedural Tasks. In GRAPP/IVAPP (pp. 543-550).

[26] A. Arrieta, S. Wang, U. Markiegi, G. Sagardui, L. Etxeberria. “Employing Multi-Objective Search to Enhance Reactive Test Case Generation and Prioritization for Testing Industrial Cyber-Physical Systems” in IEEE Transactions on Industrial Informatics, vol. 14, no. 3, pp. 1055-1066, March 2018, doi: 10.1109/TII.2017.2788019.

[27] A. Arrieta, S. Wang, U. Markiegi, G. Sagardui and L. Etxeberria, "Search-based test case generation for Cyber-Physical Systems," 2017 IEEE Congress on Evolutionary Computation (CEC), San Sebastian, 2017, pp. 688-697, doi: 10.1109/CEC.2017.7969377.

[28] ISO (International Organization for Standardization), “ISO 10218-1:2011 Robots and robotic devices - Safety requirements for industrial robots, Part 1: Robots”, 2011.

[29] ISO (International Organization for Standardization), “ISO 10218-2:2011 Robots and robotic devices - Safety requirements for industrial robots, Part 2: Robot systems and integration”, 2011.

[30] ISO (International Organization for Standardization), “ISO/TS 15066:2016 Robots and robotic devices - Collaborative robots”, 2016.

[31] H. silva, A. Lourenço, A. Fred, “Device and method for continuous biometric recognition based on electrocardiographic signals”, WO2013109154A1, May 25, 2013

[32] Alexandre David, Kim G. Larsen, Axel Legay, Marius Miku£ionis, and Danny Bøgsted Poulsen. Uppaal SMC tutorial. STTT, 17(4):397415, Aug 2015.

[33] T. Kuhn, T. Forster, T. Braun and R. Gotzhein, “Feral - framework for simulator coupling on requirements and architecture level,” *ACM/IEEE MEMOCODE*, p. 11–22, 2013.

[34] ISO 26262-1:2018 “Road vehicles — Functional safety — Part 1: Vocabulary” https://www.iso.org/standard/68383.html

[35] ISO 25119-2:2019 “Tractors and machinery for agriculture and forestry — Safety-related parts of control systems — Part 2: Concept phase” https://www.iso.org/standard/78306.html

[36] IEC 62304:2006, “Medical device software — Software life cycle processes” https://www.iso.org/standard/38421.html

[37] ISO 14971:2019, “Medical devices — Application of risk management to medical devices” https://www.iso.org/standard/72704.html

[38] AIT Austrian Institute of Technology, “MoMuT”, [online]. Available: https://momut.org/.

[39] “NS-3 network simulator”, [online]. Available: https://www.nsnam.org/.

[40] OpenSim Ltd., “Omnet++ discrete event simulator”, [online]. Available: https://omnetpp.org/.

[41] AIT Austrian Institute of Technology, “ThreatGet (threat analysis and risk management)”, [online]. Available: https://www.threatget.com/.

[42] Facebook, “Infer”, [online]. Available: https://fbinfer.com/.

[43] ”Frama-C software analyzers”, [online]. Available: https://frama-c.com/.

[44] The Eclipse Foundation, “CHESS tool”, [online]. Available: https://www.eclipse.org/chess/.

[45] ISO 12100:2010 Safety of machinery — General principles for design — Risk assessment and risk reduction, [online]. Available: https://www.iso.org/standard/51528.html.

[46] Europe's Rail Joint Undertaking, “Shift2Rail IP2 - X2RAIL-4”, [online]. Available: https://projects.shift2rail.org/s2r_ip2_n.aspx?p=X2RAIL-4.

[47] ISO 25119:2018 Tractors and machinery for agriculture and forestry — Safety-related parts of control systems.

[48] Yavar Bathaee, “The Artificial Intelligence Black Box and the Failure of Intent and Causation,” Harvard Journal of Law & Technology 31, no. 2 (Spring 2018).

Appendix A Questionnaire Applied to Conduct QAM

The questions listed in Table A-1 are asked to subjects to assess the use case demonstrators qualitatively.

Table A-1 QAM questionnaire

Construct	Questions	Personal opinion	Organisational Opinion
Perceived Usefulness (PU) – Likert Scale			
PU1	I or my organisation can complete the V&V procedure faster with the presented tools & applications		
PU2	I or my organisation can be more productive at field of work thanks to V&V tools & applications		
PU3	Using V&V tools & applications can make my or my organisation's job easier.		
PU4	I find V&V tools & applications useful for my or my organisation's business.		
Perceived Ease of Use (PEoU) – Likert Scale			
PEoU1	I or related teams in my organisation can easily learn to use V&V tools & applications.		
PEoU2	I or related teams in my organisation want to use V&V tools & applications to achieve what I/we want.		
PEoU3	Using V&V tools & applications do not require much mental and physical effort.		
PEoU4	I or related teams in my organisation would find V&V tools & applications easy to use.		
Motivation (MO) – Likert Scale			
MO1	Advances in the targeted technology area supported with the proposed V6V tools excite and motivate me or my company.		
MO2	V&V tools & applications are very much applicable to my tasks or the tasks of related teams in my organisation		
Compatibility (CO) – Likert Scale			
CO1	I think V&V tools & applications can easily integrate into existing mainstream systems widely adopted in our field of practice		
CO2	I think V&V tools & applications can easily communicate with each other.		
CO3	I think V&V tools & applications can easily integrate into our company's IT and OT networks.		
Return Of Investment Expectancy (ROI) – Likert Scale			

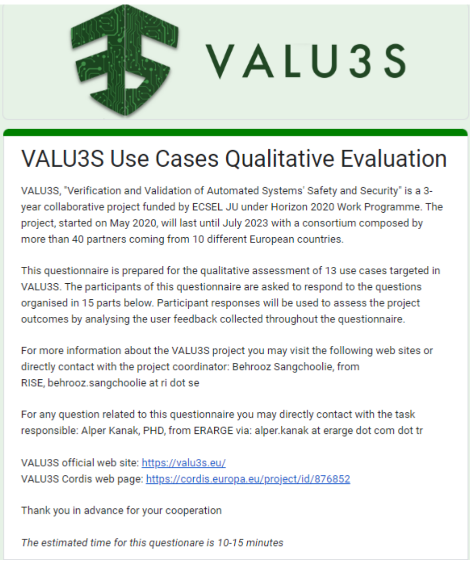
Construct	Questions	Personal opinion	Organisational Opinion
ROI1	I think/My organisation would believe the cost of setting possible V&V tools & applications is not an obstacle for our company		
ROI2	I think//My organisation would believe the benefits of V&V tools & applications will overweigh the implementation costs		
ROI3	V&V tools & applications and related technology play an essential role in reducing operational costs		
ROI4	It is possible to obtain an acceptable ROI from the application of V&V workflows.		
ROI5	V&V tools & applications would enable my organization to be more competitive and increase my market share.		
ROI6	V&V tools & applications would enable my organization to penetrate new markets.		
Performance Expectancy (PE) – Likert Scale			
PE1	V&V tools and applications enable me/my organisation to improve the quality of products, services and outputs		
PE2	V&V tools and applications enable me/my organisation to shorten the time of services or processes		
PE3	V&V tools and applications enable me/my organisation to improve efficiency of services or processes		
PE4	V&V tools and applications enable me/my organisation to reduce errors and faults		
Perceived Trust (PT) – Likert Scale			
PT1	V&V tools & applications are trustworthy.		
PT2	I/my organisation rely on the data that is used or collected by the V&V tools & applications		
PT3	In a possible project, V&V tools & applications will securely communicate with each other or with host services and other tools.		
PT4	The outputs of V&V tools & applications that I/my organisation use are error-free		
PT5	V&V tool & application providers will fulfill their commitments in a possible project		
PT6	The V&V tools & applications are self-explainable so that I/my organisation can understand what their outputs mean		
PT7	I/my organisation believe that V&V tools & applications are safe and will not cause any accident or health problem.		
PT8	I/my organisation think that the outputs of the V&V tools & applications do not create any privacy concern and/or compliant with GDPR		

Construct	Questions	Personal opinion	Organisational Opinion
PT9	I/my organisation think that the inputs or outputs of the V&V tools & applications do not raise any unfairness, bias on any social group or other ethical concerns.		
PT10	I/my organisation believe that all outputs of the V&V tools & applications are accountable and non-repudiative.		
PT11	I am confident that the V&V tool & application providers protect me/my organisation from any problems we may encounter		
Perceived Risk (PR) – Likert Scale			
PR1	Failure of V&V tools & applications can lead to complicated problems.		
PR2	A possible cyber-physical attack on the targeted infrastructure (where V&V applied) I/my organisation use may significantly affect my company's operation.		
PR3	The privacy, security, safety issues of the proposed V&V tool affect my investment plans in this Technology		
PR4	Organizations that regulate standards need to step up for better V&V procedures		
Social Influence (SI)– Likert Scale			
SI1	People who are important to me think that I or my organisation should use t he V&V tools and applications		
SI2	People who influence my behaviour think that I or my organisation should use the V&V tools and applications		
SI3	People whose opinions I value would like me or my organisation to use the V&V tools and applications		
	I would recommend a conditionally automated car to others		
Attitudes toward Using the V&V (ATU) – Likert Scale			
ATU1	Our board of directors and senior executives agree that V&V tools & applications are necessary for our company to thrive.		
ATU2	We have sufficient knowledge and staff to deploy and manage a V&V tool or application.		
ATU3	Our company has plans to implement an V&V project (or processes) in near future (2-4 years)		
Behavioral Intention (BI)– Likert Scale			
BI1	I / my company see V&V tools & applications as a benefit to our organization.		

Construct	Questions	Personal opinion	Organisational Opinion
BI2	I / my company want to use V&V tools & applications given a chance.		

Appendix B Snapshots from the Online Questionnaire

The following snapshots (see Figure B-1 - Figure B-8) are taken from the online questionnaire that is accessible via the following link: https://forms.gle/ALSRmWWwuK7xBpJQ8.



VALU3S Use Cases Qualitative Evaluation

VALU3S, "Verification and Validation of Automated Systems' Safety and Security" is a 3-year collaborative project funded by ECSEL JU under Horizon 2020 Work Programme. The project, started on May 2020, will last until July 2023 with a consortium composed by more than 40 partners coming from 10 different European countries.

This questionnaire is prepared for the qualitative assessment of 13 use cases targeted in VALU3S. The participants of this questionnaire are asked to respond to the questions organised in 15 parts below. Participant responses will be used to assess the project outcomes by analysing the user feedback collected throughout the questionnaire.

For more information about the VALU3S project you may visit the following web sites or directly contact with the project coordinator: Behrooz Sangchoolie, from
RISE, behrooz.sangchoolie at ri dot se

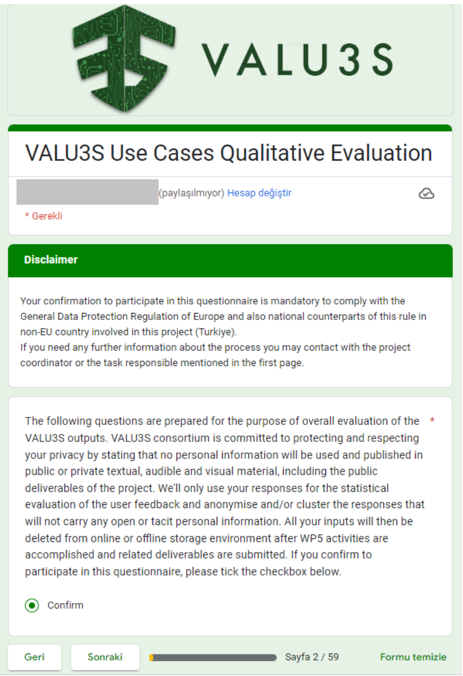
For any question related to this questionnaire you may directly contact with the task responsible: Alper Kanak, PHD, from ERARGE via: alper.kanak at erarge dot com dot tr

VALU3S official web site: https://valu3s.eu/
VALU3S Cordis web page: https://cordis.europa.eu/project/id/876852

Thank you in advance for your cooperation

The estimated time for this questionare is 10-15 minutes

Figure B-1 Opening Page.



VALU3S Use Cases Qualitative Evaluation

[redacted] (paylaşılmıyor) Hesap değiştir

* Gerekli

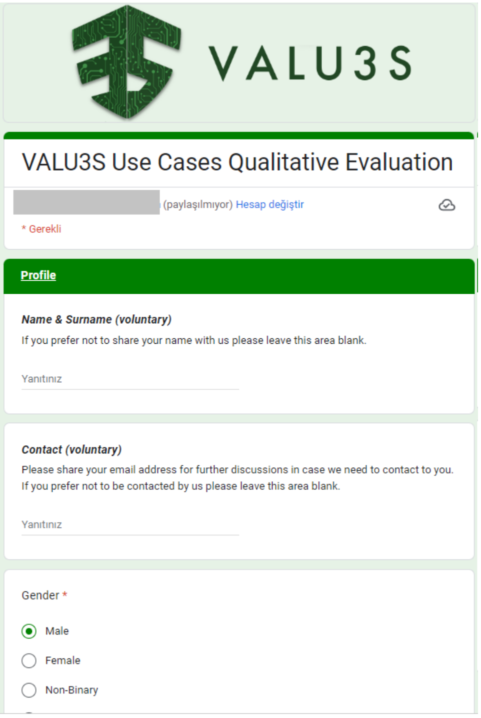
Disclaimer

Your confirmation to participate in this questionnaire is mandatory to comply with the General Data Protection Regulation of Europe and also national counterparts of this rule in non-EU country involved in this project (Turkiye).
If you need any further information about the process you may contact with the project coordinator or the task responsible mentioned in the first page.

The following questions are prepared for the purpose of overall evaluation of the VALU3S outputs. VALU3S consortium is committed to protecting and respecting your privacy by stating that no personal information will be used and published in public or private textual, audible and visual material, including the public deliverables of the project. We'll only use your responses for the statistical evaluation of the user feedback and anonymise and/or cluster the responses that will not carry any open or tacit personal information. All your inputs will then be deleted from online or offline storage environment after WP5 activities are accomplished and related deliverables are submitted. If you confirm to participate in this questionnaire, please tick the checkbox below. *

● Confirm

Figure B-2 Disclaimer Page.



VALU3S Use Cases Qualitative Evaluation

[redacted] (paylaşılmıyor) Hesap değiştir

* Gerekli

Profile

Name & Surname (voluntary)

If you prefer not to share your name with us please leave this area blank.

Yanıtınız

Contact (voluntary)

Please share your email address for further discussions in case we need to contact to you.
If you prefer not to be contacted by us please leave this area blank.

Yanıtınız

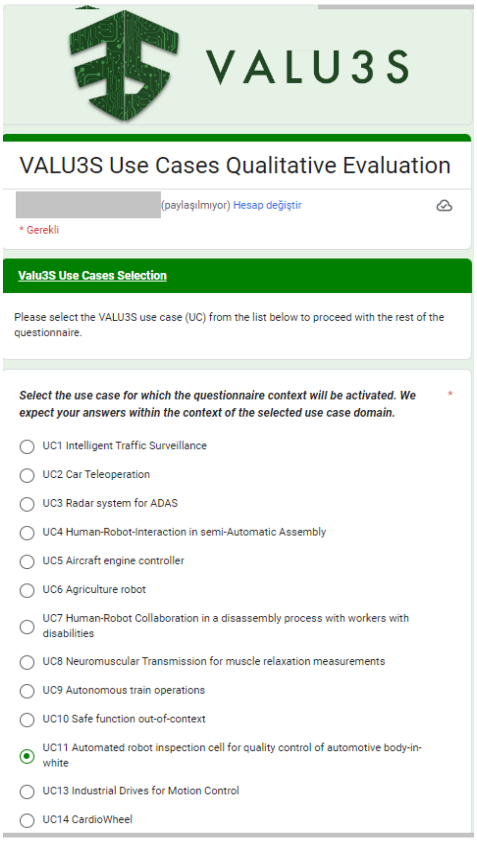
Gender *

● Male

○ Female

○ Non-Binary

Figure B-3 Profile Page.



VALU3S Use Cases Qualitative Evaluation

[redacted] (paylaşılmıyor) Hesap değiştir

* Gerekli

Valu3S Use Cases Selection

Please select the VALU3S use case (UC) from the list below to proceed with the rest of the questionnaire.

Select the use case for which the questionnaire context will be activated. We expect your answers within the context of the selected use case domain. *

○ UC1 Intelligent Traffic Surveillance

○ UC2 Car Teleoperation

○ UC3 Radar system for ADAS

○ UC4 Human-Robot-Interaction in semi-Automatic Assembly

○ UC5 Aircraft engine controller

○ UC6 Agriculture robot

○ UC7 Human-Robot Collaboration in a disassembly process with workers with disabilities

○ UC8 Neuromuscular Transmission for muscle relaxation measurements

○ UC9 Autonomous train operations

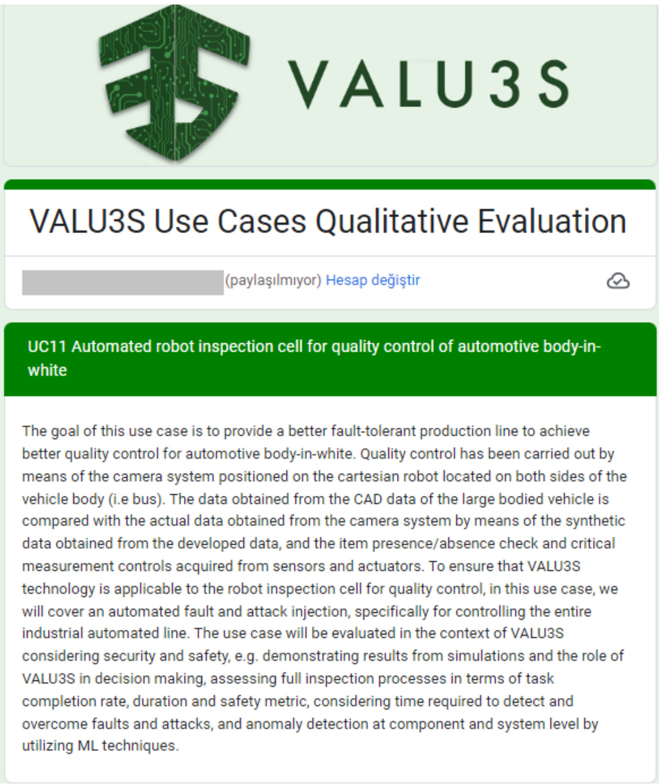
○ UC10 Safe function out-of-context

● UC11 Automated robot inspection cell for quality control of automotive body-in-white

○ UC13 Industrial Drives for Motion Control

○ UC14 CardioWheel

Figure B-4 Use Case Selection.



VALU3S

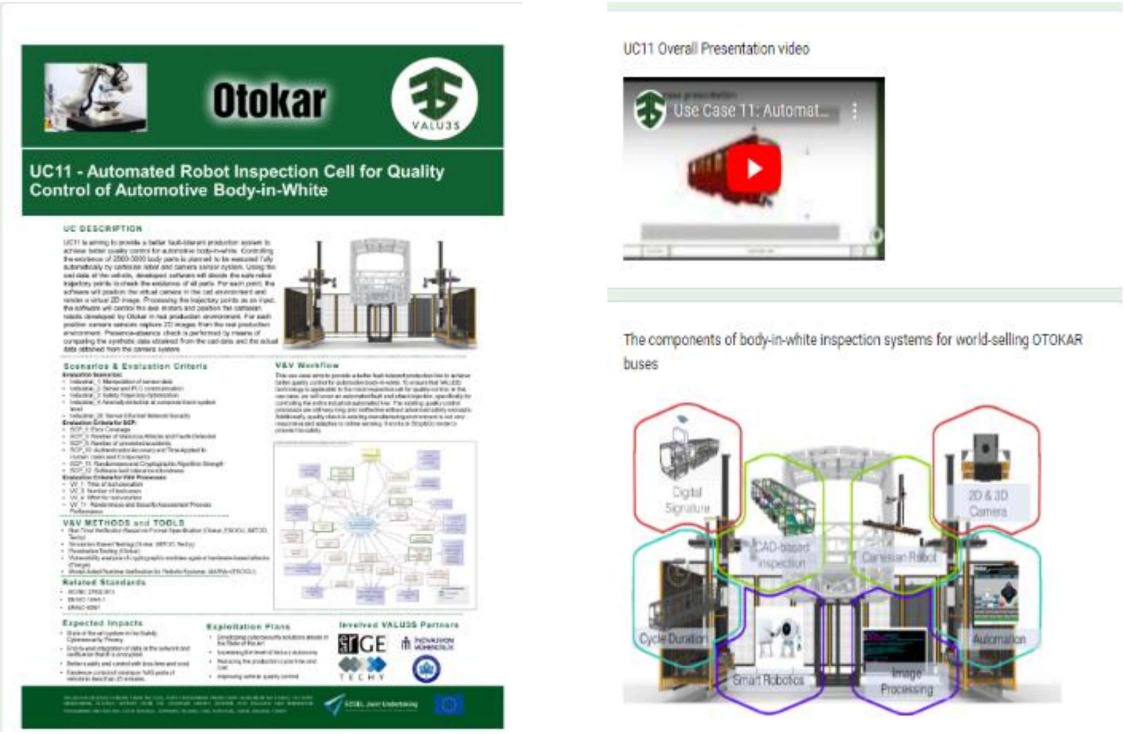
VALU3S Use Cases Qualitative Evaluation

[redacted](paylaşılmıyor) Hesap değiştir

UC11 Automated robot inspection cell for quality control of automotive body-in-white

The goal of this use case is to provide a better fault-tolerant production line to achieve better quality control for automotive body-in-white. Quality control has been carried out by means of the camera system positioned on the cartesian robot located on both sides of the vehicle body (i.e bus). The data obtained from the CAD data of the large bodied vehicle is compared with the actual data obtained from the camera system by means of the synthetic data obtained from the developed data, and the item presence/absence check and critical measurement controls acquired from sensors and actuators. To ensure that VALU3S technology is applicable to the robot inspection cell for quality control, in this use case, we will cover an automated fault and attack injection, specifically for controlling the entire industrial automated line. The use case will be evaluated in the context of VALU3S considering security and safety, e.g. demonstrating results from simulations and the role of VALU3S in decision making, assessing full inspection processes in terms of task completion rate, duration and safety metric, considering time required to detect and overcome faults and attacks, and anomaly detection at component and system level by utilizing ML techniques.

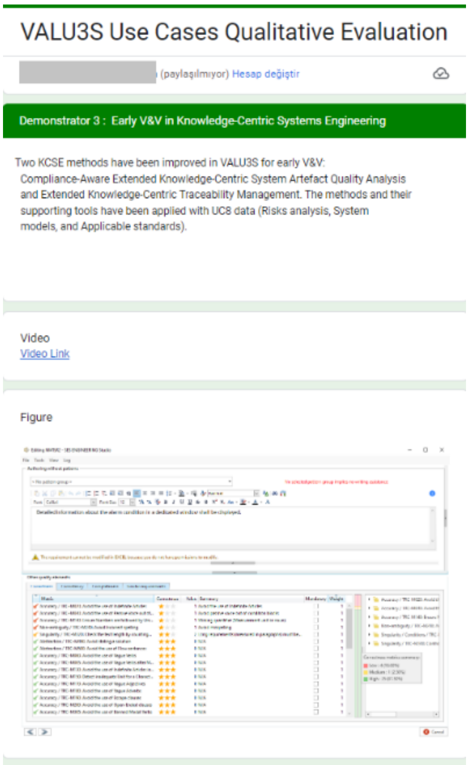
Figure B-5 Use Case Description.



UC11 Overall Presentation video

The components of body-in-white inspection systems for world-selling OTOKAR buses

Figure B-6 Use Case poster and visual materials.



VALU3S Use Cases Qualitative Evaluation

[redacted] (paylaşılmıyor) Hesap değiştir

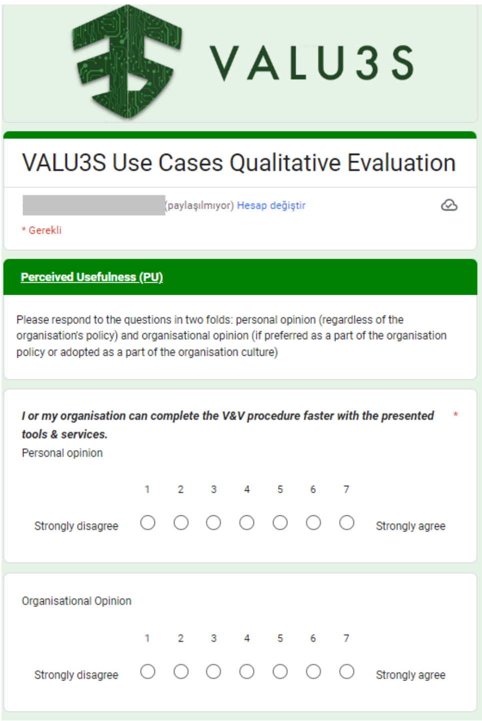
Demonstrator 3 : Early V&V in Knowledge-Centric Systems Engineering

Two KCSE methods have been improved in VALU3S for early V&V: Compliance-Aware Extended Knowledge-Centric System Artefact Quality Analysis and Extended Knowledge-Centric Traceability Management. The methods and their supporting tools have been applied with UC8 data (Risks analysis, System models, and Applicable standards).

Video
Video Link

Figure

Figure B-7 Demonstrator description (under use cases) and visual materials.



VALU3S Use Cases Qualitative Evaluation

[redacted] (paylaşılmıyor) Hesap değiştir

* Gerekli

Perceived Usefulness (PU)

Please respond to the questions in two folds: personal opinion (regardless of the organisation's policy) and organisational opinion (if preferred as a part of the organisation policy or adopted as a part of the organisation culture)

I or my organisation can complete the V&V procedure faster with the presented tools & services. *

Personal opinion

	1	2	3	4	5	6	7	
Strongly disagree	○	○	○	○	○	○	○	Strongly agree

Organisational Opinion

	1	2	3	4	5	6	7	
Strongly disagree	○	○	○	○	○	○	○	Strongly agree

Figure B-8 Questions given in Likert Scale.

VALU3S

www.valu3s.eu

ECSEL Joint Undertaking
Electronic Components and Systems for European Leadership

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876852. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Austria, Czech Republic, Germany, Ireland, Italy, Portugal, Spain, Sweden, Turkey.