Intelligent Verification/Validation for XR Based Systems

Research and Innovation Action

Grant agreement no.: 856716

D1.3 – Final Project Report

iv4XR – WP1 – D1.3

Version 1.4

December 2022

Project Reference	EU H2020-ICT-2018-3 - 856716
Due Date	31/12/2022
Actual Date	30/12/2021
Document Author/s	Marta Couto (INESC-ID), Pedro M. Fernandes (INESC-ID), Rui Prada (INESC-ID), Beatriz Marin (UPV), Fernando Pastor (UPV), Tanja Vos (UPV), Joseph Davidson (GA), Jeremy Cooke (GWE) Wisnhu Prasetya (UU), Fitsum Kifetew (FBK), Ilda Ribeiro (INESC-ID)
Version	1.4
Dissemination level	Public
Status	Final

This project has received funding from the European Union's Horizon 2020 Research and innovation programme under grant agreement No 856716

Document Version Control			
Version	**Date**	**Change Made (and if appropriate reason for change)**	**Initials of Commentator(s) or Author(s)**
1.0	28/11/2022	Initial document structure and contents	MC
1.1	25/12/2022	Added information on WP5	JD
1.1	9/12/2022	Added information on WP6	BM
1.1	12/12/2022	Added information on WP4	PF
1.1	15/12/2022	Added information on WP3	FP
1.2	17/12/2022	Added information on WP2	WP
1.3	22/12/2022	Added information on Financial Summary	IR
1.4	30/12/2022	Final Arrangements for submission	RP

Document Quality Control			
Version QA	**Date**	**Comments (and if appropriate reason for change)**	**Initials of QA Person**
1.3	22/12/2022	Review and edits	AS
1.3	27/12/2022	Review and minor edits	RP

Document Authors and Quality Assurance Checks		
Author Initials	**Name of Author**	**Institution**
RP	Rui Prada	INESC-ID
MC	Marta Couto	INESC-ID
FP	Fernando Pastor	UPV
TV	Tanja Vos	UPV
BM	Beatriz Marín	UPV
FK	Fitsum Kifetew	FBK
PF	Pedro M. Fernandes	INESC-ID

WP	Wishnu Prasetya	WP
JD	Joseph Davidson	GA
JC	Jeremy Cooke	GWE
IR	Ilda Ribeiro	INESC-ID
AS	Angelo Susi	FBK

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This deliverable presents the third and final project report. WP1 is a management package, therefore, the deliverable focuses on management and the project's overall progress. Since this is a follow-up from the previous reports, some information provided here overlaps with the last deliverables as this document was created to be read as a stand-alone document.

1. OVERVIEW OF PROGRESS

1.1 OBJECTIVES

The iv4XR project aims to build a novel verification and validation technology for Extended Reality (XR) systems based on AI techniques to provide learning and reasoning over a virtual world. With this technology, XR developers can deploy powerful test agents to automatically explore and test their virtual worlds' correct parameters as they iteratively develop and refine them. Given the importance of user experience for XR systems, we will also develop socio-emotional AI to enable test agents to conduct an automated assessment of User Experience.

1.2 EXPLANATION OF THE WORK CARRIED OUT BY WORK PACKAGE

1.2.1 WORK PACKAGE 1

WP1 is responsible for all project management activities, covering the project's execution, monitoring progress, controlling the budget, managing risks, coordinating reviews, the consortium agreement, communication between the partners, and liaising with the EC. To facilitate this, we created a management structure at the first consortium meeting, which includes a management board (MB), the quality assurance (QA) team, and the Science and Technology Committee (S&T).

The Management Board meets during the consortium meetings and aids the project coordinator with administrative and financial tasks. We have had seven MB meetings: the first and second meetings took place in person, one in October 2019 at Lisbon and the other in January 2020 at Utrecht. The following meetings were virtual due to COVID-19 and took place in July 2020, February 2021, and June 2021. The 2022 meetings took place in person again in February 2022 at Valencia and in June 2022 at Prague.

The Science and Technology Committee focuses on R&D and executive decisions and meets once a month, using teleconference technology (on Skype), to discuss and solve possible conflicts or problems regarding technical issues, identify necessary changes to the direction in which the project is going concerning technical content development, and share progress. The QA team is responsible for ensuring that the project members deliver high-quality outputs in all project phases. All project deliverables are subject to quality control by at least one member of the QA team. We have a series of channels to facilitate communication between the project's members. We have a Slack channel for all members, mailing lists for all members, for the S&T Committee and the Management Board. To increase communication between the members, the project's coordinator also promoted the creation of working groups. Each WP coordinator sets up monthly meetings to discuss their progress and foster collaboration between the different partners. The meetings are open to all iv4XR members who want to join. The COVID-19 pandemic imposed constraints on most partners. We requested an amendment to extend the project for three months. The amendment was approved in July 2020.

The project's second year was focused on meeting the second project milestone and reach an intermediate framework with testing agents with mature capabilities and running in the case studies.

WP1 coordinated the preparation of the first project review which took place online on September 17, 2021.

All deliverables for this second year were submitted on time. We asked for a 6-month extension on the market study deliverable (D6.5), and this has been approved by the PO. The reviewers asked for more details on two deliverables (D3.1 and D5.2) the modified deliverables were submitted in January (D3.1) and February (D5.2) 2022.

During the third year, we worked towards our 3rd and 4th milestones. MS3 is the full iv4XR framework with all the features completed. As part of MS3 we also have to evaluate the iv4XR solution using the pilots. That work can be read on D5.4. MS4 is an extension of MS3 and WP1 will ensure its completion by ensuring submission of all the remaining deliverables. We are also tasked with coordination of the final project evaluation. We organised a meeting at Utrecht in November 2022 to prepare the project review and discuss the project's future exploitation and final dissemination activities.

1.2.2 WORK PACKAGE 2

WP2 is responsible for the construction of the infrastructure/core part of the iv4xr Framework, also called the **Framework Core**. This includes, for example, basic agents, the agent runtime system, different types of assertions for checking during a test run, and an interface scheme with the XR system under test. Together with WP3, WP2 also develops a Domain Specific Language (DSL) to program the agents. The DSL supports tactical- and goal-based programming. We implemented an advanced information collection scheme that allows not only test verdicts (pass/fail) to be collected, but also execution traces. With the latter, state information during an agent run can be written to a trace file. Trace files are very useful, e.g. it enables runtime verification[1] and visualization of the data in the traces. Linear Temporal Logic (LTL) based runtime verification is supported by the Framework. Basic visualization scripts are provided, or else one can use a generic data science library like Pandas to process the traces (they are in the generic csv-format). More on the Framework Core can be found in the Report **D2.4**.

During the first two years of the project WP2 has focused on the development of various components of the Framework Core. In the final year we focused on improving and testing the components of the Core and supporting and coordinating the integration of modules produced by other Work Packages into the final iv4xr Framework.

- Improving and testing components of the Core is mainly done by UU. This includes e.g. work to improve and extend the aforementioned Domain Specific Language (DSL) for programming the agents. E.g., It now allows dynamic goals to be more easily programmed. We also extend LTL to allow spatial areas and aggregation (e.g., average and coverage) to be expressed and mixed with traditional LTL formulas. A Nethack-like game called MiniDungeon has been built too, to facilitate faster testing of the DSL constructs as well as features from WP3 (online goal solving) and WP4 (user experience testing). MiniDungeon is implemented natively in Java with tight integration to the Core, so features from WP3 and WP4 can be rapidly tested using a simple deployment setup (without needing a bigger setup such as using the Lab Recruits demo game).
- We support and coordinate the integration of explorative agent (WP3), model based testing (WP3), reinforcement learning environment and reinforcement learning based

[1] Contrary to its name, "*runtime verification*" does not refer to verification while a program is running. Rather, it is verification based on extracting information from a running system. The approach has been researched, with various approaches have been developed.

testing (WP3), and affective testing modules (WP4). Partners involved in this integration work are UU, INESC-ID, UPV, FBK, THALES-SIX, and THALES-AVS.

- We support the deployment of the Framework Core for final year case studies validation in WP5. Partners involved: GoodAI, UU, UPV, FBK, GW, THALES-SIX, and THALES-AVS.

1.2.3 WORK PACKAGE 3

Work Package 3 focuses on the development and integration of Functional Test Agents (FTAs) intended to verify if the behaviour of the XR systems responds correctly according to the interaction of users. The final integration of the FTAs with the Systems Under Test (SUT) of the WP5 use cases allows the observation of the XR environment through the WOM, the navigation to reach and interact with the virtual entities, and the verification of the correctness of the functionality. Deliverable 3.5 - “Report describing Functional Test Agents (FTAs)” includes a detailed explanation of the final agents resulting from WP3.

The implementation of a **Test Specification Language (TSL)** in *Task 3.1* allows developers to declaratively express a testing task. This specification language is used to indicate to iv4XR Functional Test Agents (FTAs) what entities and properties to test and which goal structure to follow to achieve the testing goal objective (see D3.1 - “Test Specification Language”).

In the iv4xr framework, there exist multiple types of FTAs that use different approaches to interact and test XR systems:

1. **Goal Solving FTA** (*Task 3.2*)
   These agents follow goal structures expressed by using the TSL developed in Task 3.1 to deliberate about which strategies will allow them to navigate and interact with different entities to solve their goals. Stakeholders can formulate structures of tasks/tests for multiple scenarios, and these agents will gradually complete these tasks to verify that the system evolves in the correct states.

2. **Reinforcement Learning FTA** (*Task 3.4*)
   These agents execute tactics and goals to interact with the XR environment without following a predefined structure. The objective is to gradually learn from the executed interactions to complete the sub-task that allows them to verify that it is possible to achieve a final state.

3. **Scriptless Exploratory FTA** (*Task 3.2*)
   These agents do not follow specific instructions to interact with the XR systems. The objective is based on the execution of non-sequential actions to test that the SUT and its functional aspects are robust enough to respond to multiple and unexpected user interactions.

XR systems may contain **hazardous entities** (*Task 3.3*) that can sabotage an agent's progress or even cause it to fall into an inescapable stuck state. Solving goals to test specific tasks, means that dealing with hazardous elements will not be very different from dealing with other XR application elements. Consequently, FTA agents can handle these hazardous elements by deliberating which strategies will allow them to navigate and interact with different entities to solve their goals.

Task 3.4 is intended to measure how good FTAs are in exercising the SUT and to develop test generation strategies that increase the **coverage** obtained from the tests. Several approaches can be employed to measure the space of interactions, or the code statement coverage reached by the FTAs in the SUT. Hence, there are several notions of coverage.

The development of a model-based testing (MBT) tool allows us to abstractly represent the range of possible interactions to be carried out in the SUT. Based on these models, **state and transition coverage** is used to measure the space of interactions reached by the agents in the systems, which try to cover all possible targets of the model.

The agents have the capability to store the information relative to the floor areas and the positions and types of functional entities as they navigate, observe, and interact with an XR system. This agent's data obtained at runtime can be used together with the scenario information to obtain **spatial and heat-map coverage** metrics that indicate which space was explored and which entities interacted.

The agent's actions in the virtual environment execute the internal methods that compose the code software of the XR systems. It is possible to integrate open-source tools within the iv4XR framework to obtain **code coverage** metrics as the FTAs test the SUT.

The iv4XR framework allows **multiple agents** to communicate and realize simultaneous interaction with the XR systems (*Task 3.5*). This is an essential feature that benefits all the iv4XR use cases.

First, it enables the possibility of verifying that the system and the entities respond correctly to the interaction of multiple users. Second, it opens the door to integrating distributed approaches, Reinforcement learning (RL) algorithms, and strategies that diversify the agent's workload and speed up the exploratory and training procedures. And finally, it allows socio-emotional agents to consider ways of measuring user aptitudes, such as collaboration.

In the project review, reviewers commented on the absence of **Augmented Reality (AR)** systems. To respond to this comment and demonstrate the inclusion of agents for AR, we collaborated between UPV (leading partner for WP3), UU (leading partner for WP2), and INESC-ID (leading partner for WP4) to create an AR demo.

Once all these capabilities were developed for the FTAs, an overall **integration** was done (*Task 3.6*), together with WP4 socio-emotional agents, to integrate all iv4XR agents in the WP2 core framework, which allows the verification and validation of WP5 use cases.

1.2.4 WORK PACKAGE 4

Work Package 4 focuses on developing socio-emotional test agents (SETAs) to aid the systematic assessment of user experience of XR systems while minimizing the manual effort. We currently have two implementations of the SETA integrated with the framework (task 4.1), one based on the two-dimensional core affect theory, modelling the pleasure and arousal emotional dimensions (PAD), and one based on the OCC theory of emotion. We have decided to pursue these two different implementations as each has its own advantages for different testing and developmental scenarios. The OCC model is better adapted to event-based systems whereas the PAD dimensional model is better adapted to continuously evolving systems, as well as being a better fit for machine learning models.

Other than the two emotional models, WP4 has conducted studies to build testing agents capable of simulating other components of user-experience during interaction with XR systems, including:

1. Testing a model to predict the cognitive load imposed by a game. This is the first step towards developing a model that can tell designers the amount of cognitive load that their systems are imposing on the user.

2. The difficulty progression of a series of levels can have a significant impact on the user experience and learnability of a game. We are developed methods to rank a series of levels in terms of the difficulty.

3. Exploring the use of clustering techniques to classify game-play traces in terms of play-style personality and using Inverse Reinforcement Learning to create agents that play according to those different play-style personas.

4. A model to predict cybersickness in immersive VR experiences. The model was built to predict the presence of CS in slices of 1 minute of interaction given a video of the interaction and the set control inputs a user would make.

5. A model to test interactive storytelling, by exploring the potential interaction graph considering changes in internal variables that guard transitions between story nodes. The output presents the distribution of story nodes visited and the dynamics of the internal variables.

The two implementations of the SETA have already been integrated with the iv4XR Framework (Task 4.5) and used to test several maps of the Lab Recruits game. Both emotional models being used for the SETAs, the OCC and PAD, have been implemented with temporal progression in mind, and we pursued a definition of coverage for UX (task 4.3), which guides the development of self-motivated SETA.

WP4 meets on the last Monday of the month to discuss the socio-emotional test agents (SETA). Partners present the ongoing work and share research plans or results, which are then discussed with the attendees (these discussions have been linked to tasks 4.1 and 4.2). The second part of the meeting is usually about test coverage and what it means for SETA (affective coverage - task 4.3).

In this last year we have focused on integrating the different modules into a User Experience (UX) model. To achieve this goal, we created a framework for automated UX testing (more details can be seen on D4.4). In this UX framework we have a modular structure with clearly defined variables aimed at facilitating the definition of automated testing agents, which will in turn facilitate moving from user tests to automated tests. We moved on to the last integration phase (Task 4.5) where we integrate software modules developed in this WP into the iv4XR Framework from WP2. This work was carried out in collaboration with UU (leading partner for

WP2) and GoodAI (leading partner for WP5). We also bundled and documented the Application Programming Interfaces (APIs) of these modules to become part of iv4XR SDK. Alongside integration, we continued the work on the emotional prediction modules and behavioural modules, as well as creating a module for the automatic prediction of motion sickness in VR experiences. We also developed a novel linear temporal logic language that allows designers and testers to define UX goals and constraints, a key requirement for truly automated UX testing.
Details of the integration and agents can be found in the documents D4.3 and 4.4.

INESC-ID also hosted one of this year's students visit. We welcomed one student from UU and one student from UPV for a hackathon focused on advancing the AR application and its test suite and the harmony between the different UX testing modules of WP4.

1.2.5 Work Package 5

Work Package 5 focuses on the preparation of pilot programs by the industrial partners in order to showcase how the iv4XR framework can be used. These pilots are a 3D game, a simulation scenario depicting the infiltration of a nuclear power plant, and a sensor suite for the monitoring of civil engineering projects.

The work of WP5 was organised from the outset into three tasks:

- Task 5.1 - "Preparing the Pilots" where the pilots modifications were made to the pilot system in order to integrate them into the iv4XR framework.
- Task 5.2 - "Pilot Environment" which was about building the testing scaffold for the pilots so that an iv4XR test suite can be executed with minimal human intervention.
- Task 5.3 - "Demos" which was about packaging the pilots, or parts of them into a form suitable for the public.

In addition, WP5 contained D5.4 which was an evaluation of the efficacy of the resulting interfaces, agents, and tools that were developed over the course of the project.

The main body of technical work in the project that is covered by the above tasks was mainly completed in the first two years. The pilots were refactored and interfaces were developed for integration into the iv4XR framework. During this, the information requirements of the agents informed the structure of the interface, and the need of the pilot tests informed the capabilities of the agents.

Once D5.3 “Full Integration” was submitted, the work mainly shifted towards integrating iv4XR based testing into the workflows of the industrial partners. For GoodAI, a series of regression tests were identified and work began on automating those tests via scripts. At the same time, the plugin was expanded with other features like UI interaction and navigation to increase the capabilities of the TESTAR exploratory agents and model based testing agents. We worked closely with the teams at UPV and FBK to realize their testing methods in Space Engineers.

For Thales AVS, we maintained a support to our partner Thales SIX when they used the Reinforcement Learning tools integrated in the Framework to test our pilot. This support mainly consisted in adding new functionalities helping to run MAEV and to get the necessary data to perform the learning. Thales AVS also participated in the definition of the validation tests with Thales SIX and exploited the results to continue modifying our simulation means and designing the requirements that will allow, in the future, the use of AI algorithms to develop the Testing Agents that will be integrated in the development for our customer projects.

For Gameware, we integrated the work done for iv4XR, including our java tool, multi-agent script generator and processor etc, into the LiveSite system. LiveSite now offers enhanced diagnostics using the iv4XR tools developed, which analyses and processes all projects in real-time on the LiveSite servers. Detected warnings are added to the LiveSite warning system, and detected errors are sent as notifications to project recipients via email.
The integration has been running for many months now, and successfully detects warnings and errors which would otherwise have not shown up.
We are working on further enhancements to increase the efficiency of the detection, in particular looking at detecting errors which may occur in the future based on the trends of data in a project.

Deliverable 5.4 is the Project Validation Report which details for each pilot how the iv4XR framework and the agents have been used and evaluated.

1.2.6 Work Package 6

WP6 focuses on the communication, dissemination, and exploitation activities. WP6 have been executed according to the plan:

- The website is in place https://iv4xr-project.eu/ (D6.1)
- The data management plan has been delivered (D6.2). It includes an assessment of the various data generated and maintained in the project by all the partners and outlined the project's management plan with respect to the storage and dissemination of the various data produced in the scope of the project.
- The 1st Dissemination and exploitation plan describes more detailed plans for dissemination and the results for the first year for Tasks 6.1, 6.2 and 6.3 (D6.3).
- Several dissemination and communication activities (Task 6.1) have been undertaken in the second year, as planned in D6.3:
    - Organized a public workshop on the Future of XR, in collaboration with the ARETE project.
    - Organized the 12th A-TEST workshop on automated testing, as planned in the dissemination and exploitation plan.
    - Participated in the VRDays Europe Immersive Tech Week event and presented ongoing work.
- The 2nd Dissemination and exploitation plan has been delivered (D6.4). The 2nd Dissemination and exploitation plan updates the information of the first plan and provides details of the cumulative results obtained in the first and second year for Tasks 6.1, 6.2 and 6.3. Moreover, we start working in exploitation.
- A white paper that presents the main problem addressed by the iv4xr project, the solution proposed and the exploitation and valorisation opportunities with the corresponding business models was delivered during the third year of the project. This white paper helps us to better understand the market and exploitation activities.
- To deal with some delays in the exploitation activities and plans, T6.2 leadership has been assigned to GWE and fortnightly conference calls have been instigated to review progress within the task towards deliverable D6.5. In order to facilitate the market research requirement a questionnaire has been created and shared with agreed key XR market sectors: Aviation & Automotive, Education & Training, Games, Construction, Software Design & Testing, Healthcare, Retail, Manufacturing & Logistics. More than 300 surveys have been sent out to targeted individuals and 34 responses have been received. These qualified respondents input data was assessed to inform market opportunity via D6.5 (Market research report). The work carried out on task 6.2 was a joint effort between all partners who shared the questionnaire with their contacts. The questionnaire was created by GWE (task leader) in collaboration with UPV and INESC-ID. The questions and study plan were shared with all partners in the Valencia consortium meeting to gather their input.

- Regarding the scientific works, we exceed the KPIs related to conference and workshop publications since we have 24 already published works. Regarding the scientific publications submitted to journals, in the computer science community the time for reviewing and accepting papers is around a year, and published papers traditionally correspond to fully packaged research, i.e., which contains the problem, the solution, the implementation, the use in practice and the empirical validation. This situation provokes that it is more difficult to have papers published in journals since it is necessary to have more mature research. During the second year of the project, 2 scientific papers were published in peer-reviewed journals. During the third year of the project, we submitted 5 papers to peer review journals and also, we plan to submit other 5 publications during the first semester of 2023. Taking into account the reviewing time of journal publications, we consider that the KPI of submitting 6 journals has been accomplished.
- Several dissemination and communication activities (Task 6.1) have been undertaken in the third year, as planned in D6.4:
    - Organized the second version of a public workshop on the Future of XR, in collaboration with the ARETE project.
    - Research of the market for the outcomes of the iv4xr project.
    - Organized the 13th A-TEST workshop on automated testing.
    - Participated in the VRDays Europe Immersive Tech Week event and presented ongoing work.
    - Increase the submission of scientific articles to journals.
    - Publish the project in non-scientific media to address a wider audience.
- The 3rd Dissemination and exploitation plan has been delivered (D6.6). The 3rd Dissemination and exploitation plan updates the information of the first and second plan and provides details of the cumulative results obtained during the three years of the project. It also provides details about the exploitation and sustainability of the results of the project.

All partners were committed to improving dissemination and exploitation activities. We created a working group to have more frequent discussions on the dissemination and exploitation plan. We had several joint discussions at the consortium meetings about dissemination and exploitation to come up with a consensual and feasible plan that we will implement beyond the lifetime of the project. To support the future exploitation of the project's results a webpage for the iv4XR toolkit was setup at https://iv4xr-project.eu/toolkit and a new domain iv4xr-toolkit.eu was purchased.

1.2.6 WORK PACKAGE 7

Work Package 7 has been included to address ethical questions and guideline development of the project. Its main goal was to develop deliverable D7.1 - “POPD - Requirement No. 1” and assure that its principles and guidelines are followed.

We created the Ethics Advisory Board (POPD requirement No.1) to maintain information on all studies conducted with humans within the project. The Ethics Advisory Board (EAB) is responsible for ensuring that all studies that involve humans are submitted to an Ethical Review Committee (ERC) for approval. The EAB will also help the partners compile the required materials for ERC submissions, inform all partners about the decisions, and help with any modifications requested by the ERC's. This Advisory Board will also be tasked with assisting all partners in complying with the principles defined in D7.1. The user studies performed so far were submitted to the ERC.

2. FINANCIAL SUMMARY

At the moment, we are still gathering information related to the final costs, so all information shared in this report is merely an estimate. We will include a detailed financial report in the technical report due February 2023.
From the previous reporting period, we had some deviations from the original plan, namely related to travel costs. The Covid-19 pandemic imposed travel restrictions, and our project meetings had to be held online. Although during this last reporting period, we were allowed to travel and meet in person and participate in events, that did not compensate for the underspending of the previous period.
To address reviewer feedback, we changed PMs to assign task 6.2 to GWE, and we increased work efforts on work packages 2, 3 and 4 to develop studies focused on AR.

A detailed financial report with the final information will be included in the technical report.

3. CONCLUSIONS

This deliverable provides an overview of the work carried out in each work package throughout the different reporting periods. As stated in the executive summary, this report includes previous reporting periods, and it is meant to be read as a stand-alone document detailing the work of each work package.

During the third year, we aimed to complete the planned work and tried to address the comments made on the first project review.

The other deliverables submitted and referenced in this report provide the details of our work.

The outputs from the project can be found on Zenodo[2], Github[3], Project Website[4] and on Social Media: Twitter[5] and Facebook[6].

[2] https://zenodo.org/communities/iv4xr-project/?page=1&size=20
[3] https://github.com/iv4xr-project
[4] https://iv4xr-project.eu/
[5] https://twitter.com/iv4xr
[6] https://www.facebook.com/iv4xr