DATE

2026

AUTHORS

Ine van Zeeland[1]
Ana Pop Stefanija[1]
Rob Heyman[1]
Enola Constanceau[2]
Aline Roc[2]

1. imec-SMIT, Vrije Univestiteit Brussel
2. CATIE (Centre Aquitain des technologies de l'Information et Électronique)

TITLE

Collaborating with AI users in early stages of AI development

CONTACT

peer-ai.eu

peer-ai Peer_Ai_ @PEER-AI

POLICY BRIEF

The hyper expert collaborative **AI** assistant

peer

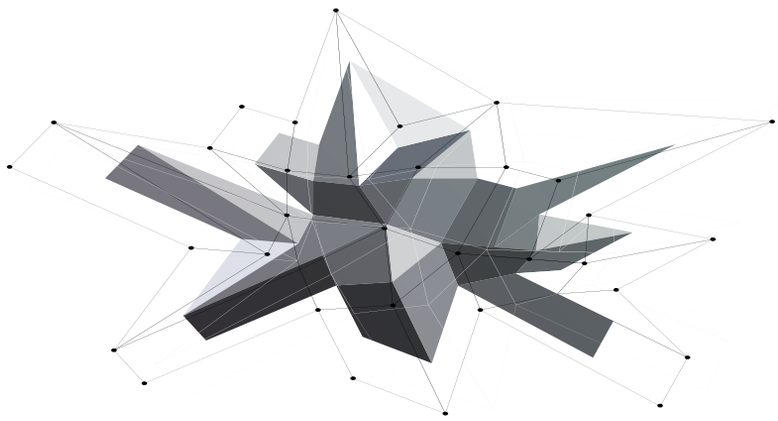
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101120406.

INTRODUCTION_

This policy brief details an approach to involving users in the early stages of AI development.

00

In response to the problem of explaining hypothetical or incomplete AI systems to potential users, necessary for their meaningful involvement, the PEER project introduced and tested low-fidelity simulation tests. The policy brief concludes with specific recommendations for user tests in early stages of AI development.



Redefining Human-AI Collaboration for Complex Decisions

Current AI solutions tackling sequential decision-making often lack flexibility and user integration, hindering their real-world impact. PEER addresses this challenge by prioritizing the user throughout the entire AI lifecycle.

PEER is not just building AI assistants; it's trying to redefine the way humans and AI interact.

ABOUT THE PROJECT_

Why PEER is different?

Accessible Design: Demystifying AI capabilities and limitations for stakeholders throughout the development process, maximizing transparency and informed decision-making.

Collaborative Problem-Solving: Enabling bidirectional communication between AI and users, creating a dynamic feedback loop that enhances user engagement and refines outcomes.

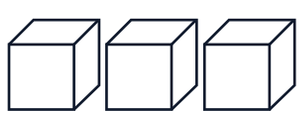
Human-Centric Metrics: Moving beyond traditional AI performance indicators, PEER develops qualitative and quantitative measures that assess interactivity, acceptance, explainability, trustworthiness, and perceived fairness, ensuring AI aligns with human values and needs.

Key objectives

Enhanced User Acceptance: Building trust through transparency and collaborative decision-making, encouraging broader adoption of AI solutions.

Optimized Outcomes: Leveraging the strengths of both humans and AI for improved problem-solving and more effective decision-making.

Responsible AI Development: Establishing reliable, transparent, and trustworthy AI frameworks, ensuring ethical and socially responsible deployment in real-world applications.



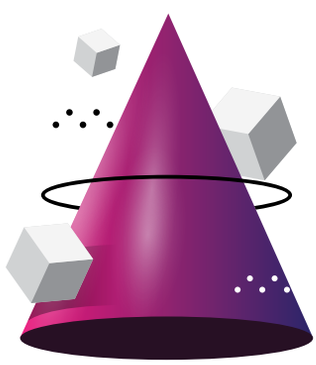
Why include users in AI development?_

Over the course of AI system development processes, one element that is often overlooked are the intended end users of the system — their needs, their visions, values, and requirements regarding the system. Often an afterthought, it is not uncommon for end users to be invited to test and provide feedback on the AI system only after it is (almost) fully developed. Intended users may imagine a system to behave in ways that are not technically feasible, or what deployers want a system to achieve may not be what intended users need. After all, the values and purposes of different stakeholders — developers, deployers, regulators, users, affected parties — may clash, and not all can prevail. This results in a need to adjust different stakeholders' expectations of the system, its technical implementation, its features, or the goals it can achieve, preferably throughout the AI development process. If this is left to later in the process, it might be too late.

Co-creation is often touted as a potential solution to prevent or mitigate such frictions. Involving intended users early in AI development through co-creation allows them to explore the capabilities and limitations of an AI system under development. For example, a city administration developing an inclusive city route planner could invite a diverse group of citizens, including people with reduced mobility, to a workshop in which they sit side-by-side with developers, city planners, local advocacy groups, and similar, to brainstorm, sketch scenarios of use, use model cards to imagine a technology, based on their own interests, needs, and experiences. A co-creation process thus creates an open space for negotiation between different interests, possibilities, and values.

Early user involvement in system development is crucial for a number of reasons, among which the **acceptance of an AI system** and its actual use after implementation we consider the most important. If included in early development, users can provide critical input that enables improving and correcting the system while the window of opportunity is not yet closed. Another reason, from a business perspective, is **reducing reparation costs** of a system's shortcomings, compared to having to retrace design in later stages. Involving end users in co-design in later development stages as well allows for harmonious integration of different stakeholders' requirements.

The problem with co-creation in AI development, however, is often **how to explain** what a system may do or achieve to potential non-expert end users, lacking advanced AI literacy or technology development skills. For instance, explaining the concepts of machine learning, algorithmic inferences, and sequential decision-making to a group of everyday supermarket customers may be quite a challenge. Differential skills have time and again proven to be a major barrier to involving potential users in the input, throughput, or even early output stages of AI system development. **The PEER project**[1] aims to tackle this problem through low-fidelity simulation tests.



Why include users in AI development?_

Within the PEER project, AI systems are developed in concert with intended users, who are active participants in the design and development process. While the AI systems developed within **PEER and their goals differ**[2], they share two common components: the development and evaluation of hyper-expert collaborative systems for sequential decision-making, and a focus on the development of human-centric AI systems through co-creation. Structured around the interaction between the user and the AI system/assistant, the PEER assistants should closely follow user preferences and constraints, by adapting dynamically and responsively. A core goal of the PEER project is building an AI system tailored to user needs, ensuring and enhancing user trust, acceptance, and collaboration, while also advancing the state of the art.

We achieve this by including potential end users from the beginning of AI system development, ensuring continuous input, evaluation, and validation. We do this by inquiring about the socio-technical requirements, conducting system simulation testing, and continuously testing the systems for achieving trustworthy AI. A number of methods and tools are used to do so: guidance ethics, model cards, think-aloud, simulation and evaluation frameworks, and a survey using the AI Acceptance Index.

While this is an important element of the project and highly relevant for wider use in society, this policy brief is centred around the parts of the user tests that focus on system requirements.

Proditec's workshop for task T2.1 of the PEER project

1 PEER stands for The Hyper Expert Collaborative AI assistant. Learn more here: https://peer-ai.eu/en/

2 The systems are developed for four different use cases: a navigation system for wheelchair users in the City of Amsterdam, a tool to support visual flaw detection in inspection systems (Proditec), a warehouse system for efficient and safe storage of dangerous goods (Datacation), and a navigation system for customers and delivery service agents in Portuguese hypermarkets (SONAE).

peer

02

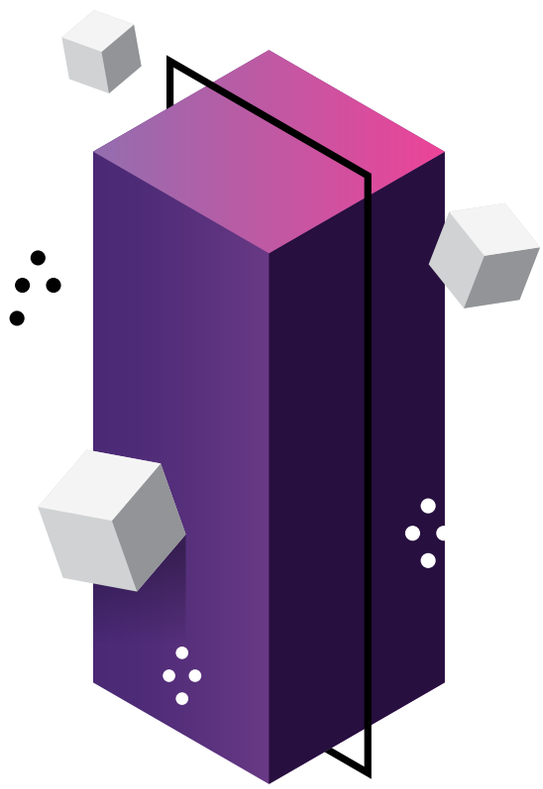
Low-fidelity simulation testing_

Research in human-computer interaction suggests an approach to involving users early in development, when a system is still in a design phase: simulation testing. Simulations can vary from very low resemblance to real-life conditions of use (low fidelity) to very realistic circumstances, such as pilot prototyping (high fidelity). For example, in a low-fidelity scenario, users may test a route-planning system by responding to drawings of its interface and functionalities, while in a high-fidelity scenario, they may interact with the system via a smartphone app while navigating a city. As fidelity is a scale, it can be set at different levels, more or less resembling the real-life circumstances in which the system will be used.

The purpose of simulation testing is not to test interface design but to ask actual users to perform tasks by using a prototype of the system. Low-fidelity prototyping is generally less costly than simulations with higher levels of fidelity, but there are more important drivers for this test approach. Any user test starts with why and what questions that form the basis of the test strategy. Low-fidelity approaches should provide answers to **three basic questions:**

1. **What is the purpose of this user test?**
2. **Who are the envisioned end users with whom we should test the system?**
3. **What is the possible and necessary level of fidelity of the prototype (at the lower end of the fidelity scale)?**

End users will rarely interact with systems in the ways developers expect. Instead, users attempt to retool designs to serve their own purposes. Testers must therefore prepare different scenarios to test interactions. Scenarios are tasks that require the use of the AI system to be solved, but for which different paths to solutions are possible, leaving space for each user's own path.



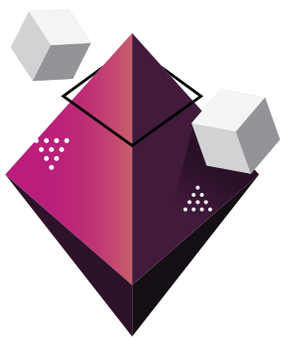
EXAMPLE SCENARIO_

In one use case within the PEER project, participants were given the following scenario: Imagine you want to do your grocery shopping at a [...] supermarket. Today, you have decided to use the new version of the [new] application, which helps you plan your route through the store based on the products you have on your shopping list. As a first step, you want to buy some products for your groceries.

02

Summary of steps, key questions, and actions for simulation testing_

STEP	KEY QUESTIONS	ESSENTIAL ACTIONS
1. Define purpose	Why test? Why at this stage? What are the goals?	List specific questions to answer • Draw up scenarios
2. Identify users	Who will use the system? Which different purposes do they have?	Define target user groups • Plan representative recruitment
3. Set fidelity	What to include or exclude? Current capabilities? Which specifications are known?	Find manageable fidelity level • Assess feasible simulation scope

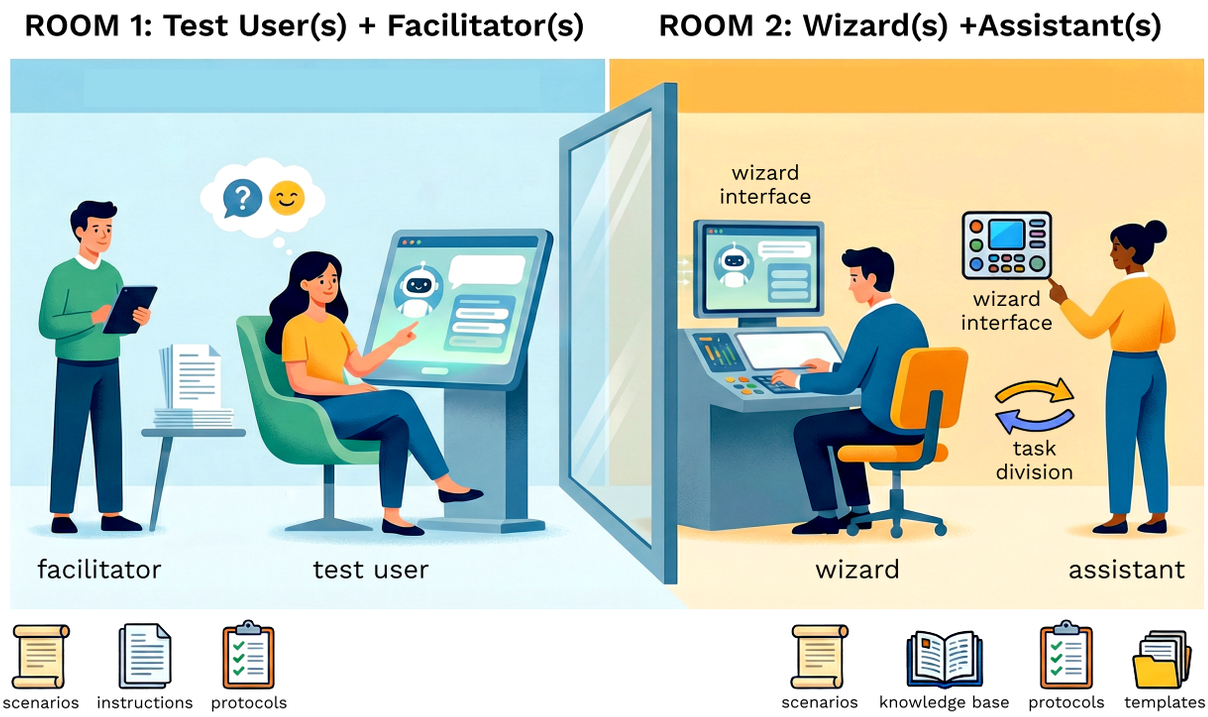


03

Paper prototyping is the simplest form of low-fidelity simulation testing. When too much of the system is still unclear, or the focus of the simulation test is on functionality rather than appearance, pen and paper can replace more elaborate prototypes. A step up in fidelity is Wizard-of-Oz prototyping. This technique implies telling the users they are interacting with an automated system while, in fact, they are interacting with another person (the "wizard"). It is a computer-based simulation test that can easily be upgraded to higher levels of fidelity. For ethical reasons, a general recommendation for Wizard-of-Oz simulations is to explain the details of the setup to participants, even if only afterwards. Preconditions are that it must be possible to specify some of the system's future behaviour, and it must be possible to convincingly simulate the automated system, given human limitations.

People pretending to be computer systems cannot take the time to type out a reply to a user, or can produce typos. A wizard will have to use canned responses and templates, and ways to quickly retrieve information as needed in the interaction with users. Multiple wizards or assistants may be needed, with clear task allocations, to recreate the speed and accuracy of responses of an AI system.

Visualisation of a Wizard-of-Oz setup



Advantages of Wizard of Oz simulations are that they improve the understanding of the system's functionality for users. Disadvantages are that creating a convincing Wizard of Oz simulation environment for multiple scenarios, training the wizard(s), and pre-empting a variety of user interactions can be costly and time-consuming.

The table below provides a description of the tasks that make up the approach.

Wizard of Oz tasks for each phase of a simulation test

PHASE	WIZARD OF OZ PROTOTYPING TASKS
Preparation	Create scenarios • Build computer interface • Create knowledge base • Train wizard(s) • Test scenarios and setup • Document every step • Prepare analysis
Simulation	Brief introduction • Guide users through scenarios • Wizard responds realistically • Record interactions • Wrap up and evaluate with participants • Document all steps and prepare analysis
Analysis	Process digital recordings • Integrate all data types and analyse coherently • Include interdisciplinary team • Evaluate wizard performance and simulation validity • Report results of simulation and evaluation

Evaluating simulation tests_

Evaluating simulation tests is important to improve subsequent iterations of the AI system design, but also to assess the extent to which the simulation characteristics have affected the outcomes of the tests, and to assess generalizability and representativeness. The user test approach must also be evaluated to assess whether it was indeed suitable for the questions that needed to be answered. The evaluation of a simulation test consists of pre-testing and post-testing phases. A very basic evaluation can be to assess the **face validity** of the procedure on paper: Does the procedure, at face value, reflect the goals of the simulation, and to be robust and coherent? This can also be done by asking outsiders to have a look at the procedure beforehand.

Important pre-tests:

Testing the scenarios for consistency, accuracy, and realism;

Testing the interfaces and knowledge base;

A trial of the full simulation procedure with pretend users (such as colleagues).

Most of the evaluation, however, will take place after the simulation test. An important part is assessing the **validity and reliability**. **Validity** refers to the absence of systemic errors, such as peculiar results caused by structural errors in the procedure. The results can be considered valid if the simulation procedure was carried out without systemic errors. Specific aspects of validity are **ecological validity** (how realistic is the simulation set-up?) and **internal validity** (which inferences can we draw from the findings?). Validity also refers to the **generalisability** of the results: Can the findings in the simulation with these particular participants and in this particular setup be generalised to the larger population of envisioned end users and use cases?

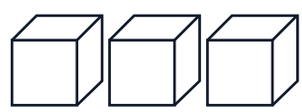
Reliability refers to the absence of accidental or occasional errors.

Its assessment includes:

Stability: is the measurement stable over time?

Internal reliability: is the scale or index of measurement consistent?

Inter-rater reliability: are different analysts conducting the analysis consistently?



Another form of evaluation is to involve participants in assessing test findings. Such **respondent validation** can provide nuance to conclusions or confirm understandings of the analysts.

Who is invited to test a simulation is very important for the validity of simulation testing. There is not always a complete view of potential users. Sometimes unexpected users emerge, such as delivery pickers next to individual hypermarket customers. For mobility assistance, emerging users can be people who temporarily need to use such assistance. Nevertheless, selected participants should closely resemble intended users, and proxy users should be a last resort. For example, the lack of job-specific experience and insights of proxy users can lead to biased test outcomes if the envisioned AI system is to be used in a work setting for work-related tasks.

Participation should not be a façade. If no real changes to the systems under development are introduced after the engagement of users in the simulations, participation may take the form of informing, consultation, or advice, without real follow-through. For meaningful user involvement, the simulation should enable true participation and lead to sincere consideration of the feedback and evaluation of the participants, and their inclusion in the envisioned system.

Conclusions and recommendations

This policy brief elaborated the purposes and benefits of involving users in AI development through simulation testing, starting early in the development process and engaging potential end users in co-creation processes. The goal is to ensure that the development of AI systems is aligned with end users' needs and requirements. Based on the PEER project research and experiences, we conclude with the recommendations below.

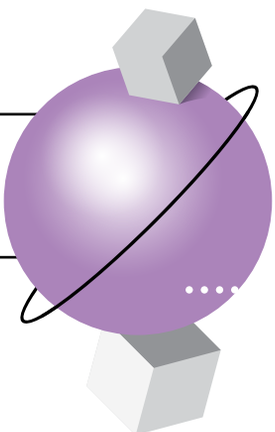
1 Meaningfully involving end users to develop trustworthy AI systems that are accepted in practice entails choosing representative participants and true involvement of end users beyond tokenism.

2 Prepare an evaluation protocol: methods used (observations, interviews, questionnaires, think aloud, etc.), choice of measuring metrics, duration.

3 Prepare participant recruitment: targeted profiles, how much access you (can) have to the end-users.

4 Choose prototype realism: define what features are simulated, what features are developed by plan a user flow.

5 Carefully evaluate simulation tests to ensure the validity and reliability of the results, potentially including the users' own evaluations.



peer

peer-ai.eu

FOLLOW THE PROJECT

peer-ai Peer_Ai_ @PEER-AI

SUBSCRIBE TO
OUR NEWSLETTER

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101120406.