

"THE HYPER EXPERT COLLABORATIVE AI ASSISTANT"

Deliverable D5.6

First report on the evaluation process

Version	2.0
Lead partner	CAT
Dissemination level	Public
Type of deliverable	Report
Contractual date	30/12/2025
Actual delivery date	28/11/2025

History of changes:

Version	Date	Changes	Pages concerned
V1.0	Submitted on 30/09/2025	First version	n/a
V2.0	Submitted on 28/11/2025	Second version included elements of Datacation use-case	p11, 22-26, 40-49, 53-55, 59-63

D5.6. First report on the evaluation process

Deliverable description as per DoA.

Authors	**Organisations**
Charles FAGE	CAT
Elise DURNERIN	CAT
Joseph GARDETTE	CAT
Aline ROC	CAT
Enola CONSTANCEAU	CAT

Contributors	
Maike EHLERT	AMS
Cláudia PINHAO	AMS
Princy ANDRIAMPANIRY	PDT
Madalena SOTTOMAYOR	SON
Floris BLEULAND VAN OORDT	DAT
Joep BEURSKENS	DAT

Revised by	
Floris BLEULAND VAN OORDT	DAT
Joep BEURSKENS	DAT
Mehrdad ASADI	VUB-AI

Disclaimer
This document reflects the authors' views only and neither the Agency nor the Commission are responsible for any use that may be made of the information contained

Table of Contents

1.	Executive summary	8
2.	Task/Work Description	9
2.1	Objectives of T5.7	9
2.2	Use-cases: Proditec, City of Amsterdam, Sonae and Datacation	9
2.2.1	Proditec: Human-in-the-loop and Explainable Anomaly Detection Trainer	9
2.2.2	City of Amsterdam: A personalized Navigation App for Wheelchair Users	10
2.2.3	Sonae: A Personalized In-Store Navigation Assistant for Customers (SigaApp)	10
2.2.4	Datacation: AI-assisted warehouse configuration optimization	11
2.3	Preparatory Work	11
2.3.1	Collaboration with use-case owners	11
2.3.2	Creation of the AIA index and the evaluation protocol	11
2.4	MVP evaluation sessions	12
2.5	Methodology of MVP Evaluation	12
2.6	Analysis	13
2.6.1	Users feedback analysis	13
2.6.2	Analysis of AIA index for prototype	13
2.7	Main outcomes	13
2.8	Conclusion and perspectives	14
2.9	Workplan	15
3.	Methodology	16
3.1	Objectives	16
3.2	Evaluation protocol	16
3.2.1	Common base	16
3.2.2	Use-case specifics	20
3.3	Methodology for analysing and processing results	26
3.3.1	Evaluation of the prototypes	26
3.3.2	Evaluation of the AIA index	26
4.	Results	27
4.1	Thematic analysis	27
4.1.1	Proditec	27
4.1.2	City of Amsterdam	32
4.1.3	Sonae	36
4.1.4	Datacation	40
4.2	AIA index	44
4.2.1	AIA index score	44
4.2.2	Patterns in separate AI trustworthiness constructs	44
4.2.3	Skipped items	47

4.2.4 Items answers homogeneity 48
4.2.5 Control variables 48

5. Discussion / Perspectives 50
5.1 Results interpretation 50
5.1.1 Overall results interpretation 50
5.1.2 Focus on the AIA Index 50
5.1.3 Trust judgment without vulnerability? 51
5.2 Perspectives on prototypes development 51
5.2.1 Proditec 51
5.2.2 City of Amsterdam 52
5.2.3 Sonae 52
5.2.4 Datacation 53
5.3 Experimental limitations 53
5.3.1 TRL level and task realness 53
5.3.2 Panel of participants 54
5.3.3 Experiment facilitation and support 54

6. Deviations occurred and corrective actions taken 55
6.1 The fourth use-case was delayed 55
6.2 Fewer participants than expected 55
6.3 Early results presented to speed up prototype improvements 56

7. Conclusion 57

8. Bibliography 58

Annex 59
Annex 1: Score distribution per construct 59
Annex 2: Answer distribution per item 61
Annex 3: Score distribution of control variables 62

Figure 1: Three screens of the Proditec mock-up. 9
Figure 2: City of Amsterdam's mock-up. 10
Figure 3: Sonae's simulator. 10
Figure 4: Datacation's mock-up. 11
Figure 5: The 10 constructs used to assess perceived AI trustworthiness. 17
Figure 6: The experimental session steps. 19
Figure 7: Participants distribution by age category. 20
Figure 8: Participants distribution by socio-professional category. 20
Figure 9: Participants opinions about AI technologies for Proditec use-case. 21
Figure 10: Participants opinions about AI technologies for City of Amsterdam use-case. 22
Figure 11: Participants opinions about AI technologies for Sonae use-case. 22
Figure 12: Participants opinions about AI technologies for Datacation use-case. 23
Figure 13: Proditec user flow. 24
Figure 14: City of Amsterdam user flow. 24
Figure 15: Sonae user flow. 25
Figure 16: Datacation user flow. 26
Figure 17: Proditec mock-up - Homepage. 27
Figure 18: Proditec mock-up - Labelling page. 28
Figure 19: Proditec mock-up - Set expertise page. 28
Figure 20: Proditec mock-up - Set accuracy page. 28
Figure 21: Proditec mock-up - Performance indicator and threshold page. 29
Figure 22: Proditec mock-up - Heatmap page. 30
Figure 23: Proditec mock-up - Success page. 30
Figure 24: City of Amsterdam mock-up - Wheelchair dimension page. 32
Figure 25: City of Amsterdam mock-up - Preference questions pages. 33
Figure 26: Alternative design for collecting information part. 33
Figure 27: City of Amsterdam mock-up - Route visualisation page. 33
Figure 28: Alternative design for route visualisation part. 34
Figure 29: City of Amsterdam mock-up - Feedback pages. 34
Figure 30: Alternative design for feedback part. 35
Figure 31: Sonae simulator environment - Shopping list. 37
Figure 32: Sonae simulator environment - Constraints. 37
Figure 33: Sonae simulator environment - Itineraries. 38
Figure 34: Sonae simulator environment - Itinerary execution. 38
Figure 35: Sonae simulator environment - End of the itinerary execution. 39
Figure 36: Datacation mock-up - Input menu. 40
Figure 37: Datacation mock-up - WMS status menu. 41
Figure 38: Datacation mock-up - Order overview menu. 41
Figure 39: Datacation mock-up - Error analysis menu. 41
Figure 40: Datacation mock-up - Optimization menu - Solutions explanations. 42

Figure 41: Datacation mock-up - Optimization menu - Chatbot. 43
Figure 42: AIA index score per use-case. 44
Figure 43: Scale-level results distribution (and median) for Proditec use-case. 59
Figure 44 : Scale-level results distribution (and median) for City of Amsterdam use-case. 59
Figure 45: Scale-level results distribution (and median) for Sonae use-case. 59
Figure 46: Scale-level results distribution (and median) for Datacation use-case. 60
Figure 47: Distribution of users' responses to Likert-scale items. 61
Figure 48: Distribution (and median) of control items for Proditec use-case. 62
Figure 49: Distribution (and median) of control items for Amsterdam use-case. 62
Figure 50: Distribution (and median) of control items for Sonae use-case. 62
Figure 51: Distribution (and median) of control items for Datacation use-case. 63

Table 1: Summary of deliverables of T5.7. 9
Table 2: Workplan. 15
Table 3: AIA index score per use-case. 44
Table 4: Details on scale-level results per use-case. 45
Table 5: Details of strong skew toward positive rating per use-case. 46
Table 6: Details of strong skew toward negative rating per use-case. 47
Table 7: Details on skipped items per use-case. 48
Table 8: Details of items with homogeneity per use-case. 48
Table 9: Experienced trust and intention to use. 49
Table 10: Situation awareness during the interaction. 49

1. Executive summary

Within the PEER project, the **WP5,** entitled **"Learn: Piloting and Evaluation"** presents **the detailed usage scenarios, integrates the components of the PEER project in the use-cases, and obtain feedback from users.** WP5 contains **Task T5.7 Management of the Evaluation Process**, related to every aspect of the evaluation process.

CATIE is responsible for **T5.7 "Management of the Evaluation Process"**, starting in M10 and finishing in M48. This task encompasses the development of an evaluation protocol of **acceptance levels for each AI (Artificial Intelligence) prototype** from the project, the **actual performance of the evaluation** and the centralization of the results to present before the consortium. This sequence will be performed 3 times during the project, improving the prototypes from TRL 3 (Technical Readiness Level) to TRL 5. Additionally, this task encompasses **the testing of the 1st version of AIA (Artificial Intelligence Acceptance) index to be developed in WP4**, for which CATIE is also responsible. These data will be presented extensively in the deliverable D4.2 which is due in M36.

The present deliverable **D5.6 Report on the Evaluation Process** presents the **methodology, the results and perspectives** on the evaluation of the MVP (Minimal Viable Product) for each use-case. This evaluation will be performed 3 times iteratively to improve the prototypes, from partially simulated (TRL 3, Year 2) to functional in real conditions (TRL 5, Year 4) prototypes.

The results related to the use of MVPs revealed that participants reported sufficient comprehensibility from the prototype to achieve the usage scenario. Noteworthy, **all participants but one was able to complete the task**. However, most participants reported a limited understanding of the overall functioning of the prototypes.

We collected feedback specific to each use-case and elements to improve each prototype as well. **The main improvement avenue is related to the timely provided explanations by the AI assistant**.

We also collected feedback on the first version of the AIA Index. The results related to the AIA index were crucial to improve it for the next evaluation. The constructs focused on in the PEER project – namely **comprehensibility**, **user agency**, **overall trustworthiness** – exhibited satisfactory results given the elements collected through the corresponding items. **We were able to identify avenues for improvements, especially on social-oriented constructs**, such as **cultural aspects**, **ethics**, **privacy and data governance**.

2. Task/Work Description

2.1 Objectives of T5.7

T5.7 pursues two main objectives, from M10 to M48 (Table 1). First, it aims at providing all elements of **evaluation relevant to improving the different versions of prototypes of each use-case**, with a focus on aspects such as human-AI collaboration, comprehensibility, and overall trustworthiness. Second, it aims at **providing elements to improve the AIA index** developed in WP4.

	Task	Deliverable
M10-M48	T5.7 | **Transparent and reliable measurement scales for the evaluation of trustworthy AI.**	D5.6 | **First report on the evaluation process (M24).**
		D5.7 | Second report on the evaluation process (M36).
		D5.8 | Third report on the evaluation process (M48).

Table 1: Summary of deliverables of T5.7.

2.2 Use-cases: Proditec, City of Amsterdam, Sonae and Datacation

According to the Grant Agreement, the evaluation should have involved 4 use-cases. These use-cases are heterogeneous in terms of application domains, users and context of use.

This subsection presents the mock-ups evaluated for T5.7.

2.2.1 Proditec: Human-in-the-loop and Explainable Anomaly Detection Trainer

The goal of Proditec is to improve a coin sorting machine. They suggest the **integration of an AI assistant** into their machine's interface. This AI assistant aims to provide contextual information, guide the operator through the various stages of sorting process and facilitate the resolution of any errors.

An initial **interactive mock-up** was **created on Figma** to visualize the integration of this AI assistant into the user interface. This mock-up illustrates the main screens of the journey (Figure 1), the possible interactions with the AI assistant, as well as the visual elements implemented to support the operator in their daily tasks.



Figure 1: Three screens of the Proditec mock-up.

2.2.2 City of Amsterdam: A personalized Navigation App for Wheelchair Users

The City of Amsterdam's goal is to develop a **mobile application** specifically designed for the **navigation of wheelchair users**. In this application, an AI assistant considers key constraints such as the type of wheelchair (manual or electric), the number of crossings or the accessibility of sidewalks. The goal is to offer personalized, safe and usable routes to facilitate daily travel and improve user's independence.

An initial **interactive mock-up** was created on **Figma** to illustrate the user experience on mobile (Figure 2). This mock-up simulates the entire journey, from setting preferences to displaying the appropriate route.

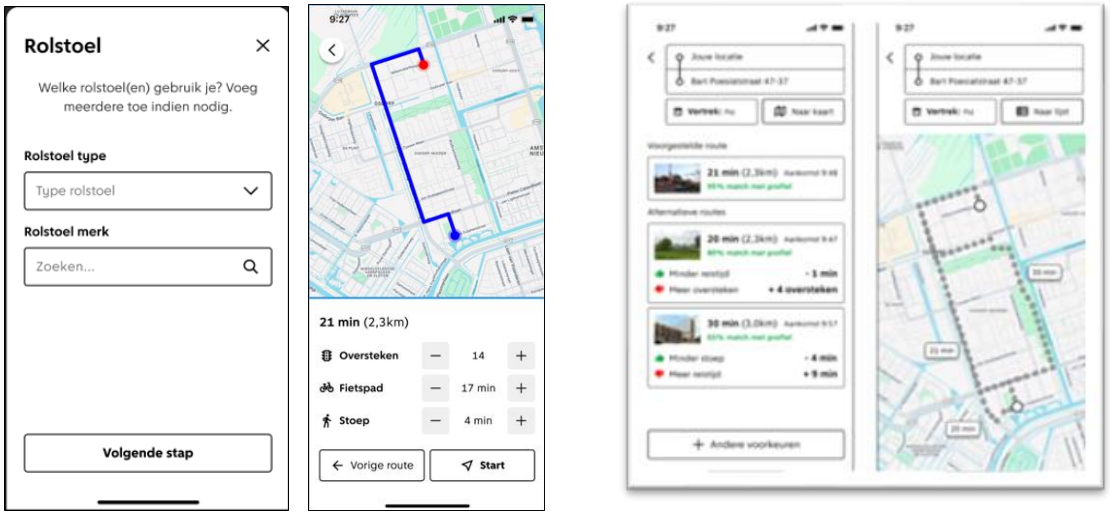


Figure 2: City of Amsterdam's mock-up.

2.2.3 Sonae: A Personalized In-Store Navigation Assistant for Customers (SigaApp)

Sonae's goal is to offer a **route planning solution within a supermarket**, based on the **user's shopping list**. The goal is to facilitate travel by optimizing the route between store aisles, based on various customizable criteria: route type (shortest in distance or time) or order of passage (picking up fresh produce last). This approach aims to improve the shopping experience, reduce time spend in-store, and limit unnecessary trips.

A simulator (Figure 3) was developed to model the behaviour of the planning algorithm (AI assistant) and test different scenarios. It allows to visualize routes generated based on shopping list and defined preferences, as well as updates if there are any changes.



Figure 3: Sonae's simulator.

2.2.4 Datacation: AI-assisted warehouse configuration optimization

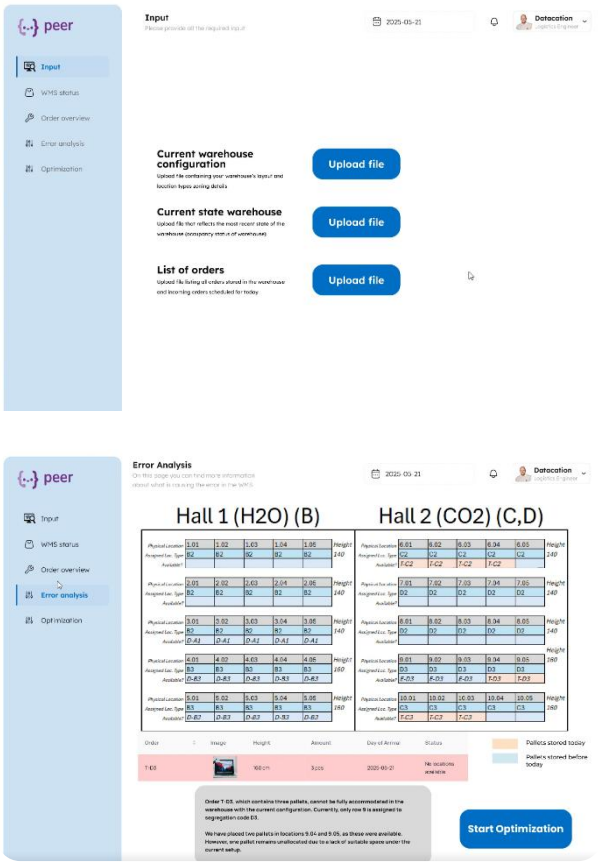


Figure 4: Datacation's mock-up.

The Datacation goal is to provide an **AI assistant system** that helps **logistics engineers and systems experts** in **optimizing product allocation and location type configuration on future orders**. The AI assistant considers some inputs (warehouse's layout, location types zoning details, occupancy status of warehouse, list of all orders stored in the warehouse and incoming order) to propose **solutions suggestion** for re-allocating or reconfiguring the warehouse.

An initial **interactive mock-up** was created on **Figma** to simulate the behaviour of the AI assistant (Figure 4). This mock-up illustrates the visual elements available to logistics engineers to carry out their job, the solutions proposed by the AI assistant and the possible interactions with it.

2.3 Preparatory Work

This subsection presents the tasks performed to conduct the first phase of T5.7.

2.3.1 Collaboration with use-case owners

Throughout the development of MVP 1, **close collaboration was maintained with the use-case owners** to ensure a coherent design aligned with real-world needs. Regular meetings were held to monitor the progress of the mock-ups, discuss functional and ergonomic decisions, and prepare for initial evaluation phases. A key stage in the process was **the in-depth testing phase**: CATIE accessed all mock-ups for detailed analysis. These tests helped identify concrete areas for improvement, and structured feedback was provided to the different use-case owners. Several **design iterations** were carried out until the mock-ups reached a sufficient level of maturity to be evaluated with end-users.

2.3.2 Creation of the AIA index and the evaluation protocol

As a reminder, the main objective of the project is **to build and evaluate various AI assistants**, with a particular focus on user trust. To support this goal, a dedicated questionnaire is developed to assess the assistant's overall **trustworthiness** as a predictor of acceptance: the AIA index.

2.3.2.1 AIA index

The 1st version AIA index includes **around sixty questions** covering several constructs, such as human-AI collaboration, auditability, transparency or data governance. **The current version is a work in progress** (a first prototype version), which will be progressively improved over the next iterations based on user feedback and expert inputs. A detailed presentation of the AIA index as it stands will be provided in the future deliverable D4.2. **The final validated version of this index is scheduled for M36 of the project**.

2.3.2.2 Experiment protocol

A standardized experimental protocol was created for all use-cases to ensure consistency. The protocol includes five main steps:

- **(1)** participant welcome and informed consent,
- **(2)** completion of a task using the corresponding mock-up,
- **(3)** completion of the AIA index questionnaire,
- **(4)** a semi-structured interview divided into two parts – the first focusing on the predictors of trustworthiness and the AIA index, the second addressing the interface and user experience - and finally
- **(5)** a socio-demographic questionnaire.

The sessions were audio recorded as well as the screen of the mock-up.

2.4 MVP evaluation sessions

To evaluate the different AI assistants, we implemented the experimental protocol described in the previous section across the 4 different use-cases. The evaluation sessions were **conducted on-site** at each partner location, allowing us to test the mock-up in context. All evaluations were conducted using Peac²h, our dedicated tool for managing the protocol and **collecting data in a structured and consistent manner**.

In total, 28 sessions were organized at four locations:

- Proditec facility in Pessac (France): 3 days working with 11 company employees.
- Sonae facility in Porto (Portugal): 2 days working with 6 company employees.
- City of Amsterdam (Netherlands): 2 days working with 7 end-users (wheelchair users) recruited by Cliëntenbelang Amsterdam, a dedicated partner of the City of Amsterdam in this project.
- Broekman Logistics Venlo (Netherlands): 1 day working with 4 company employees recruited by Datacation.

Each session lasted **approximately 1.5 hours per participant** and followed the same structure. While the participants in Proditec and Sonae were not representative of the final target users (as they were company employees), they still provided valuable feedback on usability and perceived trustworthiness. In contrast, the City of Amsterdam and Datacation sessions allowed us to collect insights directly from the **intended user group**, enriching the evaluation of the AI assistant's relevance and trustworthiness.

2.5 Methodology of MVP Evaluation

The overall methodology adopted by CATIE to perform the evaluations of MVPs is **based on actual usage of prototypes by participants**. Participants were recruited amongst employees of use-case owners, namely Proditec and Sonae, wheelchair users for City of Amsterdam recruited together with the help of Cliëntenbelang Amsterdam, and Broekman Logistics employees for Datacation.

The experimental sessions were individual. They consisted of the participant following the usage scenario with the prototype (outcomes of T5.1) and answering surveys on a computer equipped with Peac²h, and a facilitator, either directly from CATIE – for Sonae (English), Proditec (French) and Datacation (English) – or from City of Amsterdam (Dutch).

Each session was performed as follow:

1. Presentation of the usage scenario to the participant by the facilitator.
2. The task performed by the participant with the prototype.
3. The AIA index answered by the participant.
4. A semi-structured interview with the participant by the facilitator:
    a. Discussing the AIA index.
    b. Discussing the prototype.
5. A socio-demographic survey with mitigation variables.

Each session has been audio recorded, as well as the screen of the prototype when the participant was using it. **The recordings were analysed thematically**: themes such as Human-AI Collaboration, Comprehensibility or User Agency.

2.6 Analysis

The analysis took two different forms: analysis of user feedback and analysis of the AIA index.

2.6.1 Users feedback analysis

An effort was dedicated to processing and analysing the data that was collected during the evaluation sessions, with the goal of providing actionable feedback to each use-case team on aspects such as user experience, clarity of interaction, and overall interface usability. To do so, we relied on a combination of **direct observations** during mock-up usage and **semi-structured interviews**. A thematic analysis was conducted to identify key recurring themes, supported by representative *verbatim* excerpts from participants. The results of this analysis were presented to the project consortium, with the aim of **accelerating the prototype refinement process** carried out within WP5.

2.6.2 Analysis of AIA index for prototype

Alongside the **qualitative analysis**, data from the AIA index questionnaire were processed quantitatively to generate clear indicators of perceived trustworthiness of the AI assistants embedded in each mock-up. This analysis helped **identify the strengths and weaknesses perceived by users** for each use-case, based on the various dimensions covered by the index (*e.g.*, collaboration, auditability, transparency). The results were used to complementing the UX (Use eXperience) analysis and helping the project teams to prioritize improvements of the prototypes.

2.7 Main outcomes

This first evaluation of the partially simulated AI prototypes led to crucial findings on two different aspects. First, we identified avenues for improving the prototypes themselves. Second, we collected relevant insights to improve the AIA index as a first functioning version at this stage.

The results related to the use of MVPs revealed that participants reported sufficient comprehensibility from the prototype to achieve the usage scenario. Noteworthy, **all participants but one was able to complete the task**. **However, most participants reported a limited understanding of the overall functioning of the prototype** (especially for the Proditec use-case) or even the usefulness of the AI assistant as a support in their (iterative) decision-making process.

We collected feedback specific to each use-case and elements to improve each prototype as well. **The main improvement avenue is related to the explanations provided by the AI assistant**. These insights encompass both the contents of explanations, their timing during the actual usage, and the positioning on the screen, reinforcing the overall User Experience.

The results related to the AIA index were crucial to improve it for the next evaluation. Overall, **all participants but one were able to complete this first version of 59 items**. The constructs focused on in the PEER project – namely **comprehensibility**, **user agency**, **overall trustworthiness** – exhibited satisfactory results given the elements collected through the corresponding items. **We were able to identify avenues for improvements, especially on social-oriented constructs**, such as **cultural aspects**, **ethics**, **privacy and data governance**.

The results reported in this deliverable are to be **carefully interpreted** since **prototypes were tested in controlled environments**, **mostly by employees of the use-case owner involved in the creation of the prototypes**. Coupled with the improvements of the AIA index, next evaluation will provide more realistic results on the actual trustworthiness of the AI prototypes.

2.8 Conclusion and perspectives

This **first round of evaluation** presented in this deliverable offered **positive perspectives both the first versions of the prototypes and the first version of the AIA index**. Results reported in this document provide avenues for improvements in these two areas for higher TRLs and therefore getting closer to ecological conditions, realistic features and more adopted systems.

For the **prototypes**, efforts will be focused on **strengthening their comprehensibility**. We (CATIE) recommend **working on explanations**, especially with better timing, form and contents. For the **AIA index**, items related to **socially oriented constructs should be reworked**, especially on cultural aspects, ethics and privacy and data governance. These last considerations are linked with activities of WP4, and will be extensively presented in D4.2 due M36 of the project.

The next phase of evaluation that will follow this phase of improvements will take into account the limitations highlighted in this document, especially on the participants' profiles and the realness of the usage scenarios.

2.9 Workplan

The table below (Table 2) present all actions performed to accomplish the first part of the task T5.7.

	Year 1			Year 2												Year 3		
	Q4			Q1			Q2			Q3			Q4			Q1		
	M10	**M11**	**M12**	**M13**	**M14**	**M15**	**M16**	**M17**	**M18**	**M19**	**M20**	**M21**	**M22**	**M23**	**M24**	**M25**	**M26**	**M27**
Creating the evaluation tools																		
Creating the AIA index																		
Creating the protocol																		
Translations of the protocol (FR, NL)																		
Implementation of the protocol in Peac²h																		
Experiments of the first 3 uses-cases																		
Transcription and translations of the experiments data																		
Data analysis (Thematic analysis)																		
Data analysis (AIA index)																		
Presentation of early results for prototypes																		
Writing deliverable																		
Review of the deliverable																		
D5.6 V1															✓			
Experiments of Datacation use-case																		
Transcription of the experiments data																		
Data analysis (Thematic analysis)																		
Data analysis (AIA index)																		
Presentation of early results for Datacation use-case																		
Writing deliverable																		
Review of the deliverable																		
D5.6 V2																		✓

Table 2: Workplan.

3. Methodology

This section presents the methodology used in the project, including **the objectives of the tests, evaluation protocol** and the **methodology for analysing and processing results**.

3.1 Objectives

The overall goal of the PEER project is to build human-AI collaborations that facilitate acceptability and adoption of AI tools perceived as trustworthy by their users. Thus, the evaluation process of the first MVPs focused on validating trustworthiness-related notions – notably those related to explainability and comprehensibility, but also usefulness, user agency, perceived collaboration, or other aspects of the user experience.

The goal of the different evaluations was to deliver results to technical partners to improve their AI systems, and use users' insights to improve the definition of scales and metric to measure acceptance level. Additionally, the City of Amsterdam use-case wanted to gather user feedback regarding alternative versions of the MVP's interface.

3.2 Evaluation protocol

The evaluation protocol consists of a **common basis** and **adaptations for each use case**.

3.2.1 Common base

A common base has been built to facilitate the experiments. The **methodological tools**, the **experimental sessions** and the **material** are presented in this section. All sessions were conducted using the Peac²h platform[1].

3.2.1.1 Methodological tools

Observation grid

To guide the observations, we created an observation grid, a standardized template that ensures different **experimenters collect qualitative data in a consistent and comparable manner** (*e.g.* problems encountered, task duration, *etc.*). The template of the observation grid is available on the Peac²h platform: https://app.peac2h.io/surveys/8532/test_protocol.

The grid sections are aligned with the different stages of the test scenario based on the different stages of prototype testing making it easy to trace insights back to specific points in the user journey. We therefore adapted this grid according to each use case, thus creating 3 different versions.

These tools are available on the Peac²h platform:

[1] You are invited to create a Peac²h account to have access at all the links related to the Peac²h platform.

- Proditec observation grid:
  - In English: https://app.peac2h.io/surveys/9380/test_protocol.
  - In French: https://app.peac2h.io/surveys/9071/test_protocol.
- City of Amsterdam observation grid:
  - In English: https://app.peac2h.io/surveys/9072/test_protocol.
  - In Dutch: https://app.peac2h.io/surveys/9379/test_protocol.
- Sonae observation grid:
  - In English: https://app.peac2h.io/surveys/9726/test_protocol.
- Datacation observation grid:
  - In English: https://app.peac2h.io/surveys/13666/test_protocol.

AIA index

With the aim of evaluating the perceived trustworthiness, we built the AIA index, as described in the future deliverable D4.2 due in M36, developed based on the foundational elements identified in the review presented in deliverable D4.1 (Fage et *al.* 2025). This index revolves around 10 constructs (Figure 5) and is composed of 59 questions. The index is still being improved in accordance with the objectives of WP4. The first version of the AIA index is available on the Peac²h platform: https://app.peac2h.io/surveys/9181/test_protocol.



Figure 5: The 10 constructs used to assess perceived AI trustworthiness.

The questionnaire starts with 51 items assessing **perceived AI attributes** to compute the composite reality of **perceived trustworthiness**:

- 3 items related to **user agency**.
- 5 items related to **technical functioning**.
- 5 items related to **safety, security, and well-being**.
- 4 items related to **privacy and data governance**.
- 8 items related to **comprehensibility**.
- 6 items related to **ethical aspects**.
- 3 items related to **auditability**.
- 8 items related to **human-AI collaboration**.
- 2 items related to **anthropomorphism**.
- 7 items related to **overall satisfaction** and **other components of user experience**.

Then, participants answer 8 items used for **situation awareness** and **acceptance-related control variables**:

- 3 items measuring **actual self-reported trust and acceptance** for said interaction.
- 1 item directly asking for **overall perceived trustworthiness**.
- 4 items measuring situation awareness as a control of trust-related measures.

Each item uses a 5-point Likert format (“Strongly disagree”, “Disagree”, “Neutral”, “Agree”, “Strongly agree”).

Interview guide

An interview guide helps structure data collection during **semi-structured interviews**. It helps guide the discussion while ensuring the homogeneity in the responses collected. It also allows different experimenters to contribute to the same protocol by collecting reliable, coherent, and uniform data.
The interview guide is **divided into two parts**:

- Questions about the AIA index to improve it (WP4).
- Questions about the prototype (WP5).

The template is available on the Peac²h platform: https://app.peac2h.io/surveys/9067/test_protocol.

The interview guide was **adapted for each use case:**

- Proditec interview guide:
  - In English: https://app.peac2h.io/surveys/9086/test_protocol.
  - In French: https://app.peac2h.io/surveys/9125/test_protocol.
- City of Amsterdam interview guide:
  - In English: https://app.peac2h.io/surveys/9088/test_protocol.
  - In Dutch: https://app.peac2h.io/surveys/9126/test_protocol.
- Sonae interview guide:
  - In English: https://app.peac2h.io/surveys/9085/test_protocol.
- Datacation interview guide:
  - In English: https://app.peac2h.io/surveys/13654/test_protocol.

Socio-demographic and mitigation variables questionnaire

The socio-demographic and mitigation variables questionnaire are divided into two parts. The first part collects **information on the general characteristics of the participants** (age, gender, socio-professional category). The objective of this questionnaire is to better understand the profile of the participants to contextualize the results and help interpretation. The second part of the questionnaire assesses various **variables regarding general user-AI relationship or user-technology relationship**: trust propensity with these technologies, willingness to accept the use of AI technologies, motivation to use AI technologies, effort expectancy, AI literacy, and company reputation. **These variables help controlling for potential biases and confounding factors** that could unintentionally cause systematic errors or distortions and lead to inaccurate or unrepresentative results.

The template of this questionnaire is available on the Peac²h platform:
https://app.peac2h.io/surveys/8018/test_protocol.

3.2.1.2 Experimental session

For each of the 4 use-cases, **the experimental session took place in a controlled environment** (a separate, quiet room with no traffic, where the participant will not be distracted by noise or the presence of other people than the experimenters) with the same modalities for all participants.

Steps of the experimental session (Figure 6):

1. **Welcome and consent form:** the experimenter welcomes the participant and introduces him/her to the study. Then a consent form is presented to the participant. If the participant accepts the conditions, the experimenter can begin the study and its recording.

2. **Task scenario:** the experimenter presents the prototype, the task to be carried out to the participant and its objectives. Then the participant must carry out the task in autonomy, the experimenter is here if the

participant needs help. The participant is asked to think aloud during the task to help the experimenter understand his/her thinking.

3. **AIA index: questionnaire focused on acceptance / trustworthiness:** once the task is completed, the participant is invited to complete the AIA index. For each item the participant can select one of the following responses: "Strongly disagree", "Disagree", "Neutral", "Agree", "Strongly agree". There is also a check box "I can't answer" available for each question. If the participant clicks on this check box, he/she must tell the experimenter why he/she thinks he/she can't answer this question. The instructions given to the participant are to think aloud during the completion of the questionnaire so that the experimenter can collect as much information as possible.

4. **Semi-structured interview:** it is an exchange between the experimenter and the participant. The experimenter asks the participant a series of questions.

   The interview is divided into two parts:
   - **First part - Predictors of trustworthiness:** the experimenter wants to know what the participant thinks about the AIA index.
   - **Second part - Behaviour with the application:** the experimenter wants to know what the participant thinks about the prototype.

5. **Socio-demographic questionnaire:** the participant must answer some questions on his/her profile (age, gender, employment) as well as mitigation variables on generic human-AI relationship.

6. **Acknowledgments:** at the end of the session, the participant is thanked for his/her participation.



Figure 6: The experimental session steps.

All experimental sessions are available on the Peac²h platform:

- Proditec:
  - Participant study in English: https://app.peac2h.io/time_periods/12429/test_protocol.
  - Experimenter study in English: https://app.peac2h.io/time_periods/12427/test_protocol.
  - Participant study in French: https://app.peac2h.io/time_periods/12113/test_protocol.
  - Experimenter study in French: https://app.peac2h.io/time_periods/12114/test_protocol.
- City of Amsterdam:
  - Participant study in English: https://app.peac2h.io/time_periods/12144/test_protocol.
  - Experimenter study in English: https://app.peac2h.io/time_periods/12146/test_protocol.
  - Participant study in Dutch: https://app.peac2h.io/time_periods/12431/test_protocol.
  - Experimenter study in Dutch: https://app.peac2h.io/time_periods/12433/test_protocol.
- Sonae:
  - Participant study in English: https://app.peac2h.io/time_periods/12117/test_protocol.
  - Experimenter study in English: https://app.peac2h.io/time_periods/12141/test_protocol.
- Datacation:
  - Participant study in English: https://app.peac2h.io/time_periods/16913/test_protocol.
  - Experimenter study in English: https://app.peac2h.io/time_periods/16914/test_protocol.

3.2.1.3 Material

To carry out the experimental session, we used the material below:

- Printed **consent form** in the language spoken by the participant.
- **Computer 1 ("participant")**: to display the mock-up and the Peac²h "participant study".

- **Computer 2 ("experimenter")**: to monitor and coordinate the session through the Peac²h "experimenter study".
- **Screen capture tool** (on Computer 1) used to record participants interactions step-by-step, making it possible to later analyse usability issues and decision paths in the user journey.
- **Smartphone** (only for Amsterdam) with the prototype on it, and a screen capture tool.
- **Dictaphone** to capture auditory information to be transcribed later.
- **Stopwatch** / visible clock (to track task duration).
- **Internet access** to be able to use the Figma mock-ups, the simulator environment and the Peac²h platform.

3.2.2 Use-case specifics

The use-cases were different due to the specificities that were provided for each experiment. We describe below the participants, the experimental conditions and the scenarios.

3.2.2.1 Description of the participants

General description

A total of 28 participants were recruited for this first evaluation round. The gender distribution is almost balanced with 10 women and 18 men. All age groups are represented, and more than half of the participants are over 30 years old (Figure 7). In terms of socio-professional category, two-thirds of the participants are employees (Figure 8). This was expected, as the participants recruited for the Proditec, Sonae and Datacation use-cases are employees. Nine participants have a disability.

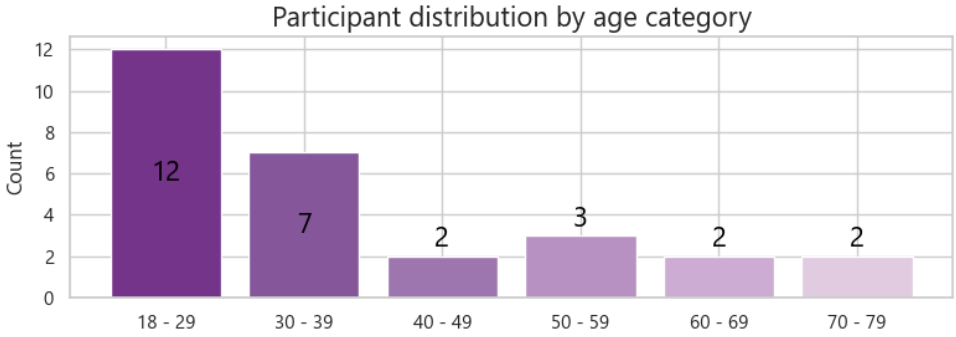


Figure 7: Participants distribution by age category.

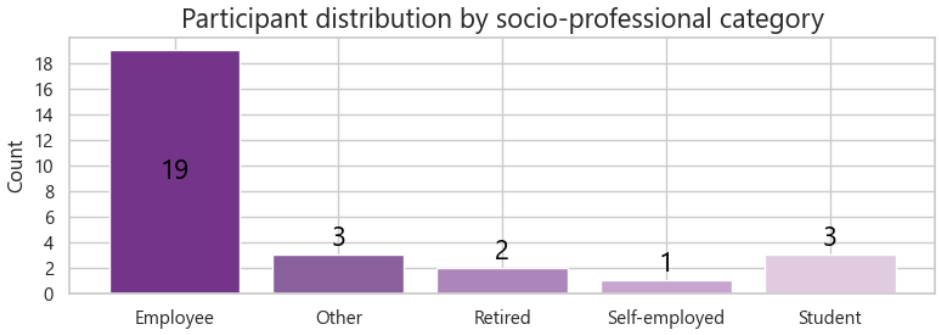


Figure 8: Participants distribution by socio-professional category.

Proditec

- 11 participants (2 women, 9 men), 1 participant colourblind.
- Average age category = 18 – 29 years old.
- Proditec employees (non-representative of end-users).
- 3 profiles:
    - 5 novices (non-technical profiles: non-engineers, non-AI experts).
    - 4 non-AI engineers.
    - 2 machine learning experts.
- All MVP-naïve: never tested the mock-up before.
- According with the participants answers to the socio-demographic and mitigation variables questionnaire, all had an overall positive opinion of AI in general (Figure 9):
    - 81.8% of participants usually trust AI technologies until they give them reasons not to.
    - 100% of participants are willing to use AI technologies.
    - 90.9% of participants are willing to take advice from AI technologies.
    - 72.8 % of participants are entertained by interacting with AI technologies.
    - For 81.8 % of participants, it is usually easy to interact with AI technologies.
    - 63.7% of participants think they are able to evaluate the capabilities and limitations of an AI technology after using it.
    - 63.7% of participants think they are able to choose the best option for them from various recommendations provided by an AI technology.
    - 72.7% of participants have a favourable opinion about the company that developed the AI assistant tested during the evaluation.

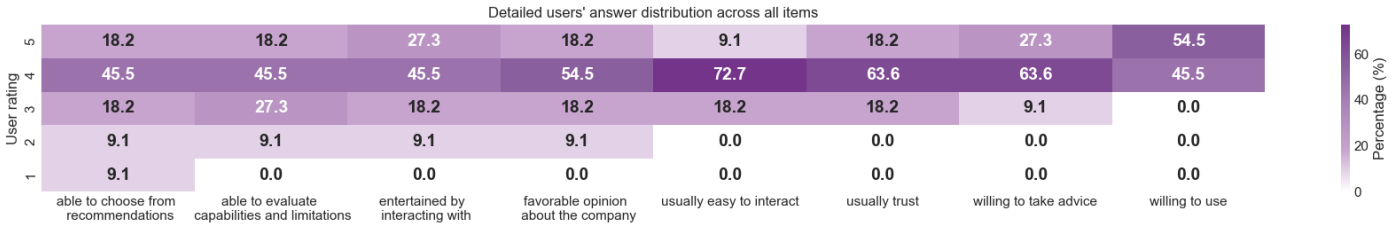


Figure 9: Participants opinions about AI technologies for Proditec use-case.

City of Amsterdam

- 7 participants (5 women, 2 men), who use wheelchair: representative of end-users.
- Average age category = 50 – 59 years old.
- 1 self-employed, 2 retired, 1 employed, 3 others.
- Familiarity with Google maps of Apple maps.
- All MVP-naïve: never tested the mock-up before.
- According with the participants answers to the socio-demographic and mitigation variables questionnaire, all had an overall positive opinion of AI in general (Figure 10):
    - 85.7% of participants usually trust AI technologies until they give them reasons not to.
    - 85.7% of participants are willing to use AI technologies.
    - 57.1% of participants are willing to take advice from AI technologies.
    - 28.6 % of participants are entertained by interacting with AI technologies.
    - For 71.4 % of participants, it is usually easy to interact with AI technologies.
    - 85.7% of participants think they are able to evaluate the capabilities and limitations of an AI technology after using it.
    - 85.7% of participants think they are able to choose the best option for them from various recommendations provided by an AI technology.

- 100% of participants have a favourable opinion about the company that developed the AI assistant tested during the evaluation.

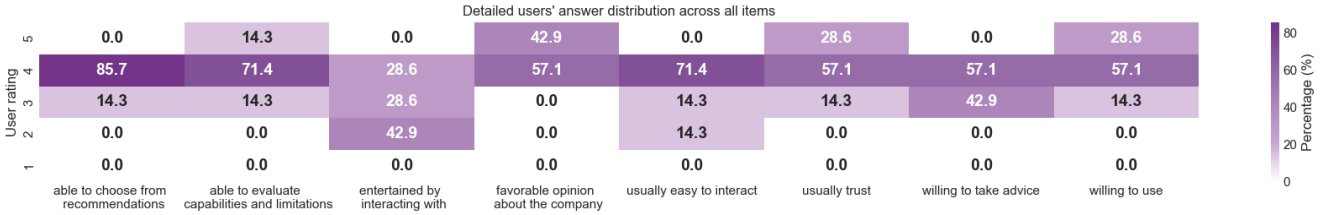


Figure 10: Participants opinions about AI technologies for City of Amsterdam use-case.

Sonae

- 6 participants (3 women, 3 men), 1 participant with limited hearing.
- Average age category = 18 – 29 years old.
- Sonae employees (non-representative of end-users).
- All familiar with AI technologies in general.
- All MVP-naïve: never tested the simulator before.
- According with the participants answers to the socio-demographic and mitigation variables questionnaire, participants had a very positive opinion of AI in general (Figure 11):
  - 100% of participants usually trust AI technologies until they give them reasons not to.
  - 100% of participants are willing to use AI technologies.
  - 100% of participants are willing to take advice from AI technologies.
  - 83.4 % of participants are entertained by interacting with AI technologies.
  - For 100 % of participants, it is usually easy to interact with AI technologies.
  - 83.3% of participants think they are able to evaluate the capabilities and limitations of an AI technology after using it.
  - 100% of participants think they are able to choose the best option for them from various recommendations provided by an AI technology.
  - 66.7% of participants have a favourable opinion about the company that developed the AI assistant tested during the evaluation.

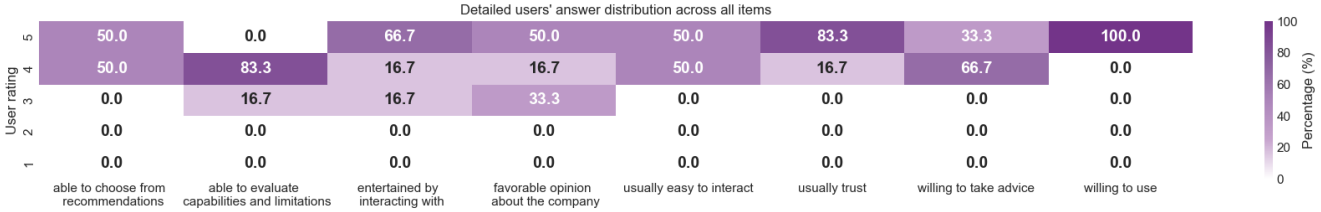


Figure 11: Participants opinions about AI technologies for Sonae use-case.

Datacation

- 4 participants (4 men).
- Average age category = 40 – 49 years old.
- Broekman Logistics employees:
  - 3 logistic engineers (end-users).
  - 1 SHEQ (Safety/Security, Health, Environment, Quality) manager.
- Familiarity with AI technologies:
  - 1 participant very familiar with AI technologies.
  - 2 participants familiar with AI technologies.

    - 1 participant non-familiar with AI technologies.
- All MVP-naïve: never tested the simulator before.
- According with the participants answers to the socio-demographic and mitigation variables questionnaire, 3 out of 4 participants had an overall positive opinion of AI in general (Figure 12):
    - 50% of participants usually trust AI technologies until they give them reasons not to.
    - 100% of participants are willing to use AI technologies.
    - 100% of participants are willing to take advice from AI technologies.
    - 75% of participants are entertained by interacting with AI technologies.
    - For 50 % of participants, it is usually easy to interact with AI technologies.
    - 25% of participants think they can evaluate the capabilities and limitations of an AI technology after using it.
    - 100% of participants think they are able to choose the best option for them from various recommendations provided by an AI technology.
    - 25% of participants have a favourable opinion about the company that developed the AI assistant tested during the evaluation.

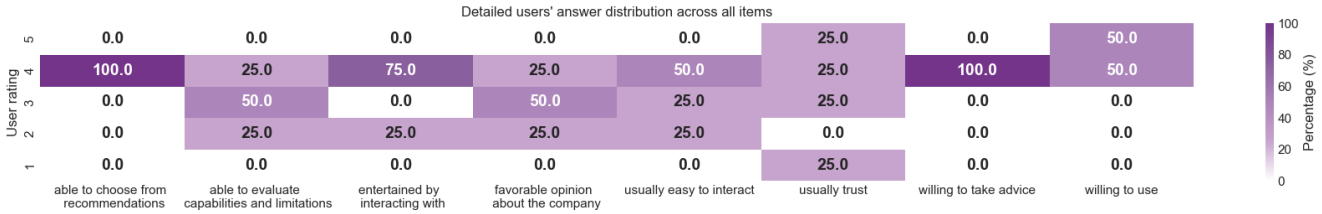


Figure 12: Participants opinions about AI technologies for Datacation use-case.

3.2.2.2 Experimental conditions

The experimentations were conducted in different languages:

- **Proditec:** French
- **City of Amsterdam:** Dutch
- **Sonae:** English
- **Datacation:** English

For each use-case, two CATIE experimenters were present to carry out the experimental sessions. For the City of Amsterdam use-case, an employee of the company, specialized in UX, was charged of driving the session because no CATIE experimenter speaks Dutch. All the tools were translated into the chosen language (observation grid, AIA index, interview guide, socio-demographic and mitigation variables questionnaire).

3.2.2.3 Description of scenarios

Proditec

The task consists of sorting coins. This sorting is performed using the new version of Proditec's sorting software. First, the participants must label the different coins. They can choose to do this themselves or use the AI assistant; in this case, they must select at least three correct and three incorrect coins. Second, the participants must train the model. They can choose their level of deep learning expertise: novice or expert. The participants then enter the degree of accuracy they wish to achieve for good and bad coins after training the model: 96% accuracy for good coins and 96% accuracy for bad coins. Third, the participants must choose the hyperparameters for their models (they can ask the AI assistant for help). Fourth, the participants can optimize

the model (threshold value, visualize images, *etc.*) until they achieve the previously defined accuracy. Finally, they must validate the model (Figure 13).

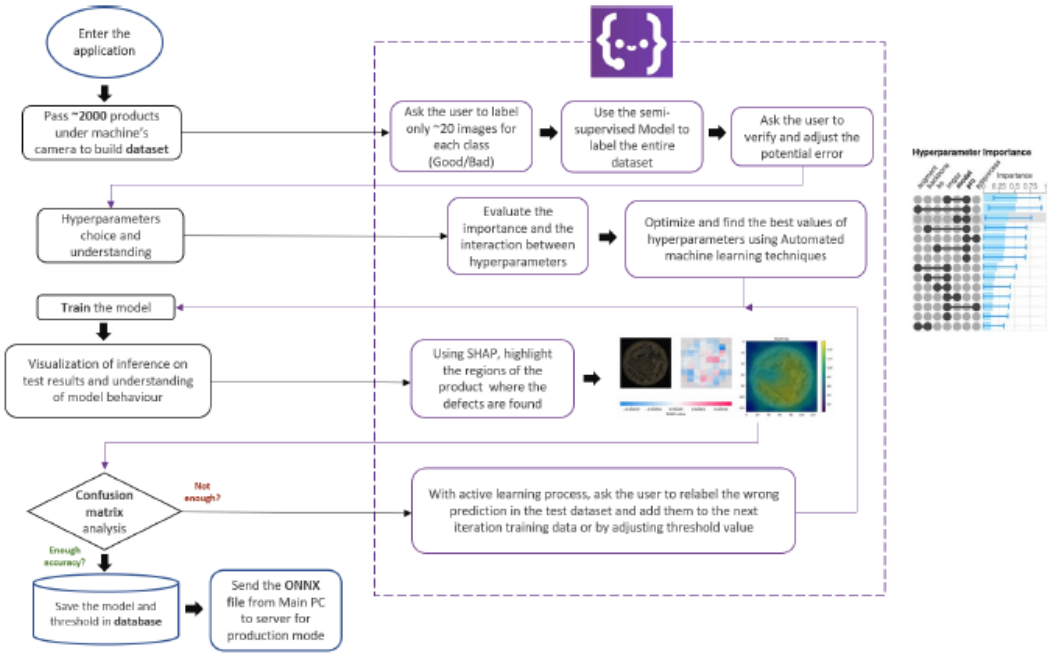


Figure 13: Proditec user flow.

City of Amsterdam

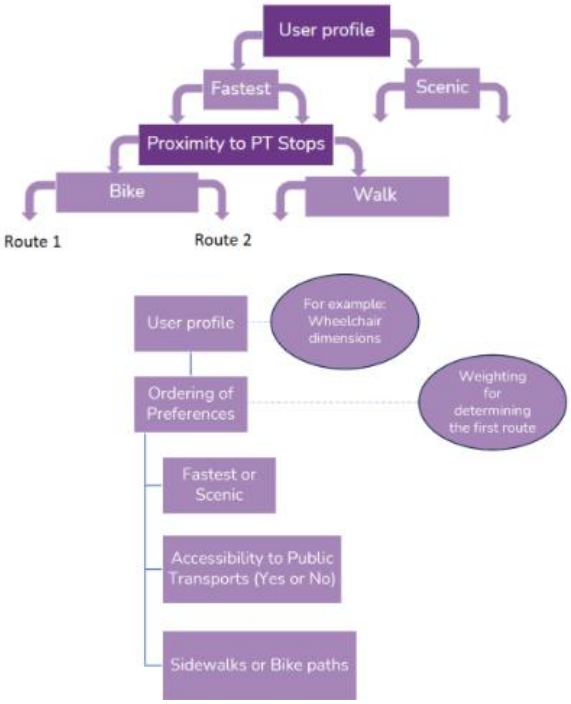


Figure 14: City of Amsterdam user flow.

First, the participants must specify their profile (wheelchair dimensions, *etc.*). Then they rank their preferences to determine the first route. There are three types of preferences:

- Do you prefer taking the fast route or the scenic route? "Fast route", "Scenic route", or "Balanced route".
- Do you prefer routes with accessible public transport options nearby? "Yes, I do prefer routes with accessible public transport options nearby", or "No preference".
- Do you prefer sidewalks or bike paths for your route? "Bike path", "Sidewalk", or "No preference".

After this, a route is presented to the participants. They can increase or decrease the number of crossroads, the time spent on the bike path, and the time spent on the sidewalk. At the end the participants complete a satisfaction question to rate the proposed route (Figure 14).

Sonae

The task consists of a route planning and navigation task in a supermarket. First, the participants go through a quick tutorial to handle the simulator. In this tutorial the experimenter introduces the functionalities of the simulator to the participants. Once the tutorial is over, the participants add some products to their shopping list: apple, carrot, canned food, ice cream and meat. The participants indicate that they want to take the ice cream last. They want to take the fastest route, so they select the "minimize shopping time" button. Then, the participants can follow the proposed route or not. They can do it as they want, like in a real supermarket. After picking the meat, participants realise they plan to buy cheese, so they decide to add it to their shopping list. At the end of the simulation, the participants can complete a satisfaction survey (Figure 15).

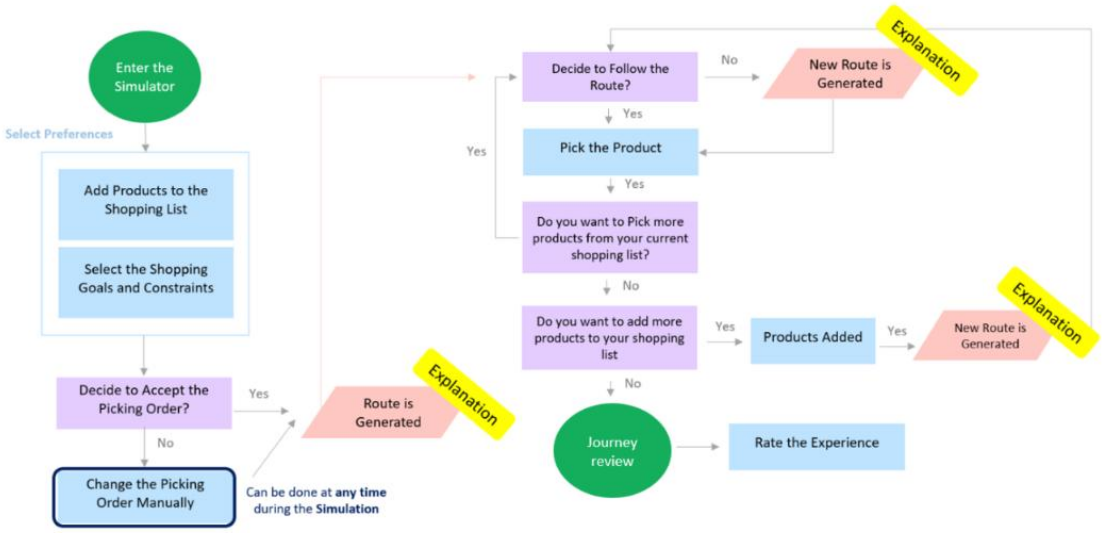


Figure 15: Sonae user flow.

Datacation

First, the participants log into the system. After that, they arrive on the "input" menu. They had to upload different files and configure the system. When the configuration is successful, they can analyse different elements:

- "WMS status" menu where they find the current state of the warehouse.
- "Order overview" menu where they find the list of all orders currently stored in the warehouse and the incoming orders scheduled.
- "Error analysis" menu where they find information about what is causing the error in the WMS.

Then, the participants can go into the "optimization" menu where the AI assistant provides two warehouse configuration solutions with explanations. At the end, the participants interact with the PEER AI assistant chatbot to choose the solution they prefer, if they want to change some elements of the solution, and to implement the final warehouse configuration solution (Figure 16).

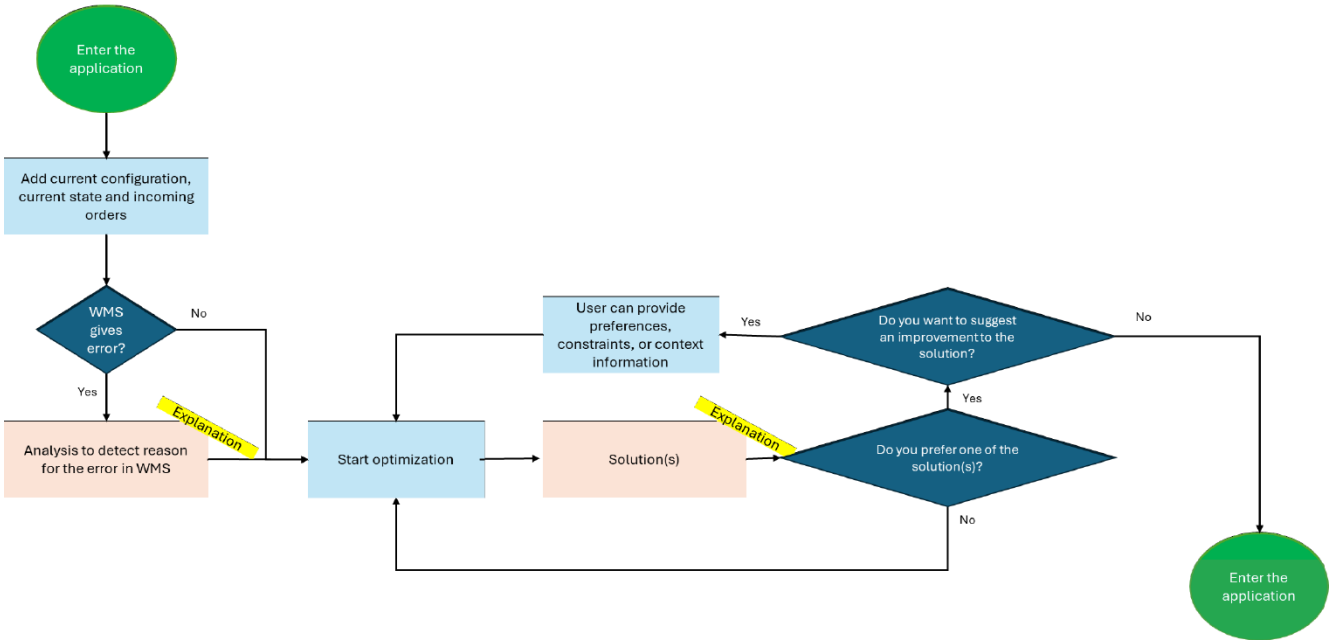


Figure 16: Datacation user flow.

3.3 Methodology for analysing and processing results

The aim of this study is double: on the one hand it is to evaluate the 4 prototypes produced by each use-case and on the other hand to evaluate the AIA index. In this sense we can distinguish two different treatments of the results.

3.3.1 Evaluation of the prototypes

We used the same way to treat the results for each use-case: a thematic analysis of the qualitative results coupled with an analysis of the quantitative results.

The 4 thematic analyses are based on transcripts of *verbatim* recorded all along the experimentation and on observations during the task. The aim of a thematic analysis is the identification, analysis and interpretation of patterns of meaning within qualitative data. First, we wrote the themes we are interested, close to the framework underpinning the AIA index (*e.g.*, collaboration, comprehensibility, *etc.*). Second, we linked each qualitative data to at least one theme. For example: *"That's it for me. [...] I don't really know what I did."* is linked to the theme **"Comprehensibility"**.

The total feedback that we have treated for each use-case is:

- **Proditec:** 550 feedback.
- **City of Amsterdam:** 150 feedback.
- **Sonae:** 160 feedback.
- **Datacation:** 144 feedback.

3.3.2 Evaluation of the AIA index

The objective is assessing the relevance of the AIA index by analysing the participants behaviours and their scores from the completion of the AIA index. We looked at the completeness rate, items scored as "I can't answer to this question" and the reasons behind it. The results will be available in the deliverable D4.2, describing the methodology used to develop the AIA index as well as its final version (TRL 5).

4. Results

This section presents the results of the experiments, first presenting the theoretical analysis of each use-case and the results concerning the AIA index.

4.1 Thematic analysis

This section presents the results of the thematic analysis for each use-case.

4.1.1 Proditec

4.1.1.1 Global results

All participants **finished the task** (mean duration: 13 min, max duration 17 min). For the thematic analysis, the most common feedback concerns the **comprehensibility of the AI assistant and the clarity of its explanations**, which are **perceived negatively**.

Some **interactions** with the prototype were **not intuitive** and some **technical issues** had a negative impact on interactions with the tool. Still, the AI assistant was perceived as **trustworthy and useful** by most of the participants.

4.1.1.2 Homepage

On the home page (Figure 17), there is **too much information** (4 participants) that is **not clear** (3 participants): *"It's a lot of information at once. I think the lambda user[...] doesn't want to come back from the beginning to better understand what he's doing after. Just the question mark, if we could have, for example you click on the question mark. It makes you a bit like a didactic on all the sites." (S02); "You have to write this for someone who knows nothing." (S01).* Moreover, for 1 participant, the **disposition of the text is not clear**. He read the steps in the order 1 – 4 – 2 – 3. The current interface leads to **a lack of comprehensibility** of the information by users.

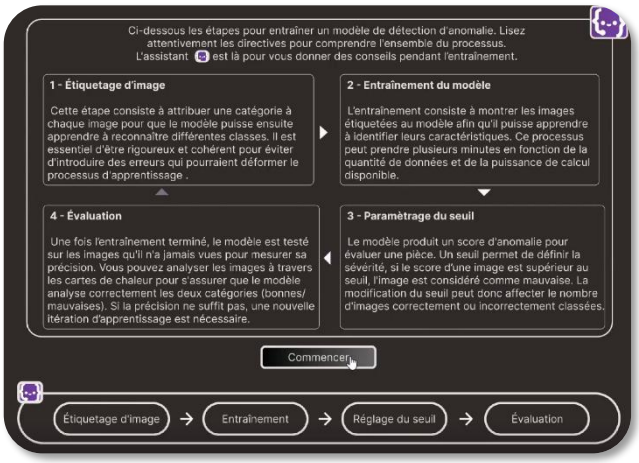


Figure 17: Proditec mock-up - Homepage.

4.1.1.3 Labelling task



Figure 18: Proditec mock-up - Labelling page.

For the labelling task (Figure 18), 2 participants find the **AI assistant useful** because it enables **coins to be labelled more quickly**: the user labels a small number of coins himself, the rest is done automatically by the AI assistant.

However, the **"auto label" button** is **not clear**. The participants did **not notice its presence directly**. They only saw it after they had labelled some coins, and clicking on it had the effect of deleting the labelling in progress (**technical problem**): *"The auto labelling button has disappeared. I'm going to have to do it manually"; "Is there a way to make this [the selector] down faster somehow?" (S09).*

Four participants found it **difficult to label coins**: to know what a good coin is and what a bad coin is: *"Well, whether something is good or bad depends on our own judgment. [..] I'm not an expert in coins." (S11).* Lot of participants (5 participants) **didn't notice the recommendation** of the AI assistant to label 3 good coins and 3 bad coins and label all the coins manually: *"It's a bit... it's a bit tiring to do this!" (S01).* This shows a **lack of comprehensibility**.

4.1.1.4 Set expertise level

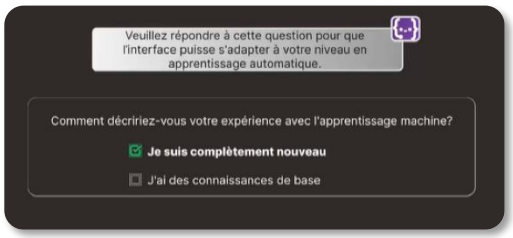


Figure 19: Proditec mock-up - Set expertise page.

During the "set expertise level" (Figure 19), 2 participants try to click on the text, not on the check box. **More work** needs to be done on the **user interface design** to make it easier for the user to identify what is clickable and what is not.

4.1.1.5 Set accuracy

For the "set accuracy" part (Figure 20), we can notice there is a **lack of comprehensibility**, because **novice participants don't understand what the percentage values mean** (percentage of good coins and percentage of bad coins), and **why** they should **select a percentage for both categories**: *"We said we want 96% on both right? On both of them?" (S01).* Also, 4 participants **didn't follow the instructions** to set 96% of good coins and 96% of bad coins: *"Customer demand was 96%. But I think the customer expects to get a higher score on good coins, so I'm going to put a higher score on the good coins." (S04).*

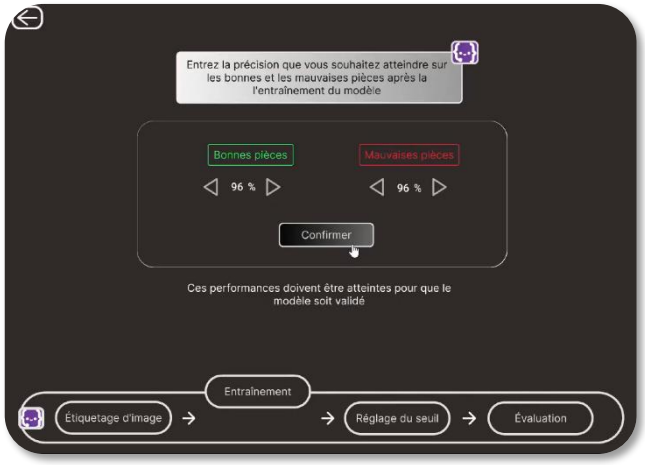


Figure 20: Proditec mock-up - Set accuracy page.

One participant found **using arrows might be tiresome**: *"Is there a way to make this [the selector] down faster somehow?" (S09).*

4.1.1.6 Hyperparameters

Five participants had access to the hyperparameters part because they had selected they had some knowledge in deep learning during the "set expertise level" part. These participants think it is **relevant to use the AI assistant for the task of hyperparameters definition** because it is a hard task. Three participants used the AI assistant directly, and 2 participants used it after failing without it: *"I will ask the assistant." (S07 after failing without the AI assistant).*

There is a **lack of explanation** regarding the types of algorithms proposed, the image size, and the backbone: *"I don't really know how to put it, but yeah, it just doesn't mean anything to me. And the image sizes, uh..." (S11).* There is also **missing information**: *"So here, he's training on how many images?" (S07).* The vocabulary is **hard to understand for novice and expert profiles**: *"Optimization of the HYPER parameters? what are HYPER parameters, like, parameters that are HYPER?" (S06).*

The **button design** needs to be **reworked**; participants did **not understand that they could interact** with certain objects: *"Oh OK... It's a button!" for the « Launch optimisation » button (S03).* For example, S07 wait for some time for an indication after the loading/training is over. Then, he saw the "show model" button and sighed before clicking.

4.1.1.7 Performance indicator and threshold

For the performance indicator and threshold part (Figure 21), participants appreciate to have a **control by the AI assistant**. The AI assistant **adapts the results displayed** based on the results obtained. Indeed, if the accuracy is too far from the objective, the AI assistant does not propose adjusting the threshold but rather returning to the labelling task or hyperparameters stage: *"That's good! Before he told me I really had to redo the labelling and not just the threshold, and then now it's still not perfect but we're close so he says you can just do the threshold and that should be enough, this is a good thing!" (S01); "Oh right so I can relabel" (S11).*

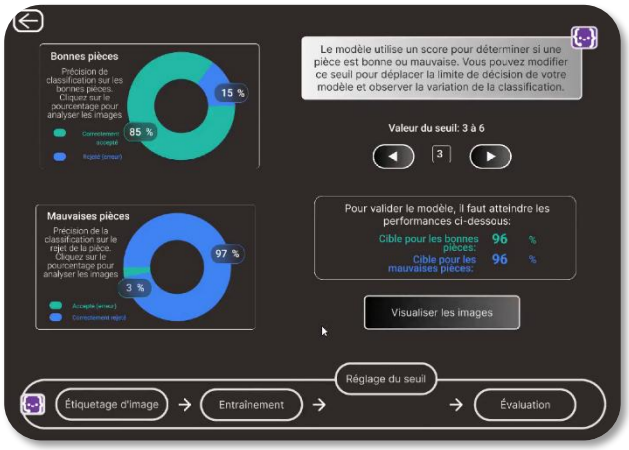


Figure 21: Proditec mock-up - Performance indicator and threshold page.

However, the **threshold notion is not understood** by participants (3 participants): *"But suddenly, I was on the threshold. And suddenly, I didn't understand. And I went back several times. I didn't really understand what I was doing. And I didn't understand this threshold setting." (S03)*, like **the visualisations of data** (the graphs) (4 participants): *"I don't understand the meaning of accepted or rejected. Is it related to my choices? Are coins supposed to be part of the bad ones that I've put in the correct ones? (S10)*. There is a **lack of explanation about the threshold and the graphs**. For example, the **colour code was not clear** for 1 participant: *"The disturbing thing is that you have to be in the green section, the accepted coins, and here you should not be in the green you need to be in the blue and then you think... what is good? On the bottom you have the blue and you think damn I've only 4% right but it's the good of the bads..." (S11).*

For this page, there is also a problem about the **button design**. The percentages are clickable but sometimes it doesn't work: *"Can we not...? OK. Because it was written to click on the percentage... I think since it was the demo, we cannot do it?" (S01 try repeated clicks).*

4.1.1.8 Visualise images close to the threshold

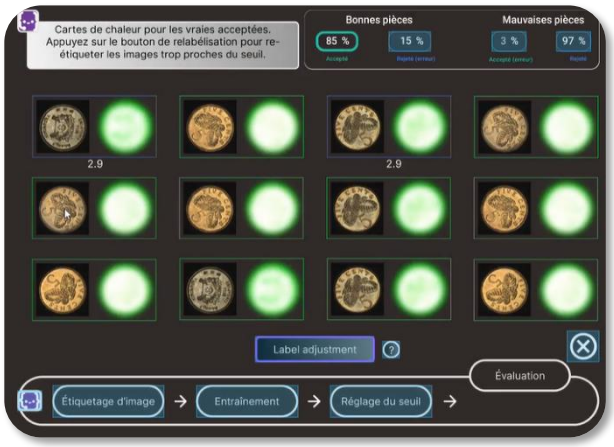


Figure 22: Proditec mock-up - Heatmap page.

The visualisation of images close to the threshold part (Figure 22) is hard to understand for participants because they **didn't understand what they were supposed to do**, they were **not familiar with heat map** (4 participants): *"There, I don't really understand... heatmaps for accepted coins? RE-labelling?" (S06);* and they **had no repository** to know what a good coin is and what a bad coin is: *"Should I still refer to a document telling me that a given coin has an acceptation limit? Let's say this one is not good..." (S09).* Participants **didn't understand the usefulness of the heatmap** (3 participants): *"But what tells me if coins are good or bad here? the colour (the heatmap), is it the green circle that tells me it's good?" (S11);* what needs to be **changed to achieve a better score** (4 participants): *"I would have liked them to put me a warning message. Be careful, you've been doing the same thing twice. It doesn't change anything." (S02);* and what are the **consequences of adjustments** (3 participants): *"Hmm. I don't know what I did. I think I didn't do anything. I think it's going to give me the same thing. I never used this machine, so..." (S03).*

It was also **difficult to understand how to navigate** between the different sections: accepted good coins, rejected good coins, accepted bad coins, rejected bad coins, and how to leave the page (3 participants): *"Why am I only seeing the good ones? Aaahhhh OK!!" S10 was blocked for some time because he did not see he could switch screens here, then understood he could relabel other coins than the good ones.* Two participants tried to **zoom on the coins** during the relabelling: *"I would have been able to zoom in on the coins. Understand exactly why this one is rejected. Understand exactly why this one is going." (S02).*

Furthermore, there is a **technical issue**, when participants click on the PEER button, they are taken back to the home page and have to start the process again from the beginning.

4.1.1.9 Success

All participants have completed the task (Figure 23). Some participants were **satisfied** (2 participants*): "Yeah, I did it!!" (S01);* others did **not understand why and how they had managed to succeed** (4 participants): *"It's OK but I don't really understand what I did." (S06);* and finally, some succeeded but **not in the way they wanted** (3 participants): *"So, indeed, I have performance achieved. But since I can't know... I feel like I was forced to change a label to be able to achieve my performance. Although it wasn't necessarily what I wanted to do. So, I can't visualize what the real impact is on the sorting. Are the pieces that were border and that I considered as good, will be considered as bad to achieve the results." (S04).*

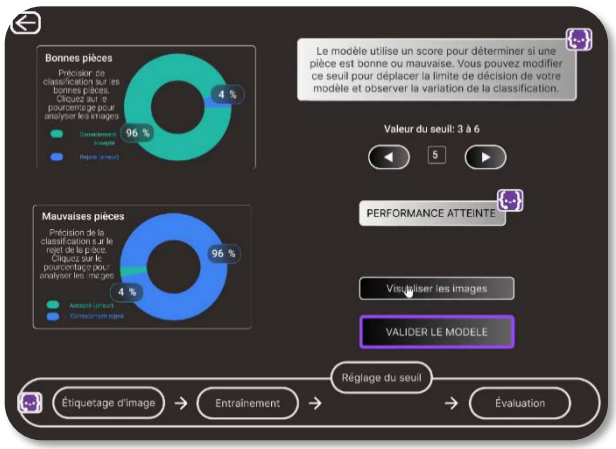


Figure 23: Proditec mock-up - Success page.

4.1.1.10 Key learnings

There is a **lack of comprehensibility** for novice and expert profiles because the **explanation is not clear**, the **amount of information is not adapted**, the **users are not guided**, and the **design of the user interface is limited**.

The AI assistant must **provide more clear explanation** about the process and **guide users** to achieve their goal: *"I think it could be improved by having more explanations, Or it's a summary that explains that I got it because I did this action [...] something that would tell me that, I don't know, maybe an attention, you have this result because of that." (S10); "I think it would be really interesting to know how it works exactly." (S02); "I would have liked them to put me a warning message. Be careful, you've been doing the same thing twice. It doesn't change anything. Try to modify that" (S02).*

Also, **buttons** must be adjusted by **increasing affordance**: buttons should look like buttons and non-clickable elements should not look like buttons: *"Why, if it's not clickable? I mean, boxes that could look like buttons." (S03).*

For the labelling task and the heatmap, users would like to have the possibility to **zoom in on the images** of the coins, to **have examples of what good coins are and what bad coins are**: *"I think it's better to understand the criteria of good pieces and bad pieces. I think it would be really important. And already, what categorizes a real piece." (S02).* They also want the **multi-selection feature** for the labelling task: *"All images at once to avoid multiplies clicks when labelling the good/bad coins." (S01).*

4.1.2 City of Amsterdam

4.1.2.1 Global results

Participants **like the principle of the application**: *an inclusive route planner with an AI assistant*. They think it is **useful**, and they are **satisfied** with the **interactions** they had with it, because the AI assistant **considers their preferences** (fast route vs scenic route, bike path vs sidewalk, accessible public transport), they can **give feedback** on the route, and in theory, the **proposed routes are accessible** for us: *"It's always nice when a route has been chosen for you that matches the wheelchair." (S01); "I think it looks good. I think it looks clear." (S03); "Yes, I thought it was very good, very useful." (S06).*

As a reminder, the participants did not carry out the navigation part. It was difficult for them to give feedback on the routes suggested by the AI assistant. One participant was able to say that he liked the proposed route because he knew on the streets proposed by the AI assistant and knew that it is accessible: *"But I know the Rijmerswaalstraat well. And I think it is a very good accessible street." (S06).*

For the thematic analysis, the most common feedback concerns the **comprehensibility of the AI assistant and the clarity of its explanations**, which are **perceived negatively**.

4.1.2.2 Collecting information

For the collecting information part, one participant thinks it is **necessary to enter** the brand and **the dimension of the wheelchair** (Figure 24), because this enables the AI assistant to suggest **routes and places adapted to the size of the wheelchair**: *"And I also thought it was very good that you also asked in the beginning how big it is. Because indeed, I fortunately don't have that, but I can also have people who can't put their legs down anymore." (S06).*

However, 2 participants think it is not necessary because they **don't know the dimension** of their wheelchair, and they **don't understand the purpose** of providing this information. This reflects a **lack of explanation** of "why the AI assistant is asking for this information" for each kind of data collected, typically here the **width** of the wheelchair. Still, for this part, one participant suggests that collecting another kind of information (the **speed** of the wheelchair) would be perceived as useful because it could be considered by the AI assistant and better estimate travel duration.



Figure 24: City of Amsterdam mock-up - Wheelchair dimension page.



Figure 25: City of Amsterdam mock-up - Preference questions pages.

Still in the collecting information part, participants **like the "preference questions" feature** (Figure 25), they think it is **relevant** because this allows the AI assistant to suggest **routes adapted to their use**. They also like it because it is a **quick and easy step**.

One participant suggests having more choices than "bike path" and "sidewalk", **like "road"**. One participant wants to **indicate "avoid place"** during the "preference questions".

For the alternative design (Figure 26),

- 2 participants prefer option 1,
- 1 participant prefers option 2,
- 3 participants prefer option 3.

Most of the participants prefer option 1 and 3 because they are **simpler**, and **fast to complete** with **all the necessary features**.

For the alternative design, the total number of participants is less than 7, because not all participants gave their opinion on it.

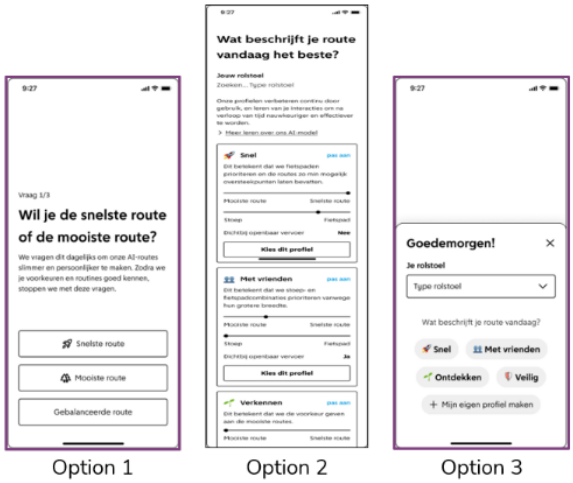


Figure 26: Alternative design for collecting information part.

4.1.2.3 Route visualisation

For the route visualisation part (Figure 27), it is **not clear** for the participants **how to interact** with the application **to generate an alternative route**, because they **don't understand that the buttons** "minus" and "plus" **are clickable**: *"I don't know how to use this, actually." (S05); "Minus three crossroads. No suitable." (S02); "In itself, I don't see exactly what has changed." (S03).* One participant suggests having **directly different options of the route** on the map, with a maximum of 3 options.

Participants are interesting to have the number of crossings, time on the bike path and the sidewalk but they **don't know** where the sections of route on the bike path are, where the sections of route on the sidewalk are, where the crossings are and how the crossings look like, because **the information is not on the map** and the **map is too small**: *"This is indistinguishable. Where is the stop? Where am I on the bicycle path? And where am I on the stop?" (S02); "Yes, it depends on how those crossings are, whether you can easily see if it is clear and things like that." (S06).*

We can notice that there is a **lack of explanation** regarding the route proposed by the AI assistant. The AI assistant **must provide more explanations and information about the proposed routes on the map**, for example:

- Whether the route is accessible or not for the dimension of the wheelchair (4 participants).
- Places localisation (restaurants, public toilets, shops, transport), and their accessibility for wheelchairs (6 participants).

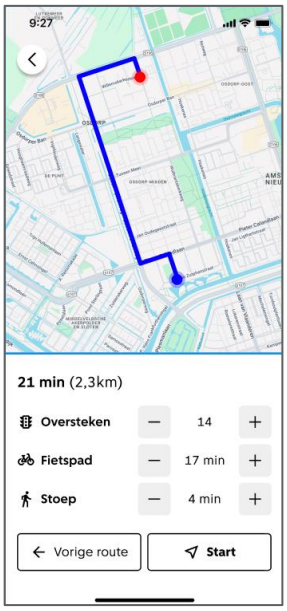


Figure 27: City of Amsterdam mock-up - Route visualisation page.

- Streets name (1 participant).
- "Street view" feature (2 participants).
- Localisation of crossings and how they look like (1 participant).
- Localisation and duration of the bike lane and footpath sections (1 participant).
- Localisation of obstacles (permanent obstacle or not) (3 participants).
- Localisation and information (ex: easy slope or not) about the transitions between bike lane, footpath, crossings (2 participants).

To have all this information, **the map must be bigger**.

In addition to the lack of explanation on the route, participants **don't understand the AI assistant's decisions** because the AI assistant doesn't explain them: *"It says up here that the total is 21 minutes. And of that, 17 minutes is the cycle path and 4 minutes is the sidewalk. The distribution of these routes. I don't think it's that clear." (S05); "Well, I would like to see what I have entered and what he uses." (S03)*. It must **provide explanation on why it suggests this route** (1 participant).

For the alternative design (Figure 28),

- 1 participant prefers option 1,
- **3 participants prefer option 2,**
- 1 participant prefers option 3.

Most participants prefer option 2 because the map is bigger than option 1, different options of the route are directly generated, and there is some information directly on the map.

For the alternative design, the total number of participants is less than 7, because not all participants gave their opinion on it.

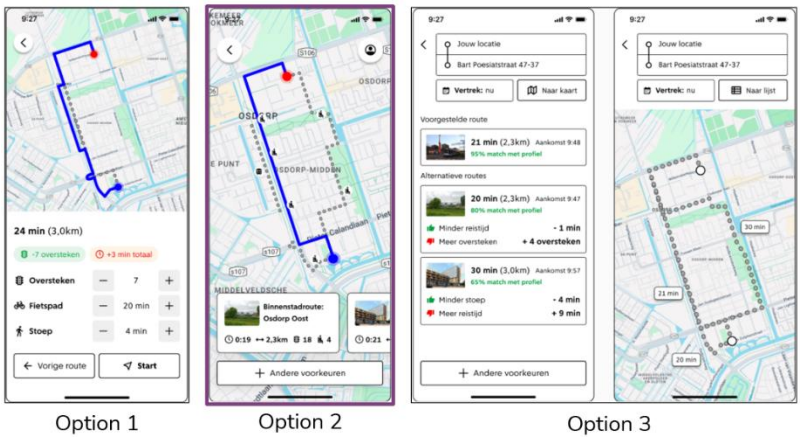


Figure 28: Alternative design for route visualisation part.

4.1.2.4 Feedback

Participants **like the "feedback" feature** (Figure 29), they think it is **relevant**, and they have the **intention to use it** because this allows the AI assistant **to adapt future routes** based on their feedback. Also, this enables **problems to be reported** so that they can be resolved: *"I could do that. I would like to do that." (S03); "Yes, I think it is useful to ask, because I actually hope for such a learning capacity." (S03).*

In the feedback part, participants would like to **indicate some specific problems** like obstacles (garbage, bike) (2 participants), invalid parking spaces (1 participant), invalid toilet (1 participant), bad eight of footpath (1 participant). One participant would like to have the **link of the municipality website** directly on the application to report problems, so that they can be treated.

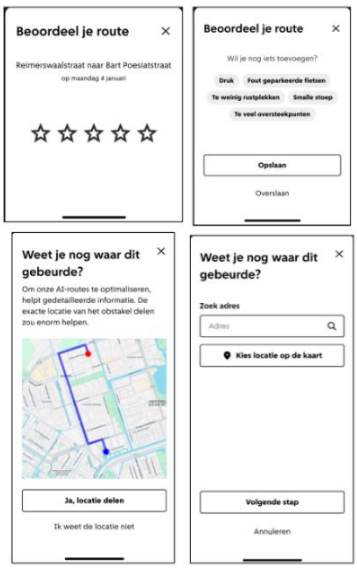


Figure 29: City of Amsterdam mock-up - Feedback pages.

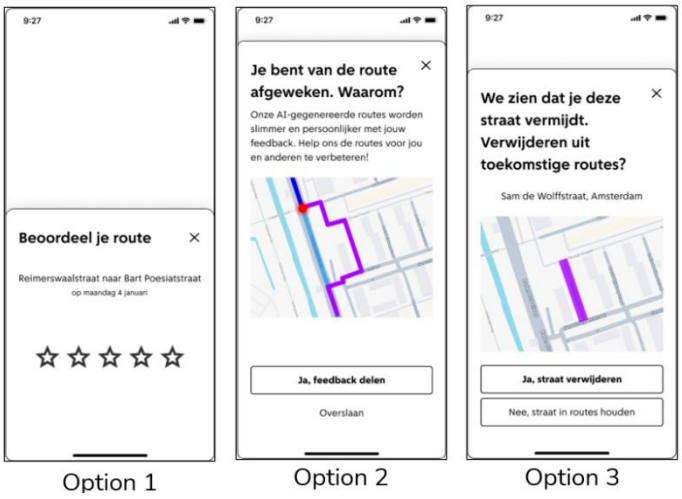


Figure 30: Alternative design for feedback part.

For the alternative design (Figure 30),

- 2 participants prefer option 1,
- 2 participants prefer option 2,
- 1 participant prefer option 3.

Two participants agreed that the AI assistant should ask them why they had deviated from the initial route because it makes the feedback task easier (option 2), but 1 participant disagreed because he would feel too stalked by the AI assistant. Participants prefer indicated "avoid place" during the preference questions step than during feedback step (option 3).

For the alternative design, the total number of participants is less than 7, because not all participants gave their opinion on it.

4.1.2.5 Key learnings

Collecting information

- Users are **willing to provide some personal information** (type of wheelchair, dimensions, areas to avoid, preferences) only i**f they know** what this information is used for. **The AI assistant must provide explanation on its functioning**.
- The collecting information step must **be quick and easy.**

Route visualisation

- **Users want to understand** how their data is used to arrive at the routes generated, **so the AI assistant must provide explanation**.
- Users want to have concrete information (*e.g.* localisation of crossings) about the routes generated directly on the map.

Feedback

- Users are willing to give feedback after using the application to indicate specific problems they have encountered during the route.

4.1.3 Sonae

4.1.3.1 Global results

For the thematic analysis, the most common feedback concerns **the satisfaction, and the user experience the AI assistant** which is **perceived positively**.

The **AI assistant** is perceived as **useful** by all the participants because it allows to **know the location of products** (**key feature**), and it is a **quick way to do shopping**: *"It is good for me to know **where are the things** I want to buy, and which steps I need to take to get the things I want the **quickest way**. It is useful because when I'm going to do my groceries I always think: **where are the things I want to buy**?" (S05); "This will help a lot because I don't know **where many of the products are**." (S04); "I'd use it tomorrow if I could. **It just looks to be quicker**. So, I think this would be really helpful." (S01).* But some participants are **unsure** if they want to use it **for recurrent use**: *"But to be fair, I **think I'll use it one time in each store** and then perhaps not anymore." (S03).*

Moreover, depending on the users' profile and preferences, participants **might or might be open to have product recommendations**. **The user agency should be preserved**: *"Because if I want to be **the fastest**, I don't want him to put some item or some path where I go through a **promotion** that the store wants me to go." (S02); "Maybe in the future **I want to try new products**, I'm guessing this will be an algorithm and then **it could be quite fun** to be like I want to take a look at **this week's new products** based on what I tried last. Perhaps it could help me **think of other products** while I edit something if I'm considering it to be an assistant for shopping and not just routing. The assistant could help me do the shopping list based on **my personal data**, previous purchases or even other people like me." (S03).*

All participants have a **good understanding of the functioning of the AI assistant**. It gave **explanations** to the participants that are **a key component to build trustworthiness**: *"I mean, **it did pretty much give me a lot of information**, the expected time, the current time, the explanation it had. **If it didn't have any of it, maybe I wouldn't trust it as much**." (S02); "**The most important is to understand**, to see why he's doing the things that he's doing and **why** is he recommending the things that he's recommending. The way the information was presented, it was clear for me why he was making the decisions. So, **I think he gained my trust**." (S01).* For this evaluation, in a controlled environment with a simulator, all the participants **trust the AI assistant**: *"I would trust the AI system again, **but not blindly**." (S03); "Personally, I trust it like **I trust GPS with my life** when it says that the path saves us time." (S04).* **It might be different in a real situation.**

It is important to notice there were **a few technical problems** with the simulator environment during the tests. Participants tend to click very fast to move in the simulator environment what caused lag (**performance problem**): *"I think it's not very fast." (S06)*. Some items could not be picked (**bug**): when the "frozen products at the end" option was checked, the ice cream could not be taken before, for example, in the penultimate product. One session had to be reset because of the **crash of the simulation environment**.

It was difficult for participants to separate the content of the application from the display of the interface: *"It's difficult to comment on this and not considering the interface. So, I think the information is ok but could be passed on in a more human-like manner." (S03).*

The simulator environment has also some **usability faults**. For example, the **functioning of the fish / meat queue** is **not clear**. The ticket for the fish / meat queue is automatically picked at the entrance in the simulation but there is **no explanation**. Moreover, **menus, selectors and controls** of the simulation environment **caused some struggles**: *"Adding and editing the products. So that's one thing that I would improve: to be **more friendly to the consumer**." (S03).* **The usability of the interface must be improved** to have a better user experience: *"I think the interface might be the key point to **improving this experience**." (S01).*

4.1.3.2 Shopping list

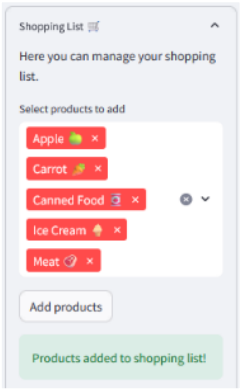


Figure 31: Sonae simulator environment - Shopping list.

For the shopping list (Figure 31), participant noticed that **products** were **not classified** by category or aisle: *"The shopping list should be like **ordered in alphabetic order** because it's really confusing to find the things" (S04).* Moreover, **interactions** to add and remove products from the shopping list were **not adapted**: *"When you told me to add another product, I had to clear all the products and then edit again [...] **it could be easier**" (S01).*

4.1.3.3 Configurability: constraints

"Constraints" feature (having options) (Figure 32) is **perceived positively** by the participants: *"I feel like in order for people to **actually use this**, the **constraints** are the most necessary part." (S02).* For example, the "collect **frozen products in the end**" checkbox was **really appreciated** by the participants: *"I like that because it's a thing that **we don't always think about**. I go and then I realize it shouldn't be the second item I choose; it should be at last. So, I go picking the others and then come back there." (S04); "The constraint of the ice cream. I think that was really clever. I think **it's maybe the one that I would use more** because it's a really interesting use case. **When I'm shopping, I do the same.** I tend to let that type of products for the last. **Sometimes I don't do it** because I just passed the products, but I think this one is really important." (S01).* They would like to **have other constraints** like this, for example "collect heavy products in the end". One participant proposes that the AI assistant should suggest default constraints: *"I think it should be the AI assistant's job to **do that for you**, and you say OK or not." (S03).*

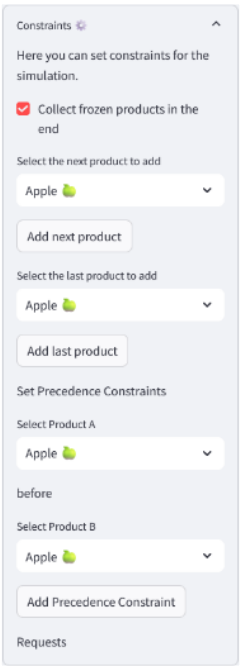


Figure 32: Sonae simulator environment - Constraints.

4.1.3.4 Target: routes comparison

For the target feature, it was **not easy** for the participants **to see the difference between the two routes**: minimize shopping time and minimize shopping distance (Figure 33). They made **many backs and forth** to compare the two routes.

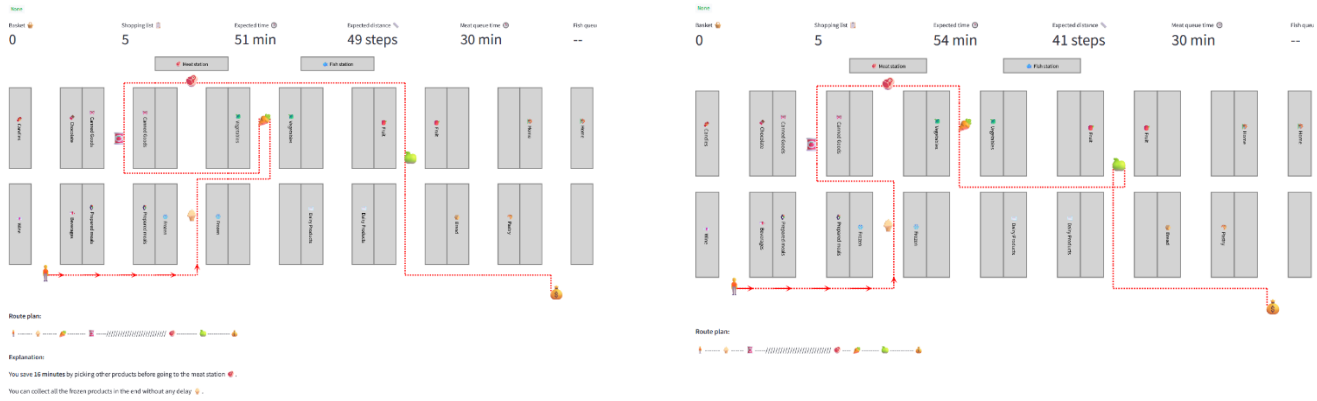


Figure 33: Sonae simulator environment - Itineraries.

4.1.3.5 Route execution: deviations

During the route execution (Figure 34), participants followed the itinerary proposed when it corresponded to **what they would do**: *"It needs to **adapt to human behaviours**!" (S05).* They **relied on the map to make decisions:** *"I knew the cashier was on the right side. **I knew that because I had a map**. So, I think the map is important to show to the customer." (S03)*; and **prioritized directness** over efficiency: *"For example, when I did not follow the instructions the AI gave me, I think it could have been interesting to **add in red -or another colour- what changed**. It could have been interesting to add the expected time when I did the change, for example, plus 3 minutes. **Just so I can see how not following the AI's recommendation changes my path**." (S02).*

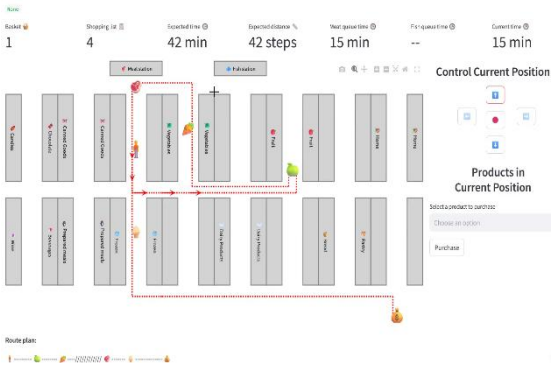


Figure 34: Sonae simulator environment - Itinerary execution.

When participants deviated from the route, the AI assistant didn't update the route immediately. It had a **lack of responsiveness**: *"When **I choose to ignore one of the decisions** that he was making, I noticed that it did not change my route when I clicked the first time ignoring his decision. **It only updated my route on the second click. I was hoping to be on the first click.** I decided to go up and not down and it **instantly** updated the route." (S01).*

Two participants decided avoiding going 2 times in the same aisle because they think they would have acted the same way in a supermarket: *"I didn't want to **go to the main corridor two times**, so I went straight. Don't pass through the same thing twice" (S03); "It looked easier for me to go straight so **I don't go back where I went before**. [...] I really think the recommendation was the one that made the route quicker, but I just imagined me in the store and **that's the way that I do it**." (S01).* **It is important to notice that this is a projection of user behaviour. In a real situation, in a real supermarket with a shopping trolley, users' behaviours might be different.**

At the end of the shopping journey (Figure 35), when there were two products left to pick up, 3 participants decided to not take the ice cream last to avoid them making back and forth: *"**I did not follow the route** because you want to take the cheese on the road, direct to the cash. **It wasn't as linear as it should be**." (S02); "The thing about the ice cream is **I didn't think it would be that much difference for the melting** if I went for the cheese first." (S03).*



Figure 35: Sonae simulator environment - End of the itinerary execution.

4.1.3.6 Rating

For the rating part, the participants think this feature is **relevant to report problems**: *"I think it was ok. It was **expected** as well..." (S06)*; but it could be **optional**: *"**I didn't really look at it**. If I were at the store, I would use it the first time. Maybe the first three times if I had an issue, I would use it again, and then the other times **I would probably just skip it.**" (S02)*.

4.1.3.7 Explanations

For the explanations part, the participants had access to the number of items in the basket and in the shopping list, the expected time, the expected distance, the meat / fish queue time and the current time. These explanations are **perceived as good** by participants: *"**I think it was quite simple**." (S02); "I think it's very **easy to understand**. I think **it's relevant**." (S06);* but the **interface design** needs to be **improved**. Participants would like to have also the price of the products: *"Just the cost of the products, of my cart here." (S06).*

The participants had also access to other explanations below the map of the supermarket: the route plan and some information provided by the AI assistant to explain the proposed route. Participants **needed to scroll down** to see these explanations and **only 2 participants** checked them.

For the participants it is important to have **explanation,** but it should be **timely:** *"I do think **it's important to have some kind of explanation**, like it doesn't have to necessarily appear on screen, but it is important to be there for people to **understand why the route is that way**." (S02); "If I arrive at the meat station just in time, **that's all I need to know**. I don't need to know if there are 20 minutes left. Give that information to the user **only if he asks for it**." (S03).* One participant suggests having **a chatbot**: *"But it's fed me like some verbal cues or something like **more of a conversation** and suggestions and questions before. Instead of just input and output-based assistance**, I think it would feel more trustworthy** if it asks questions and response with answers or actions or suggestions **more human** here in directions." (S03).*

4.1.3.8 Key learnings

- **Humaneness and directness** rule over algorithm efficiency.
- **User agency** has to be preserved.
- **Explanations** are a key factor of acceptance.
- **Content clarity** is important, and the **interface design** as well.
- Knowing **where products are** is an answer to real user need.
- An **AI assistant for products localisation** could be preferred over route planner.

4.1.4 Datacation

4.1.4.1 Global results

Participants see the **interest** of using the AI assistant and are **willing to use it in the future**: *"It **saves us a lot of time**. I think **I see opportunities in that**." (S03).*

All participants have a **good understanding of the functioning of the AI assistant**, but this one **doesn't give explanations of its functioning** to the participants: *"But to be honest, **it's an assumption** because **I don't know what's feed** into the assistant." (S02).*

In addition to the lack of transparency of the AI assistant, the Wizard-of-Oz simulation made it **difficult for participants to really trust** the solution: *"So that's a bit, **I want to trust** it and I think that can be helpful. **But it's fully dependable on what you put in and who's putting it in**. And so that's why **I was not surely enthusiastic** with answering because there are **too many factors I don't know**." (S03)*. Moreover, they **didn't do the implementation part**. So, it is difficult for them to **really judge the relevance** of the chosen solution: *"I can't answer still. Because, yeah, **we didn't have the chance to apply or to implement the solution**. To see does it work or not. We have just the planification." (S01)*. And they **tested the AI assistant only once**.

For the thematic analysis, the most common feedback concerns **the clarity of the explanations provided by the AI assistant** and the **comprehensibility of them**, which are **perceived positively**.

Participants warned about the **loss of skills** and the limits that can entail. If no engineers can do the work anymore, there will be **no one left to assess the relevance** of the solutions proposed by the AI assistant: *"We should make sure to somehow maintain that knowledge." (S03).*

4.1.4.2 Input

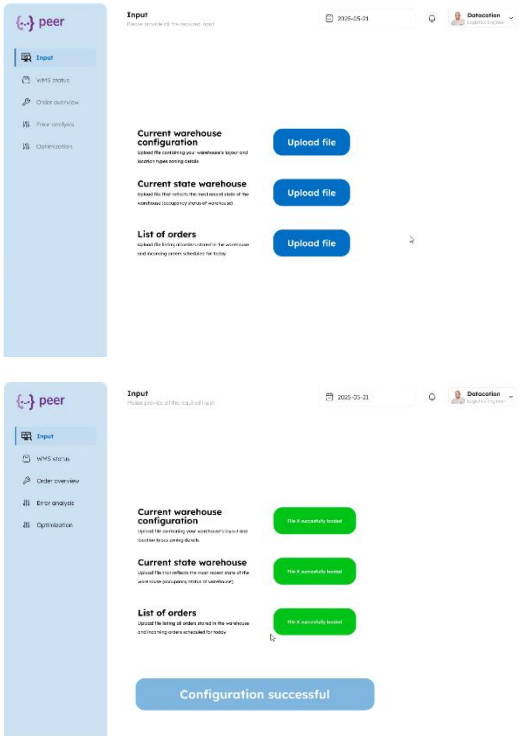


Figure 36: Datacation mock-up - Input menu.

In the "input" menu (Figure 36), the participants have a **lack of user control**. The Wizard-of-Oz simulation feels like acting too fast: *"Because this is **completely automated**; [I would want] confirmation that it has been done, or a check because I didn't see it [....] **like give approval to execute these three actions**, instead of giving three options in one go. It feels like you click the button, and you think, I did something wrong. Something happens. **I only want to do one thing instead of all three at once**." (S04).*

Moreover, participants are lost and **don't know what to do** in this menu: *"At beginning I wasn't sure what I'm doing." (S01); "I thought: OK, well, what am I going to do? [...] Do I need to upload?" (S02)*. They would like **clearer instructions**: *"Because there was **no kind of guidance** or instructions for me." (S01); "Maybe some text there, just "click here"." (S02)*. After uploading the documents and clicking on the "configure tool" button, the 4 participants **wait and remain idle without taking further action**. They have seen that the button has changed but they don't know what to do next: *"Configure tool. Successful. And now..." (S01)*. These elements reflect a **lack of guidance on the interface**.

4.1.4.3 WMS status

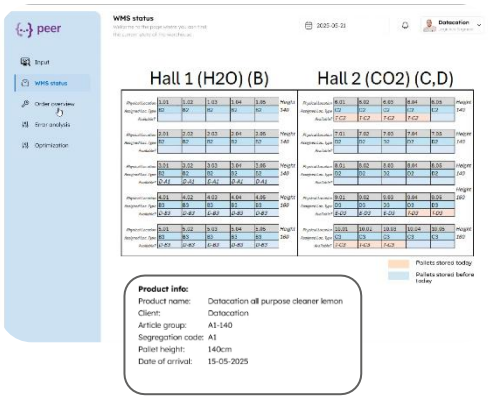


Figure 37: Datacation mock-up - WMS status menu.

In the "WMS status" menu (Figure 37), all participants take the time to look at the different information. But the participants think this menu **lacks details**: "*Not enough, **it should be more elaborated**. Like it should have reporting, like utilizations. For example, 50% free space in that area. You can store 500 pallets D4, for example, in that area [...] indicate the number of free space for each row: "You have 1, 2, 3, 4, 5, yeah, 6 places". **Here now I need to look at it and count**." (S01); "**Can I also see** whether they are occupied, yes or no?" (S02).*

In addition, the "product info" element was not used at all. Maybe it was not seen or not useful.

4.1.4.4 Order overview

For the "order overview" menu (Figure 38), all participants were engaged in the scenario and took the time to look at the list of orders. But they don't use the "filter" feature. Maybe it was not seen or not useful in this scenario.

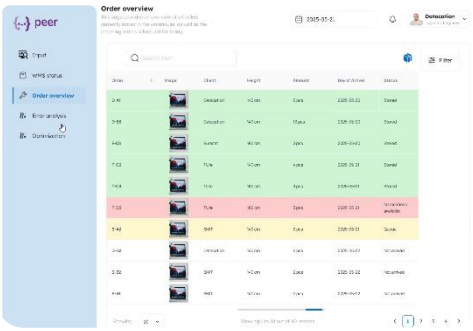


Figure 38: Datacation mock-up - Order overview menu.

4.1.4.5 Error analysis

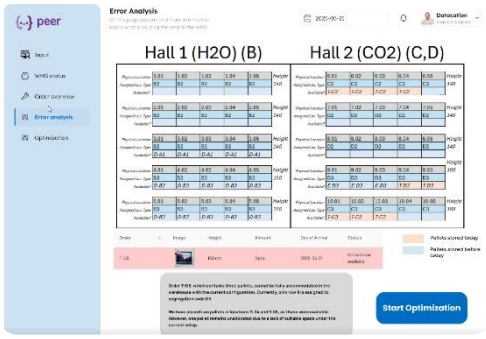


Figure 39: Datacation mock-up - Error analysis menu.

For the "error analysis" menu (Figure 39), all participants take the time to read the explanations provided. These **explanations** are well received and **seem to help to understand the situation**: *"I can see the error. I can see the issue [...] **this is well explained**." (S01).*

4.1.4.6 Optimization

For the "optimization" menu, we first detail the elements about the UI/UX, then the elements related to the solutions explanations (Figure 40), and after that the chatbot feature (Figure 41).

UI/UX

The **visual elements** of the "optimization" menu are **positively appreciated**: *"What I really like is that it gives not only explanation in words but also visual. I really like it because **a picture tells you more than a thousand words in real life but also in AI**." (S04).*

But the **architecture of the menu is unclear**: the 4 pages have very **different kinds of elements**: *"I was confused like is this moving to the next page going to give me another option or but for me **it was something completely different, without indications, without telling me what I'm going to see**." (S01)*; *"I think chatbot need to be here somewhere [in the menu]." (S01).* Moreover, after reading the explanations for solution 1 (pop-in), participants landed automatically on page 2, while they were previously on page 1. They were surprised.

Solutions explanations

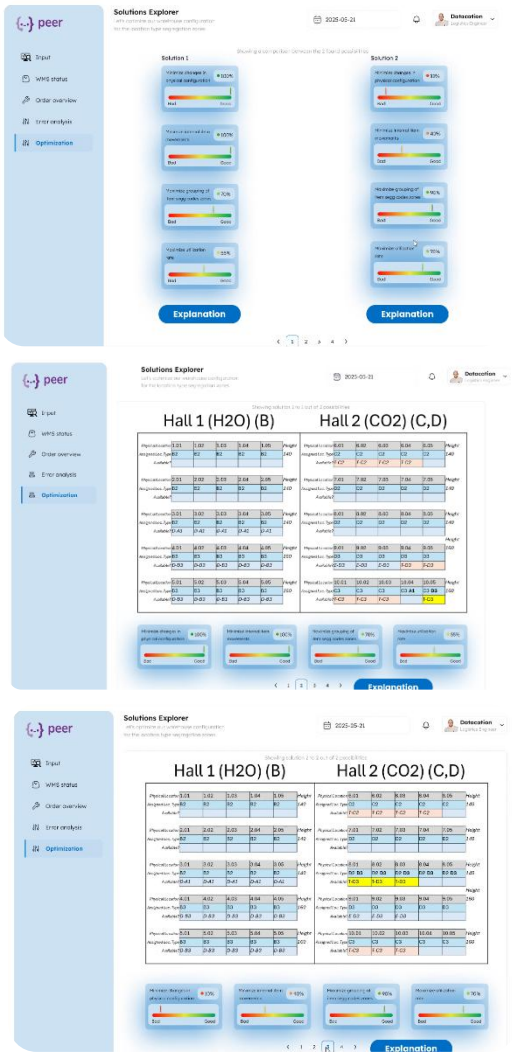


Figure 40: Datacation mock-up - Optimization menu - Solutions explanations.

The participants looked at all the different **explanations** and **liked** them: *"The fact that **multiple solutions** were **provided** with both **their pluses and minuses**, **so in the different categories** and assumed that the categories were upfront determined." (S02); "So, **it gave you** really the **information to make a decision**." (S02).* However, **many clicks are required to go through all different information**: the participants do a lot of back-and-forth between pages 1, 2 and 3: *"I think it's important to have a screen that shows more information. [...] or just to have **all the information, easy to follow in one window**." (S01).*

Participants notice that some **explanations** are **too limited**: *"And a little bit more feedback in terms of **what the reason is, why.** I thought explanation. But for me, the explanations were too limited to understand why I preferred a different suggestion than what I thought myself." (S04)*, and they would like to **have qualitative information** rather than percentages: *"I think it's good that the system shows you that for category one, solution 1 is better than solution 2. It's good that it's feasible that one is red and the other is green**, but really quantifying it like this is hard, I think**." (S02).*

Indeed, the participants were **confused by the use or interpretation of percentages**: *"But giving it a score, yeah, I think it's hard to say if something is 50%, 55%." (S02)*; *"I think it's important to know the numbers, [...] the base of these percentages: what 50% indicates, what 80% indicates. **It was honestly, not clear to me**." (S01).*

The **color-coded gauges** help **comparing easily the solutions**: *"**I just looked at those small lines, not as details so much** so I checked the percentage afterwards, but just the lines for me was **just one glance**. I thought, OK, when I have that one, because this is more green. Here I see orange and red, so I don't want that. Then I would pick that." (S03)*. But participants think that the **differences between solutions could be highlighted**: *"This is different than this, and different from this, but for me, from the beginning I thought, okay, **this is all the same**." (S01).*

Chatbot

During the test, participants were **restricted in their interaction with the chatbot**: *"But **in the example, it was directed**. I would say, yeah, it was hard to experience that." (S01).* In fact, they had to choose "solution 2", even though they all wanted to choose "solution 1".

When users want to modify a solution, they need to have the description of the solution in front of them. During the test, S01 had to go back-and-forth between page 4 and 3, and the conversation with the chatbot was not saved: *"I **need to go back** to the layout **to see**. Because here, I'm writing. But **I don't see the things**." (S01).*

Participants have a **mixed feelings about the chatbot feature**, they like it more or less, and they don't know if it is really useful: *"It was **nice**. It gives you the **feeling** that you're **talking with somebody on the other side**. It's not like, okay, I'm talking to the computer, and I do get some feedback. [...] You have a real interaction with **somebody** who also has the same knowledge." (S04)*; *"This felt **more like a conversation** in terms of it gave me feedback and what I always see with ChatGPT. It gives you an answer, but it's not thinking with you." (S04)*; *"**I'm not sure whether we really need this chat function**. On the other hand, it makes it a little less black and white, and that **can also be nice**." (S02).* But it is important to notice that the Wizard-of-Oz chatbot provided **limited immersion** as it did not really allow for interactions: *"For example, **if you just ask anything, the system can still handle it**." (S01).*

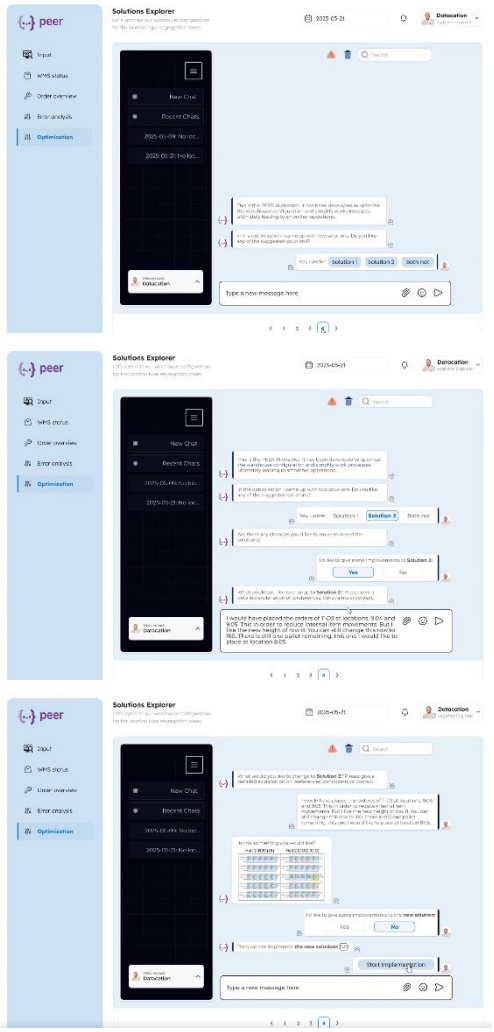


Figure 41: Datacation mock-up - Optimization menu - Chatbot.

4.1.4.7 Key learnings

- Visuals elements and text of solutions explanations are positively appreciated.
- The **functioning of the AI assistant** needs to be **explained** to the users.
- The menus architecture needs to be rethought, with **more guidance** of what users have to do.
- Input elements must be like the elements used today and users' workstations must remain similar to what they use today (multiple screens): **adding an AI assistant does not mean simplifying** all elements of a process.

4.2 AIA index

In addition to the primary qualitative methods (direct observation of the user testing sessions and semi-structured interviews with each user), **quantitative indicators** were collected during the evaluation process using the first version of the AIA index. This questionnaire measures a range of 10 key AI attributes associated with trustworthiness, with the aim of computing a composite index of perceived AI trustworthiness as a predictor of acceptance. This section presents case-specific results regarding participants' perceptions of trustworthiness.

4.2.1 AIA index score

To compute **each participant's overall score**, their average response is computed within-scale (to reflect how the participant generally perceived the dimension). Next, the average of these scale-level scores is computed to create an overall AIA index score for each participant, **representing their general evaluation of their perceived trustworthiness across the 10 key AI attributes associated with trustworthiness**. Finally, to summarize the results across participants, the median of these overall scores is used as it provides a more robust estimate (especially in non-symmetrical data). This **composite index of perceived AI trustworthiness** represents a predictor of AI acceptance and the score (out of 5) for the 1st MVP is reported below (Figure 42, Table 3).

Importantly, **the current version of the questionnaire is a work in progress** (a first prototype version), which will be progressively improved over the next iterations based on user feedback and expert inputs. **The final validated version of this index is scheduled for M36 of the project** and future score computations can highly differ due to variations in the framework and both in item wording and number.

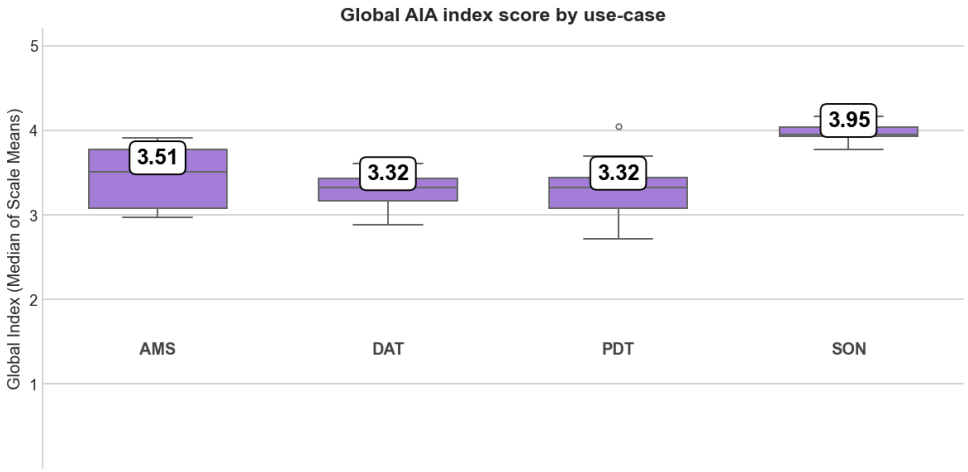


Figure 42: AIA index score per use-case.

Proditec	City of Amsterdam	Sonae	Datacation
The AIA index score is **3.32** / 5.	The AIA index score is **3.51** / 5.	The AIA index score is **3.95** / 5.	The AIA index score is **3.32** /5.

Table 3: AIA index score per use-case.

4.2.2 Patterns in separate AI trustworthiness constructs

To better understand the factors that shape the scores obtained through the AIA index, **we examine the 10 underlying constructs**, helping to identify which attributes are consistently strong and which ones may need further attention. A descriptive and distributional analysis of these quantitative indicators is deliberately prioritized over advanced inferential statistics in this report to align with best practices for exploratory analysis

given the limited sample size of the first MVP evaluation. The goal is not statistical generalization but rather to identify patterns and inform general tendencies regarding users' perception of the attributes of the AI assistant.

4.2.2.1 Score obtained for each construct

Overall, **the AIA index scales were rated rather positively in this series of MVP evaluations** (Table 4). The strengths of the PEER project, worth preserving in future implementations, are the perceptions of **"auditability"** (PDT, AMS, DAT), **"technical functioning"** (PDT, SON) except for Datacation use-case, **"collaboration"** (AMS, SON), and **"user agency"** (AMS, SON, DAT). However, **three scales scored consistently lower across use-cases**: perceived **"anthropomorphism"** (PDT, AMS, SON, DAT), **"ethics"** (PDT, AMS) and **"privacy and data governance"** (PDT, AMS, SON). For Datacation use-case, "anthropomorphism" had a better score than for other use-cases. This underlines opportunities for refinements.

	Proditec	City of Amsterdam	Sonae	Datacation
✚	The highest-rated constructs are "**auditability**" (median=4) and "**technical functioning**" (median=4).	The highest-rated constructs are "**auditability**" (median=4), "**collaboration**" (median=4) and "**user agency**" (median=4).	The highest-rated constructs are "**user agency**" (median=4.7), "**technical functioning**" (median=4.4) and "**collaboration**" (median=4.4).	The highest-rated constructs are "**auditability**" (median=3.5), "**comprehensibility**" (median=3.5) and "**user agency**" (median=3.3).
▬	A few scales (3 / 10) are not positively rated (i.e. ≤ 3), namely "**anthropomorphism**" (median=1.5), "**ethics**" (median=3) and "**privacy and data governance**" (median=3).	A few scales (3 / 10) are not positively rated (i.e. ≤ 3), namely "**anthropomorphism**" (median=2.6), "**ethics**" (median=3) and "**privacy and data governance**" (median=3).	A few scales (2 / 10) are not positively rated (i.e. ≤ 3), namely "**anthropomorphism**" (median=2) and "**privacy and data governance**" (median=3).	A few scales (3 / 10) are not positively rated (i.e. ≤ 3), namely "**anthropomorphism**" (median=3), "**safety, security and well-being**" (median=2.9) and "**technical functioning**" (median=2.9).

Table 4: Details on scale-level results per use-case.

The aggregated scale score distribution is shown in Annex 1, where boxplots allow for a granular inspection of response concentration across constructs.

4.2.2.2 Trends in score distribution for each construct

Overall, a **strong skew toward positive rating is observed in ~70% of the questions** (*i.e.* more than half of participants agreed or strongly agreed with these statements, with over 50% of responses concentrated at values 4 or 5) (Table 5).

	Proditec	City of Amsterdam	Sonae	Datacation
✚	Overall, most items (31 / 51) show **a strong skew toward agreement**. This is especially the case for "The AI assistant would always behave the same under similar conditions." (**technical functioning – reproducibility**) with 20% of "strongly agree", "The type of algorithm used by the AI assistant can be reviewed in an inspection procedure." (**auditability – traceability of**	Overall, most items (38 / 51) show **a strong skew toward agreement**. This is especially the case for "The AI assistant considers my preferences." (**collaboration – consider user's feedback**) with 17% of "strongly agree", "The AI assistant takes into account the information I give it." (**collaboration – consider user's feedback**) with 17% of "strongly agree", and "The	Overall, most items (43 / 51) show **a strong skew toward agreement**. This is especially the case for "The AI assistant is easy to learn" (**comprehensibility – intelligibility**), "The AI assistant takes into account the information I give it" (**collaboration – consider user's feedback**) and "The AI assistant is useful" (**other components of UX – usefulness**), each of which	Some items (16 / 51) show **a strong skew toward agreement**. This is especially the case for "The AI assistant works quickly" (**technical functioning – efficiency**), "The AI assistant provides information to explain its recommendations" (**comprehensibility – explainability**), "The AI assistant is understandable" (**comprehensibility –**

the algorithm) with 25% of "strongly agree", "The AI assistant would be able to perform under a variety of circumstances." (**technical functioning – robustness**) with 43% of "strongly agree", "The AI assistant takes into account the information I give it." (**collaboration – consider user's feedback**) with 40% of "strongly agree", and "The AI assistant works quickly" (**technical functioning – efficiency**) with 60% of "strongly agree".	AI assistant is useful" (**other components of UX – usefulness**) with 20% of "strongly agree".	received 100% "strongly agree" responses.	**intelligibility**), "The AI assistant is easy to learn" (**comprehensibility – intelligibility**) , "The AI assistant accounts for diverse culture" (**ethics – culture**), and "The AI assistant considers my preferences" (**collaboration – consider user's feedback**),), each of which received 100% "strongly agree" responses.

Table 5: Details of strong skew toward positive rating per use-case.

In contrast, very few items (4 to 10 items depending on the use-case) show a **strong negative skew** (*i.e.* more than half of participants disagreed or strongly disagreed with these statements, with over 50% of responses concentrated at values 2 or 1) (Table 6).

	Proditec	City of Amsterdam	Sonae	Datacation
	Some items (10 / 51) items show **a strong skew toward disagreement**. This is the case for "The AI assistant is not vulnerable to attacks" (**safety, security and well-being – protection of AI against attacks and risks**) with 50% of "strongly disagree" (med=2.5), "The AI assistant is concerned about my welfare" (**safety, security and well-being – user well-being**) with 60% of "strongly disagree" (med=2), "[...] informs me of the data it collects" (**privacy and data governance – collecting data**) with 56% of "strongly disagree" (med=2), "[...] provides access to its data structure" (**privacy and data governance – access to data structure**) with 89% of "strongly disagree" (med=1), "The actions of the AI assistant can be traced by me" (**comprehensibility – traceability of the information by the user**) with 56% of "strongly disagree" (med=2), "[...] hard to be fair in dealings with others" (**ethics – non-discrimination**) with 71% of "strongly disagree" (med=2), "[...] accounts for diverse culture" (**ethics – culture**)	Some items (10 / 51) show **a strong skew toward disagreement**. This is the case for "The AI assistant is concerned about my welfare" (**safety, security and well-being – user well-being**) with 67% of "strongly disagree" (med=2), "The AI assistant provides access to its data structure" (**privacy and data governance – access to data structure**) with 57% of "strongly disagree" (med=2), "[...] provides information to explain its recommendations" (**comprehensibility – explainability**) with 50% of "strongly disagree" (med=2), "The actions of the AI assistant can be traced by me" (**comprehensibility – traceability of the information by the user**) with 50% of "strongly disagree" (med=2), "[...] accounts for diverse languages (**ethics – language**) with 100% of "strongly disagree" (med=1), "[...] accounts for diverse cultures" (**ethics – culture**) with 67% of "strongly disagree" (med=2), "[...] accounts for many disabilities" (**ethics –**	Some items (4 / 51) items show **a strong skew toward disagreement**. This is the case for "The AI assistant provides access to its data structure" (**privacy and data governance – access to data structure**) with 75% of "strongly disagree" (med=2), "The AI assistant account for many disabilities" (**ethics – accessibility, universal design**) with 80% of "strongly disagree" (med=2), "The AI assistant spontaneously communicate with me" (**collaboration – communication**) with 60% of "strongly disagree" (med=2), and "The AI assistant expresses its emotions" (**anthropomorphism – attribution of human properties**) with 80% of "strongly disagree" (med=1).	Some items (7 / 51) items show **a strong skew toward disagreement**. This is the case for "The AI assistant is correct all the time" (**technical functioning – reliability**) with 100% of "strongly disagree" (med=1), "The AI assistant warns people of potential risks" (**safety, security and well-being - protection of the user against attacks and risks**) with 66.7% of "strongly disagree" (med=2), "The actions of the AI assistant can be traced by me" (**comprehensibility - traceability of the information by the user**) with 66.7% of "strongly disagree" (med=2), "The AI assistant spontaneously communicates with me" (**collaboration – communication**) with 66.7% of "strongly disagree" (med=2), "The AI assistant suits my level of expertise" (**collaboration – personalisation**) with 66.7 pf "strongly disagree" (med=2), "The AI assistant expresses its emotions" (**anthropomorphism - attribution of human properties**) with 66.7% of

with 50% of "strongly disagree" (med=2), "[...] responds to me when I share something with it" (**collaboration – give feedback**) with 57% of "strongly disagree" (med=2), [...] expresses its emotions" (**anthropomorphism – attribution of human properties**) with 90% of "strongly disagree" (med=1), and "[...] has all functionalities I would expect from an assistant" (**others components of user experience – dependability**) with 67% of "strongly disagree" (med=2).	**accessibility**) with 60% of "strongly disagree" (med=2), "[...] do not induce the risk of de-skilling of human operators" (**ethics – social impact on jobs and skills**) with 50% of "strongly disagree" (med=2), "[...] has a sense of human behaviour" (**anthropomorphism – attribution of human properties**) with 50% of "strongly disagree" (med=2), and "[...] expresses its emotions" (**anthropomorphism – attribution of human properties**) with 100% of "strongly disagree" (med=2).		"strongly disagree" (med=2), and "The AI assistant has all functionalities I would expect from an assistant" (**others components of UX – dependability**) with 100% of "strongly agree" (med=1).

Table 6: Details of strong skew toward negative rating per use-case.

To further interpret participant's responses across the evaluated constructs, the detailed item-level distributions are provided in Annex 2.

4.2.3 Skipped items

To make sure participants were not forced to provide an answer when they didn't feel able to, the questionnaire included an "I cannot answer" checkbox presented as an option to the right of each item in the questionnaire. Overall, **the questionnaire supported high completion rates**: participants engaged well in the questionnaire and completed most items (see Table 7 for details). One notable exception was a single participant who consistently skipped items during the evaluation of the City of Amsterdam use-case.

Proditec	City of Amsterdam	Sonae	Datacation
Some items (39 / 51) were left unanswered by at least one participant. Among them, the **items that show the highest rate of missing values** (i.e. >50% participants selecting the "I cannot answer" option) were related to **safety** ("The AI assistant is not vulnerable to attacks"), **ethics** ("The AI assistant accounts for diverse languages", "[...] accounts for diverse culture"), **auditability** ("The data used [...] can be traced for an inspection procedure" , "The decisions made [...] can be traced for an inspection procedure", "The type of algorithm used [...] can be reviewed in an inspection procedure"), **privacy and data governance** ("[...] keeps collected information secure") and **technical functioning** ("The AI assistant would always behave the same under similar conditions").	Most items (46 / 51) were left unanswered by at least one participant. However, one user answered very few items. Among the unanswered items, the **items that show the highest rate of missing values** (*i.e.*, >50% participants selecting the "I cannot answer" option) were **ethics** ("The AI assistant tries hard to be fair in dealings with others", "[...] accounts for diverse languages", "[...] accounts for diverse culture", "[...] is environmental-friendly"), **comprehensibility** ("The actions of the AI assistant can be traced by me"), and **auditability** ("The decisions made by the AI assistant can be traced for an inspection procedure", "The type of algorithm used by the AI assistant can be reviewed in an inspection procedure").	Some items (21 / 51) were left unanswered by at least one participant. Among them, the **items that show the highest rate of missing values** (*i.e.*, >50% participants selecting the "I cannot answer" option) were situation related to **situation awareness** ("I understand the limitations of the AI assistant") and **ethics** ("The AI assistant tries hard to be fair in dealings with others").	Some items (26 / 51) were left unanswered by at least one participant. Among the unanswered items, the **items that show the highest rate of missing values** (*i.e.*, = 50% participants selecting the "I cannot answer" option) were related to **technical functioning** ("[...] would always behave the same under similar conditions", "[...] would be able to perform under a variety of circumstances"), **safety, security and well-being** ("[...] would not knowingly do things to hurt me or others"), **ethics** ("[...] tries hard to be fair in dealings with others", "[...] accounts for diverse culture", "[...] accounts for many disabilities"), and **auditability** ("The data used by the AI assistant can be traced for an inspection procedure", "The decisions made by the AI assistant can be traced for an

			inspection procedure", "The type of algorithm used by the AI assistant can be reviewed in an inspection procedure").

Table 7: Details on skipped items per use-case.

4.2.4 Items answers homogeneity

Overall, **participants' perceptions were quite consistent within each use-case** (*i.e.* users tend to agree with one another within use-cases). Nonetheless, some items stand out with a high response variance (> 1.5, see Table 8). These are items demonstrating more **divergent perceptions of AI attributes among respondents**, pointing to AI attributes that are either unclear to the users or whose perception is strongly user-dependent according to their needs or expectations (leading to divergent views).

Proditec	City of Amsterdam	Sonae	Datacation
A high variance (>1.5) is observed in many items (18 / 51), specifically: "The AI assistant keeps collected information secure" (**privacy and data governance – data protection**) (v=2.3) as well as "[...] informs me of the data it collects" (**privacy and data governance – collecting data**) (v=2.7), "The data used by the AI assistant can be traced for an inspection procedure" (**auditability – traceability of the data**) (v=2.7), "The decision made by the AI assistant can be traced for an inspection procedure" (**auditability – traceability of the decision**) (v=2.7), and "[...] considers my preferences" (**collaboration – consider user's feedback**) (v=2.7).	**A high variance (>1.5) is observed in only very few items** (4 / 51), specifically: "The AI assistant is clear" (**comprehensibility – intelligibility**) (v=1.6), "The information collected by the AI assistant is observable for me" (**comprehensibility – traceability of the information by the user**) (v=1.7), "The AI assistant provides information to explain its recommendations" (**comprehensibility – explainability**) (v=1.9) and "The actions of the AI assistant can be traced by me" (**comprehensibility – traceability of the information by the user**) (v=2).	**A high variance (>1.5) is observed in a few items** (6 / 51), specifically: "The AI assistant is concerned about my welfare" (**safety, security and well-being – user well-being**) (v=2.8), "[...] spontaneously communicates with me" (**collaboration – communication**) (v=2.8), "[...] accounts for diverse language" (**ethics – language**) (v=2.3), "[...] provides access to its data structure" (**privacy and data governance – access to data structure**) (v=2.3), "The functioning of the AI assistant is made clear" (**comprehensibility – interpretability**) (v=2.2) and "[...] has all functionalities I would expect from an assistant" (**others components of user experience – dependability**) (v=1.9).	**A high variance (>1.5) is observed in few items** (8 / 51), specifically: "The AI assistant follows my directions" (**user agency – user control**) (v=3.3), "[...] guarantees my autonomy" (**user agency – user autonomy**) (v=3.3), "The autonomy level of the AI assistant is satisfying" (**user agency – user autonomy**) (v=1.6), "[...] performs its job well" (**technical functioning – performance**) (v=3), "[...] warns people of potential risks (**safety, security and well-being - protection of the user against attacks and risks**) (v=3), "[...] would not knowingly do things to hurt me or others (**safety, security and well-being - no negative consequences for the user and others**) (v=2), "[...]accounts for many disabilities" (**ethics – accessibility**) (v=2) and "[...] do not induce the risk of de-skilling of human operators" (**ethics - social impact on jobs and skills**) (v=1.7).

Table 8: Details of items with homogeneity per use-case.

4.2.5 Control variables

To complement the main AIA questionnaire, **eight additional items are included as a control** along the 10 constructs of perceived trustworthiness.

First, **direct reports of experienced trust and acceptance were collected**. Median values across all items indicate generally favourable attitudes towards the tested AI prototypes, with a median of 4 or above for both "intention to trust" and "willingness to use" the system in the future. Experienced trust (situational) is collected based on the short experience that users had with the MVP and might not fully represent the potential trustworthiness of the systems (see Table 9 for details).

Proditec	City of Amsterdam	Sonae	Datacation
The reported experienced trust is rather high (med=4), and acceptance estimations are high as well: expected future trust (med=4) and willingness to use the AI in the future (med=4).	The current reported experienced trust can be improved (med=3.5) but acceptance estimations are both high: expected future trust (med=4) and willingness to use the AI in the future (med=4).	The reported experienced trust is rather high (med=4), and acceptance estimations are both very high: expected future trust (med=4.5) and willingness to use the AI in the future (med=4.5).	The current reported experienced trust can be improved (med=3), and acceptance estimations can also be improved: expected future trust (med=3) and willingness to use the AI in the future (med=4).

Table 9: Experienced trust and intention to use.

Second, **situation awareness (SA) control variables were collected**. They included four items: "I understand the capabilities of the AI assistant."; "I understand the limitations of the AI assistant."; "I understand how the collected information led to the result." and "I am able to predict the behaviour of the AI assistant.". The SA items show more discrepancies between use-cases as they are not a characteristic of the evaluated AI: they can highly depend on differences in users' domain expertise (*e.g.* coin manufacturing, urban navigation using a GPS) and MVP's mode of interaction (*e.g.* simulation environment, Figma mock-up, *etc.*) (see Table 10 for details).

Proditec	City of Amsterdam	Sonae	Datacation
Users provided high ratings regarding all their perceptions of situation awareness, although lower levels are reported in users' understanding of the AI's limits (med=3) and capability to predict the behaviour of the AI assistant (med=3). Overall, while the short experiments with MVP 1 system did inspire baseline trust, users seem to have struggled to clearly grasp where the AI's boundaries are or what are the AI's limitations. Additionally, the answers are widely distributed due to users with different backgrounds in this group.	Users provided high ratings regarding all their perceptions of situation awareness (med=4).	Users provided high ratings for all perceptions of situation awareness (med ≥ 4).	Users provided high ratings regarding all their perceptions of situation awareness, although lower levels are reported in users' understanding of the AI's capabilities (med=3.5) and capability to predict the behaviour of the AI assistant (med=3). Overall, while the short experiments with MVP 1 system did inspire baseline trust, users found it difficult to really trust it as they did not realise the implementation part.

Table 10: Situation awareness during the interaction.

The aggregated scale score distribution is shown in Annex 3, where boxplots allow for a granular inspection of response concentration across constructs.

5. Discussion / Perspectives

This section discusses and **provides perspectives on the results** presented in the previous section. We provide **in-depth interpretations related to comprehensibility**, and additional insights on the AIA index as a prototype as well as limited trustworthiness without frailty. This section also presents the avenues for **prototypes improvements**, as reported by participants. Finally, we discuss **the experiment limitations**, such as simulated behaviours of prototypes, or prototypes being evaluated only on planification task and not while performing the planed route.

5.1 Results interpretation

5.1.1 Overall results interpretation

According to the qualitive feedback, we can say that for each use-case **most of the feedback is linked to the comprehensibility**. For City of Amsterdam and Proditec we can assume that there is a **lack of comprehensibility**, but if we look at the **quantitative results**, we can see that the **comprehensibility construct** got a **satisfying score** for both. **We can admit that the positive and negative feedback related to the comprehensibility of the prototypes was the easiest to highlight by the participants**. For the Sonae and Datacation use-cases, there is a lot of **positive feedback related to the comprehensibility** construct but there is also some relevant negative feedback, and axis of improvement mentioned by users.

Please note that responses to the AIA index items may be biased. Participants tend to respond based on their projections of what AI assistants might be like in the future, rather than responding based on their current state.

Among all the feedback linked to the comprehensibility construct, we can distinguish two types:

- **The comprehensibility of the prototype by the explanation**: During the completion of the scenarios, the prototypes give some explanations to the user. The aim of those explanations is to provide the user with the **necessary information to understand how the prototype works and what to do with it**.
- **The comprehensibility of the explanation itself**: The above-mentioned explanations are distributed to users during scenarios. For the prototype to be understandable, users must be able **understand that these explanations are explanations**. Moreover, these explanations must be **placed in the right places at the right times** and be easily accessible so that the user understands that they must use this information.

In this sense, the **interest of a chatbot** has been mentioned many times by users of Proditec, City of Amsterdam and Sonae use-cases as a mean of interaction with the prototype **to have access to the explanations**. According to the qualitative feedback from these three use-cases, being able to access the desired explanation at the desired time would have a **positive effect on comprehensibility**, and on the **collaboration** perceived by users between them and the prototype. For the Datacation use-case, participants have a mixed feeling about the chatbot because they were constrained into its use. For them, the chatbot could allow for question-answer interactions.

5.1.2 Focus on the AIA Index

As part of this project, our objective was to evaluate the mock-ups by creating an index designed to measure **trust** and **trustworthiness**. Since no such index currently exists, we developed one specifically for this purpose. As explained previously, the experimentation process was therefore not only aimed at assessing the mock-ups

but also at **iteratively improving the index itself**, based on user feedback. This study represents the first phase of experimentation and thus the first version of the AIA index. As we gather more insights, the way we evaluate the constructs will evolve (through the refinement, removal, or reformulation of questions, as well as the addition of new ones). Consequently, **the methods used to evaluate the constructs are expected to change over time**, leading to a more robust and reliable measurement tool.

The **lack of response** to certain items of the AIA index (*e.g.* auditability and ethics) can either suggest **potential ambiguity** in the item or **difficulty answering** due to a lack of item relevance for the respondent. Since MVPs were prototypes (simulation environment or Wizard-of-Oz mock-ups), this is not surprising that higher-TRL items (auditability and ethics) show lower response rates.

Moreover, **some missing values could be due to some constructs did not cover by the use-cases**. The constructs "safety, security and well-being", "privacy and data governance", "ethics", "auditability", and "anthropomorphism" were not addressed during the development of the MVPs. For these MVPs, greater priority was given to aspects considered to be more important, *e.g.* "user agency", "technical functioning", "comprehensibility", "collaboration", and "other components of user experience".

5.1.3 Trust judgment without vulnerability?

For each use-case, the **score of the AIA index is great**: 3.32/5 for Proditec, 3.51/5 for City of Amsterdam, 3.95/5 for Sonae, and 3.32/5 for Datacation. **But these results should be tempered**.

A key limitation in interpreting the results of the evaluation in the context of predicting trust or acceptance lies in the nature of the test setting. Indeed, **trust requires actual stakes in a situation characterized by uncertainty and vulnerability** (Lee & See, 2004), which means that PEER testers need to evolve in a meaningful situation, within a context of risk, irrespective of their ability to monitor or control the AI party.

During this series of evaluations, participants interacted with AI systems through mock-ups or simulation environment, **without actual consequences or risk**. As a result, while participants can evaluate perceived attributes of the AI assistant (*i.e.* how acceptable and trustworthy the AI assistant is), **the absence of vulnerability may have impeded the formation of true trust**.

This means that judgments of trust, trustworthiness or acceptance in this context may reflect hypothetical impressions which may vary across users and **differ from what they would perceive in a more realistic context**.

5.2 Perspectives on prototypes development

This section details for each use-case possible improvements for the development of MVP2.

5.2.1 Proditec

In general, the AI assistant must provide **more information and more explanations** about:

- **What the user must do**. The users need to be more guided to achieve their goals.
- **How it works**. The users need to know and understand how the AI assistant works: what data it uses and how it uses them to achieve an outcome.

Please note that simply adding explanations is not enough. These explanations must be **relevant and understandable** to users. The explanations also must be **displayed at the right time** to have a **better collaboration** between the AI assistant and the users.

Regarding **interface design**, **buttons (clickable elements)** should be **standardised** so that they are easily recognisable by users. Non-clickable elements should not look like buttons.

In addition to explainability aspect, other elements can be improved to make the **AI assistant more trustworthy**. The "**privacy and data governance**" aspect could be improved by knowing whether the AI assistant adequately protects the data collected and whether it is accessible. Furthermore, **users need to have access to the history of the data** (traceability of the information by the user).

Ethical aspects need to be more considered: have a non-discriminatory AI assistant that takes into account multiple languages and cultures, is accessible to people with disabilities, *etc.*

5.2.2 City of Amsterdam

In general, the AI assistant must provide **more explanations about how it works**. The users need to know and understand how the AI assistant works: what data it uses and how it uses them to achieve an outcome.

Please note that simply adding explanations is not enough. These explanations must be **relevant and understandable** to users. The explanations also must be **displayed at the right time** to have a **better collaboration** between the AI assistant and the users.

Regarding the interface design, and especially the map, the various **routes suggested by the AI assistant** should be displayed **directly on the map**, with **relevant information to distinguish between them** (total time, section on the cycle path, section on the footpath, footpath slope, *etc.*). and make a choice.

In addition to explainability aspect, other elements can be improved to make the **AI assistant more trustworthy**. The "**privacy and data governance**" aspect could be improved by knowing whether the AI assistant adequately protects the data collected and whether it is accessible. Furthermore, **users need to have access to the history of the data** (traceability of the information by the user).

Ethical aspects need to be more considered: have a non-discriminatory AI assistant that considers multiple languages and cultures, is accessible to people with disabilities, *etc.*

5.2.3 Sonae

In general, the AI assistant must provide **more explanations about how it works**. The users need to know and understand how the AI assistant works: what data it uses and how it uses them to achieve an outcome.

Please note that simply adding explanations is not enough. These explanations must be **relevant and understandable** to users. The explanations also must be **displayed at the right time** to have a **better collaboration** between the AI assistant and the users.

Regarding the interface design, a transition must be made from the simulation environment to a **mobile interface**. **The information provided and interactions must be adapted to such a device**.

In addition to explainability aspect, other elements can be improved to make the **AI assistant more trustworthy**. The "**privacy and data governance**" aspect could be improved by knowing whether the AI assistant adequately protects the data collected and whether it is accessible.

Ethical aspects need to be more considered: have a non-discriminatory AI assistant that is accessible to people with disabilities.

5.2.4 Datacation

The AI assistant must **provide explanations** regarding its functioning: what data are used? why it suggests these solutions? The percentages presented in the solutions could benefit from more context: on what are they based? or could be replaced by qualitative indicators. Moreover, the users need to be **more guided in the system**, especially in the "input menu", to achieve their goals. Visible and clear instructions of what users have to do can be add on the header of the different menus.

Please note that simply adding explanations is not enough. These explanations must be **relevant and understandable** to users. The explanations also must be **displayed at the right time** to have a **better collaboration** between the AI assistant and the users.

Regarding the interface design, the **architecture of the system** needs to be **rethought**. More **transparency** might be needed **in navigation**: for example, a breadcrumb trail can be added, the optimization menu can be divided in sub-menus or the chatbot feature can be put in another menu. The possibility of **displaying several elements at the same time** should be considered. Today, logistic engineers work with multiple documents, especially excel sheet, that they display simultaneously on different screens. In the "optimization menu", to compare easily the different solutions proposed by the AI assistant, the users need to see the two solutions proposed and their explanations at the same time.

In addition to explainability aspect, other elements can be improved to make the **AI assistant more trustworthy**. The "**safety, security and well-being**" aspect could be improved. Reorganising a warehouse involves risks: the greater the internal item movements and physical configuration changes is, the higher the risk of accidents is. Users should be alerted to these potential risks.

Ethical aspects need to be more considered: have a non-discriminatory AI assistant that takes into account multiple languages and cultures, is accessible to people with disabilities, *etc.*

5.3 Experimental limitations

5.3.1 TRL level and task realness

The MVPs evaluated during this first round of the PEER project relied on mock-ups with scripted behaviours, which introduces significant **limitations in interpretation** and, especially**, in participant engagement**.

For **Proditec use-case**, participants were **unfamiliar with the task of sorting coins**. They did not know what criteria to use to determine whether a coin was good or bad.

For **City of Amsterdam use-case**, **only the planning activity** was performed during the experiment. Participants were not aware of the generated route; only one participant reported knowing the proposed route. Some participants therefore reported that they **could not judge the quality of the proposed route**, thus limiting the scope of the overall reliability or at least the performance evaluation of the AI assistant.

For **Sonae use-case**, participants **interacted with a simulator** rather than with the client application where future versions are planned.

For **Datacation use-case**, participants did **the planification part** but **didn't do the implementation part**. So, it is **difficult** for them to **really judge the relevance** of the chosen solution. Moreover, participants were **restricted**

in their interaction with the chatbot. Due to time limitations, one user flow was mocked up. In fact, participants had to choose "solution 2", even though they all wanted to choose "solution 1".

As a result, user feedback from interviews and questionnaires may not fully reflect the feedback that would have been obtained with a more advanced, integrated version of the AI assistant. This is especially relevant since participant tend to form holistic views of the system, **where the boundaries between the AI assistant, the interface and other functionalities might not be clear**: *"So, I still didn't understand the AI assistant was it just the AI that will provide us with the hyperparameters? Or in the term assistant, we also take into account all the help part of the IHM?"*.

5.3.2 Panel of participants

The participant profile represents a notable limitation of this experimental phase. For the use-cases conducted at Proditec and Sonae, evaluations were carried out exclusively with **company employees**, introducing a **strong bias**: these participants are not necessarily representative of the **target end-users**.

For Sonae, participants had a **very positive opinion of AI in general**, **more than** Proditec, City of Amsterdam and Datacation's participants.

In the case of the City of Amsterdam, while the participants were more aligned with the **intended user group**, most of them were in the **50–59 age range**, raising questions about the **representativeness** of this sample within the broader target population.

Furthermore, the limited number of participants in each use-case, especially in Datacation use-case, reduces the generalizability of the findings and calls for caution in interpreting the results.

5.3.3 Experiment facilitation and support

In **City of Amsterdam** use-case, the experimental sessions were facilitated by a local staff member acting as the **experiment facilitator**. This approach was necessary due to the language barrier, as the participant spoke Dutch while the evaluation team did not. The facilitator conducted the sessions while we were presented to observe, following detailed instructions provided beforehand to ensure methodological alignment. While this arrangement enabled the evaluations to be conducted smoothly, it also **introduced potential variability in the delivery of the protocol**: since the protocol was designed by our team, it was naturally more challenging for someone external to conduct the sessions without the same in-depth understanding of its rationale and objectives. This lack of familiarity may have impacted the depth of probing during the interviews. **For future iterations**, it may be beneficial to **plan more joint preparation sessions**, particularly focusing on interview techniques, to ensure that facilitators are fully equipped to capture the most relevant and detailed feedback.

In a different use-case, part of the **mock-up was modified during the course of the experimentation**. While such adjustments may be necessary in certain contexts, they can also **introduce inconsistencies in the collected data** and **affect the comparability of results across participants**. This situation highlights the **importance of structured guidance and methodological consistency during experimental phases**. For future iterations, it will be essential to reinforce CATIE's support in this area, both in terms of anticipating potential design adjustments before the testing phase and providing real-time methodological advice if changes become unavoidable. Strengthening this collaboration will help ensure that prototype modifications are managed in a way that preserves data validity while still allowing for iterative improvement.

6. Deviations occurred and corrective actions taken

This section presents the **deviations** that occurred during T5.7 regarding the Grant Agreement and the actions taken to correct these deviations.

6.1 The fourth use-case was delayed

During the first year of the project, **the use-case Continental was removed from the project**. Datacation was found to replace it, but **this delay has impacted the timing of T2.1**, namely identifying the socio-technical requirements as well as T5.1 and T5.3 to design and develop the prototype.

Consequently, the evaluation for Datacation in T5.7 has been delayed - granted by the European Commission. Contrary to the first three use-cases (Proditec, City of Amsterdam and Sonae), for which evaluations took place in April/May as scheduled in the Grant Agreement, Datacation's evaluations took place in September. This leads to the submission of version 2 of deliverable D5.6.

6.2 Fewer participants than expected

While the original grant agreement identified a target of **at least 10 participants in each experiment**, this figure was presented as a **preliminary estimate** to supported as a preliminary estimate to support initial planning. Upon further reflection and alignment with the project's core goals, there was a need to reassess this number. As reported in the previous section of the present deliverable, only one experiment involved more than 10 participants, namely Proditec with 11. We recruited 7 participants for the City of Amsterdam, 6 participants for Sonae, and 4 participants for Datacation.

First, the technologies evaluated in the project are MVPs, therefore TRL3 prototypes. **In the evaluations of such MVP, we want to collect information as to how to improve the prototype**, and only when we reach TRL6, we can recruit larger groups of participants to assess their impact. It is not the objective in the PEER project. We stop the development of the prototype at TRL5.

Second, **we aim at collecting rather qualitative than quantitative data**. We are searching for ways to improve prototypes rather than measure exact impact on representative populations**. Consequently, only few participants are required to lead to relevant insights**. The **methodology adopted to process the gathered data, namely the thematic analysis, is particularly relevant** in this search. In qualitative evaluations such as user testing, it is generally accepted that 5 participants are enough to uncover 85% of total usability problems in a product (Nielsen & Landauer, 1993).

Therefore, it appears that **involving less participants than expected** in the Grant Agreement is mitigated regarding the type of information gathered in T5.7 as well as the methodology adopted to process these data. For the next evaluation cycles, for MVP2 and MVP3, it is not a problem if the number of participants recruited per use-case is less than 10.

6.3 Early results presented to speed up prototype improvements

At a project management level, the outcomes of T5.7 are meant to orient the **next phase of development** of the AI assistant prototypes (*i.e.*, improvements based on user testing). The present deliverable was due for M24, and the next version of the prototypes due 6 months later, at M30. This timing can be problematic for AI researchers and partners responsible for **developments and integration** because they cannot ensure they will resolve scientific challenges, therefore impeding the developments based on these findings for the project to produce improved versions of the prototypes to be evaluated again 6 months later.

Consequently, CATIE offered the consortium **to speed up data analysis** by focusing first on the elements directly linked with prototypes improvements. In other words, we focused on the thematic analysis of participants *verbatim* and left processing information from the AIA index and the socio-demographic questionnaire for the present deliverable. **These early results have been processed immediately** after the experiments, and presented to the consortium in early M22, so that **partners gained 3 months on the avenue to pursue to improve the prototypes**. In-depth analyses presented in this deliverable are fully aligned with these early results.

We suggest that future iterations of experiment/improvements follow this time structure, with a middle milestone of early results and following comprehensive analysis. **We highlight the relevance of this organisation for future European projects as well.**

7. Conclusion

This deliverable D5.6 “Report on the evaluation process” results of the first iteration of T5.7 “Management of the evaluation process” in WP5. It aimed at evaluating MVP1 of each use-case of the PEER project.

Experimental methodology was designed to collect qualitative data from thinking aloud when using mock-ups, answering surveys and discussing with the experimenter in semi-structured interviews. Collected data concerned first the use of the prototypes by the participants, and second the usage of the AIA index developed in WP4. The second will be presented thoroughly in D4.2 “AIA Index: definition and methodology” due in M36 of the project.

A total of 28 participants were recruited across four use-cases, leading to the collection of almost 1000 *verbatims* which were analysed thematically. Based on this analysis, we provide the actual results for MVP1s in terms of notions such as comprehensibility, human-AI collaboration, and intention of use, as well as potential improvements suggested by participants themselves and direct observations from experimenters.

Overall, the present report exhibits promising results regarding MVP1s, especially considering they were only prototypes, mock-ups where AI behaviours were simulated to the participant. We report **two main avenues for improvements.** The first avenue for improvement is the **comprehensibility of AI functioning and capabilities** to participants. Not always that they understood how the AI assistant was able to help them or even how it worked. The second avenue for improvements is the **collaboration between humans and AI assistants**. Future users will benefit from better explanations from AI assistants, and especially well-timed and well-designed explanations to offer added value rather than additional information, potentially hard to process.

These comprehensive results will serve two purposes. **First, they will be directly used to design, arbitrate and develop improvements for each use-case for their next version**, due at M30. These versions will not be simulated behaviours but fully developed prototypes to be usable in controlled conditions (TRL4). **Second, these results will help improving the AIA index** developed in the PEER project. Notions and items will be studied in detail to serve WP4 objectives, for deliverable D4.2 due at M36.

8. Bibliography

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Nielsen, J., & Landauer, T. K. (1993, May). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 206-213).

Fage, C., Durnerin, E., Gardette, J., Roc, A., & Maranne, A. (2025). PEER D4.1 - Measurement scales and metrics for trusworthy AI. Zenodo. https://doi.org/10.5281/zenodo.15013450

Annex

Annex 1: Score distribution per construct

The AIA index revolves around 10 constructs and is composed of 59 questions. To better understand the factors that shape the scores obtained through the AIA index, **we examine the 10 underlying constructs**, helping to identify which attributes are consistently strong and which ones may need further attention. The four graphs below (Figure 43, Figure 44, Figure 45, Figure 46) display the detailed scores obtained for each construct within boxes showing the interquartile range (IQR). Individual points outside this range represent potential outliers. Text annotation on each boxplot highlights the central tendencies (median values) as they are more robust and less sensitive to outliers than means, therefore more representative of the typical response.

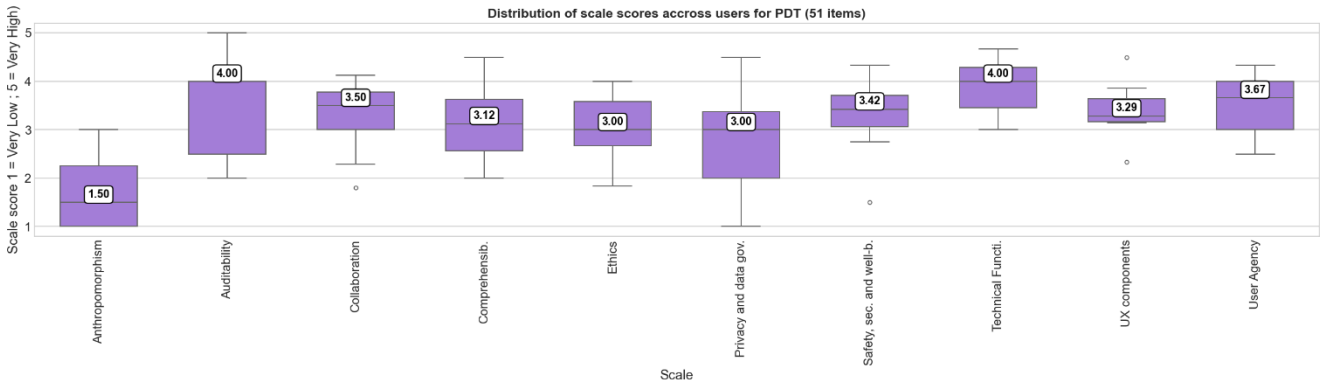


Figure 43: Scale-level results distribution (and median) for Proditec use-case.

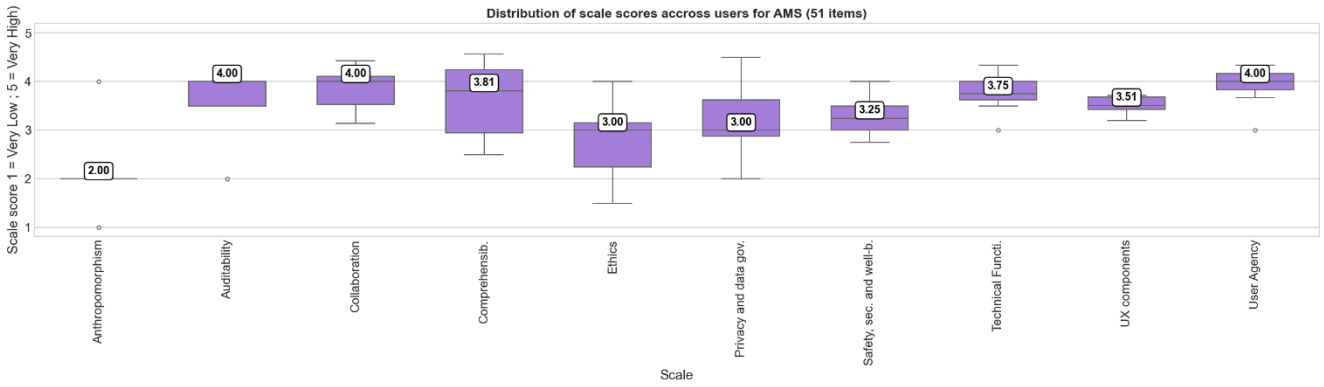


Figure 44 : Scale-level results distribution (and median) for City of Amsterdam use-case.

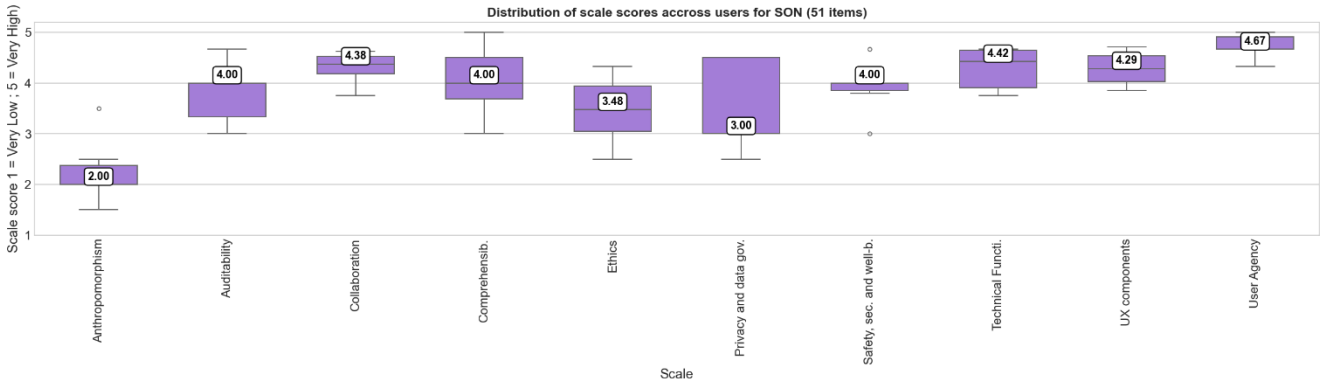


Figure 45: Scale-level results distribution (and median) for Sonae use-case.

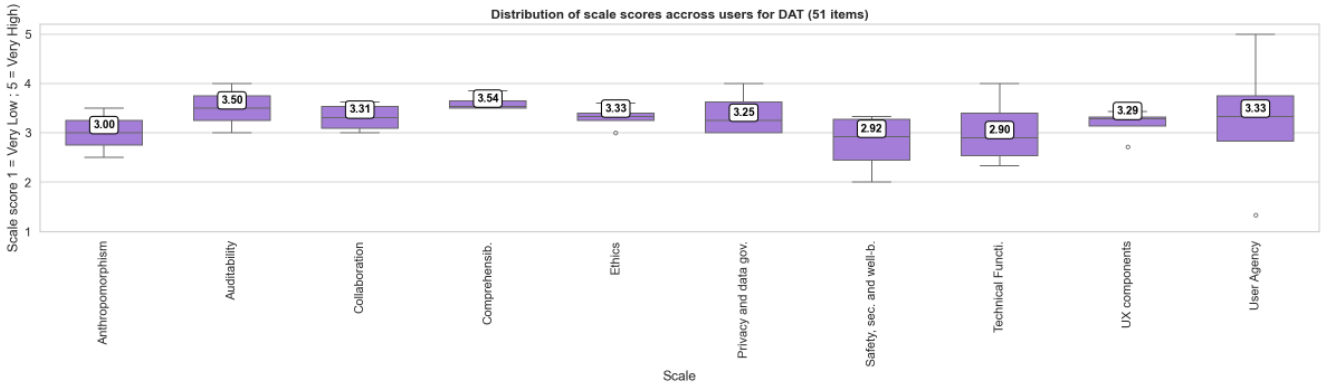


Figure 46: Scale-level results distribution (and median) for Datacation use-case.

Annex 2: Answer distribution per item

This graph (Figure 47) presents the complete distribution of answers through the questionnaire, where the colour intensity reflects the proportion of respondents selecting each value. Each column corresponds to one questionnaire item, and the rows represent the possible Likert response options (1 = strongly disagree to 5 = strongly agree). The heatmap cells show the percentage of respondents selecting each response option, with darker colours indicating higher frequencies. Percentages are annotated within each cell for clarity. This visualization allows for easy comparison of how users rated different items and highlights both central tendencies and variability across responses

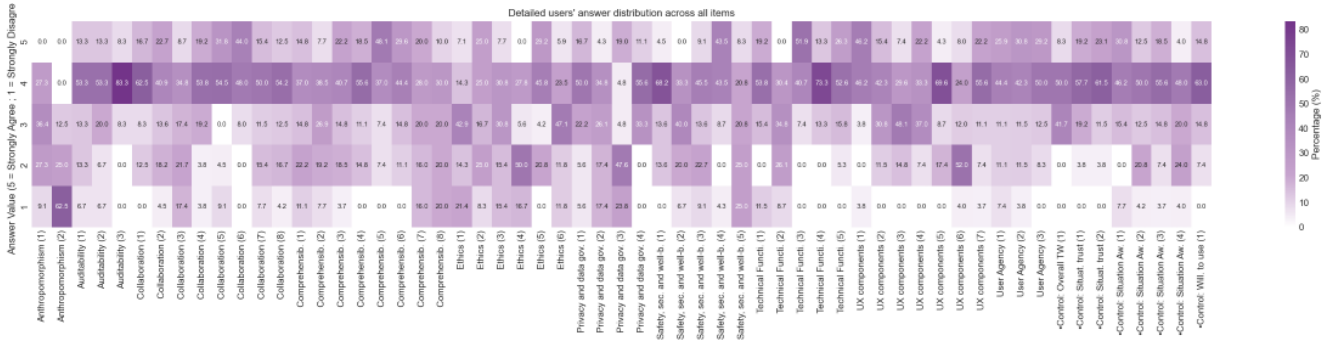


Figure 47: Distribution of users' responses to Likert-scale items.

Annex 3: Score distribution of control variables

To complement the main AIA questionnaire, **eight additional items are included as a control** along the 10 constructs of perceived trustworthiness. The four graphs (Figure 48, Figure 49, Figure 50, Figure 51) below display the detailed scores obtained for each individual item within boxes showing the interquartile range (IQR). Individual points outside this range represent potential outliers. Text annotation on each boxplot highlights the central tendencies (median values) as they are more robust and less sensitive to outliers than means, therefore more representative of the typical response.



Figure 48: Distribution (and median) of control items for Proditec use-case.



Figure 49: Distribution (and median) of control items for Amsterdam use-case.

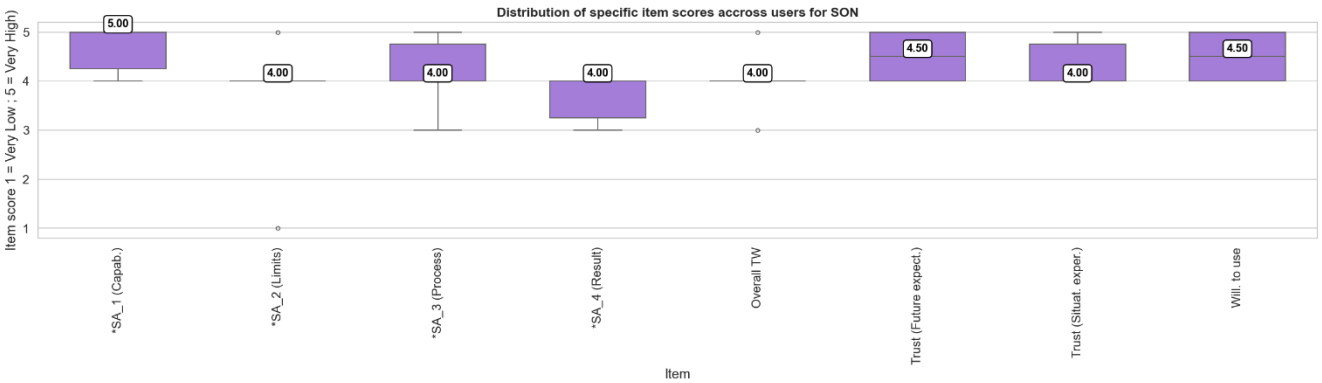


Figure 50: Distribution (and median) of control items for Sonae use-case.

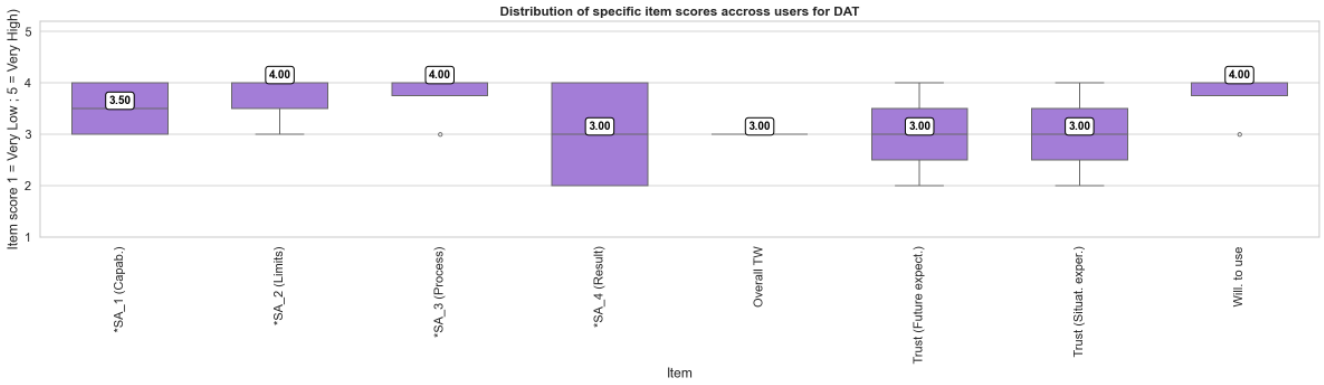


Figure 51: Distribution (and median) of control items for Datacation use-case.

PEER will focus on how to systematically put the user at the centre of the entire AI design, development, deployment, and evaluation pipeline, allowing for truly mixed human-AI initiatives on complex sequential decision-making problems. The central idea is to enable a two-way communication flow with enhanced feedback loops between users and AI, leading to improved human-AI collaboration, mutual learning and reasoning, and thus increased user trust and acceptance. As an interdisciplinary project between social sciences and artificial intelligence, PEER will facilitate novel ways of engagement by end-users with AI in the design phase; will create novel AI planning methods for sequential settings which support bidirectional conversation and collaboration between users and AI; will develop an AI acceptance index for the evaluation of AI systems from a human-centric perspective; and will conduct an integration and evaluation of these novel approaches in several real-world use cases.





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