UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering,
Mathematics & Computer Science

Enhancing AI Assistants for Routine Tasks in the Workplace: The Role of Explainable AI in Technical and Human-Centred Way

Ruben Koole
M.Sc. Thesis Interaction Technology
July 2025

Supervisors:
dr. ir. D. Reidsma
dr. J. L. Rebelo Moreira

External Supervisor:
M. van 't Sant

Planon
Building Connections

Contents

List of acronyms **vii**

1 Preface **1**
1.1 Abstract 1
1.2 Acknowledgments 2
1.3 Disclaimers 2

2 Introduction **3**
2.1 Objective 4
2.1.1 Research Questions 4
2.2 Methodology 6
2.3 Thesis structure 6

3 Artificial Intelligence Assistants **9**
3.1 Types of AI Assistants 9
3.1.1 High Severity Impact 9
3.1.2 Conversational AI 10
3.1.3 Art 10
3.1.4 Automated Writing Evaluation 11
3.1.5 Routine Tasks 12
3.2 Who Needs Explanation 13
3.2.1 High Severity Impact 13
3.2.2 Customer Service Chatbot 14
3.2.3 AI Marketing 14
3.2.4 Automated Writing Evaluation 15
3.2.5 Routine Tasks 15
3.3 State of the Art of Routine Task AI Assistants 16
3.3.1 Task Scheduling 16
3.3.2 Customer Service 17
3.3.3 Inventory Management 18
3.3.4 Office Space Management 18

4 Explainable AI **21**
4.1 Problem . . . 21
4.1.1 Definition of Explainable AI . . . 22
4.2 Typology of Explanations . . . 22
4.3 Methods of Explainable AI . . . 25
4.3.1 LRP . . . 27
4.3.2 LIME . . . 27
4.3.3 SHAP . . . 28
4.4 Human-Centred Explainable AI . . . 29
4.4.1 Design Requirements . . . 29
4.4.2 XAI-Interaction Types . . . 31
4.4.3 End User Control . . . 32
4.5 Evaluating Explainable AI . . . 33

5 The Company Planon **37**
5.1 Introduction . . . 37
5.2 Software . . . 37
5.2.1 IWMS & Campus Management Solution . . . 38
5.2.2 Facility Services Business Solution . . . 38
5.2.3 Lease Accounting Solution . . . 39
5.3 Case Studies . . . 39
5.4 Connection to the Preliminary Research . . . 40

6 Ideation **41**
6.1 Method . . . 41
6.2 Results of the survey and interview . . . 41
6.3 Ideas . . . 45
6.3.1 Brainstorming Ideas . . . 45
6.3.2 Survey Influenced ideas . . . 47
6.3.3 Design space . . . 49

7 Designed Prototypes **51**
7.1 Design Method . . . 51
7.2 Visuals . . . 52
7.2.1 Transmission Global . . . 55
7.2.2 Transmission Local . . . 56
7.2.3 Dialogue Global . . . 57
7.2.4 Dialogue Local . . . 58

8 Evaluation **59**
8.1 Objectives . . . 59
8.2 Evaluation Plan . . . 59
8.3 Results . . . 60

9 Final Reflections **69**
9.1 Conclusions . . . 69
9.2 Contributions . . . 70
9.3 Discussion and Future research . . . 71

References **73**

Appendices

A XUI Mind Map Unfolded **81**

B Questions Survey Ideation Phase **83**
B.1 General AI Experience . . . 83
B.2 Explainable Artificial Intelligence . . . 83
B.3 Usage of the Workplace App . . . 85
B.4 Human Assistance . . . 85

C Ideation Images **87**
C.1 Settings . . . 87
C.2 Global explanation . . . 88
C.3 Local explanation . . . 89
C.4 Input variable . . . 90
C.5 Satisfaction . . . 91
C.6 Help function . . . 92
C.7 Dialogue Explanations . . . 93
C.8 Homepage . . . 94

D Prototype Images **95**

E Scenario Number generator Code **99**
E.1 Code.gs . . . 99
E.2 Index.html . . . 100
E.3 JavaScript.html . . . 101

F User test **103**
F.1 Explanation Satisfaction Scale . . . 103
F.2 Self-made Questions . . . 104

F.3 Tasks 104

G Additional Analysis output **105**

List of acronyms

AI	Artificial Intelligence
HCI	Human-Computer Interaction
HCXAI	Human-Centred Explainable Artificial Intelligence
XAI	Explainable Artificial Intelligence
UX	User Experience
IWMS	Integrated Workplace Management System
EI	Explainable Interfaces
XUI	Explainable User Interface
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layer-wise Relevance Propagation
SHAP	SHapley Additive exPlanations
CAFM	Computer-Aided Facility Management
ML	Machine Learning
ESS	Explanation Satisfaction Scale
EGC	Explanation Goodness Checklist

Chapter 1

Preface

1.1 Abstract

In a world where Artificial Intelligence (AI) is developing rapidly, there is a need for people to understand what the AI is doing. Especially if we use AI in our job and routine tasks, that is why the field of Explainable Artificial Intelligence (XAI) arose. There are various techniques to explain how an AI works both technically and how to present the explanation. This thesis investigated the user needs for an Explainable User Interface (XUI) for an AI assistant in the domain of routine tasks. First delving into the technical side of XAI and investigating what guidelines there already are for making an XUI. A survey is made to find key user needs among end users of the application. Together, based on literature and user research, four prototypes are made and evaluated among the end users. This evaluation indicated a preference in a dialogue interaction for the explanation. If users want a local or global explanation, it is based on the own preference of the user. The assessment additionally gave detailed improvement suggestions for a better user experience of the overall interaction with the XUI. Overall, these findings are promising and should be considered in further research for XUI's for routine tasks.

1.2 Acknowledgments

This thesis was not possible without the help of various people. First up, I want to thank my supervisors Dennis Reidsma, Maartje van 't Sant, and João Rebelo Moreira for their advice, guidance, and enthusiasm throughout this project. I appreciate the time you all invested in me and let me do this project.

Additionally, I want to thank my friends and family for the support during the process and the sometimes difficult times. And also my colleagues at Planon who made me feel welcome in the team and made my time there so pleasant and the long train rides worthwhile.

Finally, I want to mention all the participants who took time out of their busy schedules to fill out my survey and participated in my user test.

1.3 Disclaimers

The chapters 2 to 5 of this report are based on the previously made "Research Topics" report [39].

Portions of this thesis were supported by the use of AI-based tools, to improve the research quality. The tool Writefull[1] is used for LaTeX formatting and is sparingly used for writing, aiding in correct grammar, restructuring and condensing of sentences.

[1] https://www.writefull.com/

Chapter 2

Introduction

In recent years, AI has progressed rapidly. Initially, chatbots appeared, supporting tasks such as customer service or personal assistance through platforms like Alexa, Siri, and Google Assistant [33, 78]. However, these early chatbots incorporated minimal AI. In 2022, OpenAI launched ChatGPT, a new chatbot for public use. ChatGPT is an advanced chatbot capable of generating detailed responses based on prompt instructions [81]. Its responses can be either textual or visual. In the wake of ChatGPT's release, multiple firms launched their own iterations or enhanced existing conversational agents, such as Microsoft's Copilot and Google's Gemini.

In addition to the emergence of AI chatbots, developers are also implementing AI in various software applications, such as agenda scheduling [36, 37] and AI-assisted code development [70, 73, 59, 60]. Furthermore, it supports autonomous vehicles, serves as a decision-making aid in medicine, and finds numerous other applications.

Next to AI applications as assistants, there has been increasing enthusiasm from artists, technologists, and researchers to delve into the creative possibilities of AI. Tools like ChatGPT and generative adversarial networks significantly accelerated the use of AI in creating visual art [81, 8]. By providing a prompt, an AI assistant is capable of producing not only high-quality images but also videos and music.

As companies increasingly integrate AI assistants into their software, users become accustomed to and anticipate their presence in other applications. Companies, recognizing this trend, feel pressured to adopt AI assistants to remain competitive.

Incorporating AI assistants into interactive applications necessitates careful attention to possible challenges. Given that AI does not perform optimally across all areas, it can fail. This might lead to ethical concerns like privacy violations, security threats, inappropriate content delivery, or even traffic accidents [1]. This prompts the question, 'Who should be held accountable?'. Furthermore, economic issues might arise since companies may face financial setbacks if an AI assistant malfunc-

[1] Incident database: `https://incidentdatabase.ai/apps/incidents/`

tions. Usability challenges may emerge from a User Experience (UX) perspective. Specifically, businesses could experience financial losses if the AI assistant fails. Moreover, from a UX angle, if the implementation of AI assistants isn't user-friendly, it could lead to usability concerns.

Understanding *how* and *why* an AI assistant makes decisions is crucial for identifying when issues occur, allowing explanation of the underlying reasons. However, determining how these explanations should be structured to make the AI's decisions comprehensible remains challenging. Given the diversity of AI assistants and the variety of stakeholders involved, each requires a distinct form of explanation.

That is why, since the last decade, there has been more research done in the field of XAI [43]. XAI is a research field that aims to make AI systems' results more understandable to humans [2, 35].

This thesis identifies various types of AI assistants by their functionality and compares them in terms of the necessity for clear explanations. Additionally, the study assesses multiple XAI methods, analysing their technical details and how explanations can effectively communicate AI decisions to users. Importantly, the research focuses on AI that assists with routine workplace tasks, a relatively under-researched area compared to fields like healthcare or autonomous driving. Given the anticipated growth of such AI applications, it is crucial to establish clear guidelines for explaining AI-driven decisions. This study is applied within the context of Planon, a company specialising in smart sustainable building management software, including workspace management solutions, to create examples on interfaces and to evaluate the constructed prototypes.

2.1 Objective

This thesis aims to identify essential user needs and design criteria for developing an XUI tailored for AI assistants handling workplace routine tasks, as well as to explore design strategies within this particular field.

2.1.1 Research Questions

In order to tackle this challenge, the main research question for the final project will be:

- ***"How can an XUI be designed to enhance the usability of AI assistants used for routine tasks in the workplace?"***

This research question is addressed in a number of steps. The following steps make use of two sub-questions.

Initially, it is crucial to check what is already out there in the literature of this domain. Making a foundation of knowledge on AI assistants, XAI and Human-Centred Explainable Artificial Intelligence (HCXAI). Then we need to check what is found in the literature holds up in practice. And explore specific needs and expectations within the domain, as the majority of existing literature tends to be generic. This will be done by conducting a survey to quickly get many participants. Giving an answer to the following question:

- *"What are the key user needs and expectations when interacting with AI assistants for routine tasks through an XUI?"*

Once the key user needs and expectations are established, the following step involves generating various concepts for designing an XUI tailored to the specific domain of routine tasks and within the particular software Planon. This will be done while definitely considering the guidelines outlined in the literature along with the findings from the user analysis survey. This will be done by brainstorming and input sessions with UX designers. From those ideas, Lo-Fi (Low Fidelity) prototypes will be made.

Once the initial prototypes have been completed, it is essential to evaluate the proposed designs. To see what aspects are important in an XUI. These designs will be evaluated with a user study and giving an answer to the following question:

- *"What are important design requirements and considerations to take into account when designing an XUI for AI assistants for routine tasks in the workplace?"*

After the evaluation of the prototypes, the aim is to effectively address the main research question and make a substantial contribution to the development of XUI, with the focus on the domain of routine tasks in the workplace, considering not only the application within Planons software but also the general routine tasks present in the workplace. A reflection on the future work on the development and research of this topic will conclude the thesis.

2.2 Methodology

The methodology that is used in this thesis is the Design Thinking process launched by Stanford Design School (D.school) [65]. This method has five stages: empathize, define, ideate, prototype, test, and assess; see Figure 2.1.

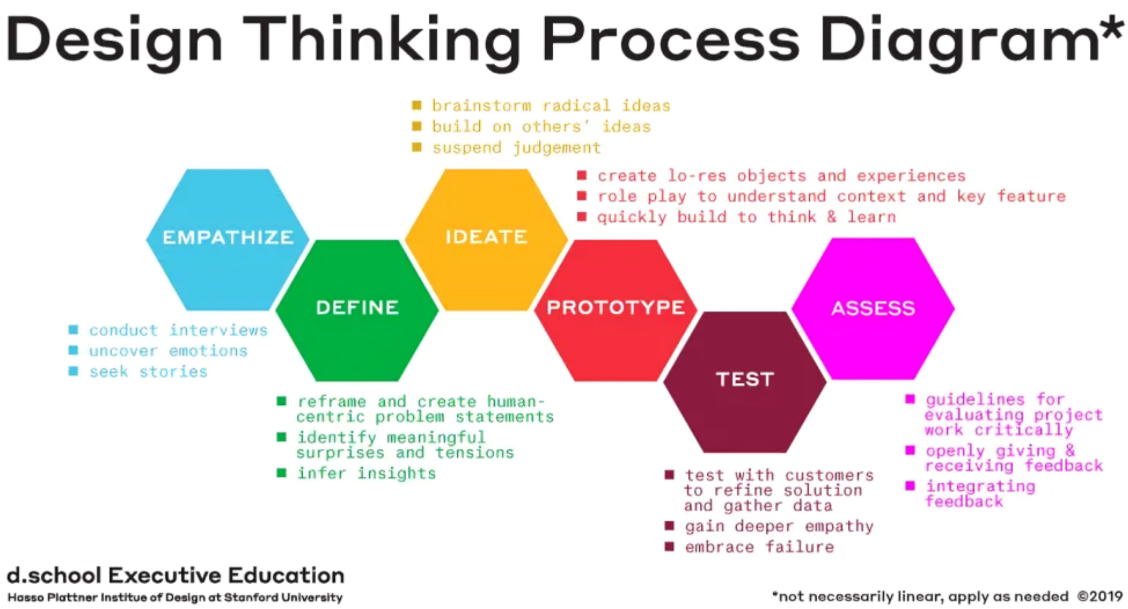


Figure 2.1: Five steps of the Design Thinking Process Diagram [65]

As the asterisk already indicated in Figure 2.1, the method is not linear and applied as needed. The method is adjusted slightly to align more effectively with this thesis. Initial empathizing and defining are conducted through literature, which also serves to enhance comprehension of the XAI concept. A subsequent round of empathizing and defining is performed with the application's end users to delve deeper into the user needs. Using insights from both rounds, the steps of ideation, prototype, test, and assess are carried out.

2.3 Thesis structure

The thesis is organised into different sections. The first chapter delves into the different types of AI assistants that are out there and categorises them. Then, per type of AI assistant, the different stakeholders are analysed for who wants an explanation of the assistant and what they expect from an explanation. Examples serve as the foundation for this. This chapter ends with the state of the art of different AI assistants that are currently out there.

The second chapter gives an introduction to XAI, in its definition and different ways of executing XAI in a technical manner. It also delves into the typology of explanations themselves. The chapter follows with a section on guidelines, suggestions, and (design) requirements for HCXAI[19, 79].

Since this graduation project is applied in the software of an external company, Chapter 5 introduces the company Planon, a software company that focuses on software for smart building management. It investigates its different software with the use of case studies from clients of Planon. This chapter concludes by describing the relationship between Chapters 3 and 4 and their connection to Planon.

This is followed by the chapter of the ideation where the first ideas are made for the prototype. Based on literature and user research among employees of Planon, brainstorming sessions are held, and a design space is created.

Chapter 7 talks about the designing part of the prototypes and Chapter 8 is dedicated to presenting the user evaluation methodology and results.

Lastly, Chapter 9 summarises and discusses the thesis's findings, outlines its academic contributions, and provides recommendations for future research.

Chapter 3

Artificial Intelligence Assistants

This chapter of the thesis will discuss different types of AI assistants with examples. The assistants are categorised and grouped together based on their functionality. Moreover, each type will be evaluated on whether utilizing this type of AI assistant results in ethical and/or economic implications.

Next, the nature of frustrations, per type, will be explored. The types of frustrations are split into two types of frustration. Frustrations when something goes wrong/that is not intended, and frustrations when people do not know what is happening.

After that, a section about *Who wants explanation* is added for each type of AI assistant where different stakeholders are analysed in what kind of explanation they want and expect. This is followed up by a state-of-the-art section, exploring what applications are already out there for AI assistants for routine tasks.

3.1 Types of AI Assistants

To categorise the different types of AI assistants, they are divided into main functionalities. In addition to that, the first group consists of AI assistants that all can have a significant impact on human lives. In the sections below, each different type is explored.

3.1.1 High Severity Impact

This category of AI assistants significantly influences human lives. The potential high impact arises from the risk to human safety if the assistant's decisions are faulty.

The AI assistants in this category include self-driving cars. Significant progress is being made in developing autonomous vehicles such as cars and buses [9, 6, 24].

Next to vehicles, this category includes hospital diagnostic assistants and military drones, all capable of making life-dependent decisions.

Ethical issues arise with this set of assistants due to the involvement of human lives, raising crucial questions such as, *'Who is responsible?'* For example, autonomous vehicles: is it the passenger, the "driver," or the manufacturer? Numerous accidents involving self-driving cars, including those with pedestrians, are documented in an AI incident database[1]. Another ethical issue is bias; as AI assistants enter medical fields, the symptoms can sometimes differ between races, which could lead to misdiagnosing patients [21].

In the medical field, frustration with AI assistants largely stems from their failure to grasp divergent outcomes. Additionally, the lack of insight into the AI's functioning contributes to this frustration [40].

3.1.2 Conversational AI

Conversational AI assistants are likely the most encountered type of AI as they are becoming more famous in the day to day lives [31]. They respond to user prompts by addressing queries and often incorporate previous interactions into their responses.

These assistants are also often referred to as chatbots or voice assistants [33]. Examples of these chatbots are Google Assistant, Siri, Alexa, and Bixby. Furthermore, chatbots are nowadays also used for customer service. In addition to these chatbots, there are generative AI chatbots like ChatGPT and Gemini.

Ethical issues arise from chatbot responses. Despite programming to avoid offensive or harmful replies, users can manipulate prompts to get to the undesired answers. Incorrect information from customer service chatbots may cause poor customer experiences, risking customer loss and financial challenges for companies. There was also an incident where a chatbot suggested to a user to take his own life [49]. Clearly, an unwanted and untechnical occurrence. Thus, one could ask, *'How could this have happened?'*

Frustrations with conversational AI assistants, and especially voice assistants, are quite feasible. Voice assistants can frustrate users when they fail to complete tasks, perform actions not based on the user's input, need to repeat commands, or if the assistant is keeping silent [32].

3.1.3 Art

Another AI assistant that can be prompt-based, are the assistants that create art. The term art is very broadly interpreted in this context. That is because AI can create

[1] Incident database: `https://incidentdatabase.ai/apps/incidents/`

a lot of different types of art. And defining what art is, is almost impossible [16].

The current AI assistants can, for example, make an image based on a prompt that a user enters. DALL-E from OpenAI [84, 81] is an assistant that is widely used. Since DALL-E is from OpenAI, ChatGPT has incorporated the functions of DALL-E. Next to making images, AI assistants can also make videos and music.

Exploring the ethical issues of using AI in art creation. As AI advances, the images and videos it produces are increasingly lifelike, leading to potential deception, commonly known as 'deepfakes' [84]. Such misinformation can influence public perception, altering views on subjects or individuals.

In addition to the deepfakes, another concern with making art with the help of an AI assistant is the copyright issue, who is actually the rightful owner of the image/video/music? Is it the person who entered the prompt or the company of the AI assistant?

Other ethical implications include issues related to bias and discrimination, privacy, job displacement, and unintended consequences [84].

According to Elgammal [20], artists might use these types of AI assistants to experiment with gaining interesting results, but too often quit using them due to the approach, which results in frustration among the artists.

In terms of other end users, they are mostly frustrated when the expectations differ from the output of the AI assistant. In addition to that, the long time it takes to generate an image is frustrating [44] or the new skills and workflow needed to use such an assistant can be frustrating [22].

3.1.4 Automated Writing Evaluation

Following this is the category of AI assistants that help users in enhancing text. Commonly employed for drafting reports or emails, these tools are also termed Automated Writing Evaluation [17].

This group of AI assistants includes assistants that help with spotting spelling and grammar mistakes in texts, or it could give your message a different tone by rephrasing and restructuring the texts. But if everybody uses these AI assistants, it could create only bland texts without a personal touch. There also exist AI assistants that try to humanize the text written by other AI assistants in order to not be detected as AI generated text.

Ethical issues may vary based on usage context. In general, quick (personal) messages create no significant concerns. However, in professional settings, ethical challenges may emerge. For instance, using extensive AI assistance in report writing may lead to questions about originality [3]. Khalifa and Albadawy [38] also highlight potential accuracy problems with references and misinformation.

Studies have found that using textual improvement AI assistants can also lead to frustration. The automated feedback can be insufficient and unspecific, which provides little help to users' writing process [12]. Also, their longer-term language development can be adversely affected by using these types of AI assistants, leaving the users frustrated [56].

3.1.5 Routine Tasks

The last type of AI assistants are the assistants that help with routine tasks in the workplace. Conceptually, a highly structured, routine task is likely a good fit for automated, efficient technology [62]. There are a lot of different routine tasks that can be automated with the use of an AI assistant. For example, scheduling tasks, room booking for a meeting or inventory management. Generally, it corresponds to tasks which follow a specific set of rules or procedures that make them suitable for AI automation [53].

An AI assistant can help with the agenda management of the workers, in order to have an efficient planning of their day. Or booking a meeting room, or a desk in some workplaces, can be made more efficient with an AI assistant. Other examples of routine tasks include filling the stock, processing invoices/documents, reporting broken facilities, etc. In short, all the little side tasks that a worker has to do next to their job description.

The use of AI assistants for task automation can lead to economic worries, first of all, if the AI assistant does the job, fewer people are needed in the company resulting in the unemployment of more people. But also when errors occur in the AI. Consider the stock replenishment scenario; if the company fully trusts the AI assistant and it fails to have enough stock of a product, the company may be unable to fulfill orders. This can result in customer dissatisfaction and potential repercussions like compensating customers or losing their business. And with fewer people working in the company, it is harder to rectify these errors.

While searching the literature on integrating routine workplace tasks with AI assistants, few papers emerged. Also, research on frustrations for AI assistants in routine tasks is therefore limited; one can imagine that in terms of (task) scheduling, it can be frustrating if one's planning changes all the time or if the meetings are scheduled in places far away from each other. During a study [41] it was found that most participants perceive AI assistants as insufficiently able to handle tasks to reach the expected standards and describe them as not quite "smart" enough to meet their specific needs.

3.2 Who Needs Explanation

As mentioned in the sections above, there are some ethical and economic concerns about the use of AI assistants. In addition to these issues, users often feel frustrated when they cannot comprehend the decisions made by the assistant when errors occur. Explanations enable them to grasp *why* the AI assistant created such an outcome. But of course, there is not one-solution-fits-all for the explanation; there are different stakeholders involved with the same assistant. Each different stakeholder wants a different explanation of the behaviour of the assistant. To identify those differences, examples from Section 3.1 will be used to make a stakeholder analysis. For each kind of AI assistant, a single example is provided, as various stakeholders are associated with each type.

3.2.1 High Severity Impact

Firstly, consider the high-impact assistant for severe cases. For autonomous vehicles, various stakeholders seek different explanations from the AI assistant, particularly when issues arise or the AI's decisions are misinterpreted. Four groups of stakeholders are identified as the following: users, legislators, operators, and manufacturers [27]. These different stakeholders can give insight in the importance of different explanations for different stakeholders.

On a day-to-day basis, the user of the car seeks an explanation of the decisions the car makes to alter its driving behaviour. Drivers who may not have direct involvement in the management of the autonomous vehicles should be able to instantaneously request accounts as intelligible explanations for such undesired actions when they occur [57]. This can include taking a different route to get to the destination faster. Furthermore, the manufacturing company and operator are also stakeholders in the day-to-day usage since they use the car's decision-making information to enhance the performance of the AI system for further development and to improve their operations [61].

However, additional stakeholders will seek clarification when situations go terribly wrong. As mentioned in Section 3.1.1 there are incidents that involve collisions with humans or other cars. In this scenario, key parties include the police and possibly a judge, the autonomous vehicle's "driver," any other party involved, insurance firms, impacted traffic, and naturally, the car's manufacturer.

The questions that all stakeholders have are: *'How could this happen?'* and *'Who is responsible?'* In order for legislators to make a good judgment call if they need to change policies, they need to know exactly what the situation was and what every party did, including the car. In addition to that, they need to ensure that the system satisfies ethical requirements and adheres to all legal requirements [26]. Was an attempt made by the vehicle to avoid the collision, and did the driver attempt to intervene? This information is crucial for both law enforcement and insurance firms to determine responsibility: is it on the driver or the manufacturer? Who gets the sentence, and who has to pay? These parties need a clear account of what exactly happened and the decisions taken, focusing more on context than technicalities.

The manufacturer/developer seeks to determine if the AI systems encountered any issues. The primary motive is quality assurance [61]. So a different, more technical explanation needs to be presented to the developers so that they can improve the system.

3.2.2 Customer Service Chatbot

In the case of chatbots for customer service, key stakeholders include clients, the company managing the operation, and the company responsible for development. The clients interact with the chatbot and might want reasons for its answers [45]. Especially when the assistant does not give an expected answer, the clients might want to know what variables of the prompt contributed to the answer, in order to adjust the prompt for better results. The operating company wants the chatbot to give correct answers based on the client's queries and if misinformation is given, why the chatbot provided that information. The developing company wants an explanation if the AI model fails on a higher level, like saying offensive things. This is also due to their desire for quality assurance [61].

3.2.3 AI Marketing

In designing marketing strategies utilizing AI assistants, various stakeholders are also engaged. In this case, the company of the AI, the company that uses the AI and their marketing professionals and artists are involved. Marketing professionals who use an AI system can be sceptical of the system if they are unclear about the motives and reasonableness of the system [63].

The marketing professionals and artists would like to know how their work is represented in the output of the AI [45]. Is the AI assistant trained based on the previous advertisements or not? Because if the work is not based on the company's own material, a copyright case can occur, and that is what they want to avoid.One could imagine that they may prefer a clearer depiction of how parts of the output connect to their contribution or input. In addition to the explanation, when the assistant incorporates artwork, the user also seeks clarification if the output deviates from expectations. This clarification should show how the wrong element in the image relates to the input prompt, to facilitate prompt adjustment for improved results.

Naturally, the manufacturer or developer is a stakeholder, as they seek insights into the assistant's behaviour to enhance system performance [61]. They expect the information on a more technical basis with very technical details.

3.2.4 Automated Writing Evaluation

When examining AI assistants for textual improvements, stakeholders are based on the usage context. Typically, the only stakeholder is the user of the AI assistant. One could imagine that they expect explanations for the mistakes they make or how they can improve the message that they want to convey. This can be spelling, grammar, sentence structure, or even the tone of the text.

In the context of writing a report and the student makes use of the AI assistant for correct referencing papers, they should watch out for fake references [38]. Furthermore, they seek straightforward descriptions of the reference's origin. It is also crucial to identify the sections from which the AI derives its paper summary [38]. Otherwise, the supervisor, another stakeholder in this context, has to make charges, giving the student an insufficient grade or even going to the exam board.

3.2.5 Routine Tasks

In the context of routine tasks managed by an AI assistant, booking meeting rooms or desks will be used as an example. Key stakeholders in this scenario include the employee, their team, the developer, facility and cleaning services, and building management.

Although literature on this subject is limited, it's reasonable to assume that workers and their colleagues likely prefer to sit close together and have meeting rooms near their desks. By employing the AI assistant, users seek clarity on the reasoning behind its selection of a particular location, particularly if the choice is unexpected. They are also interested in understanding which factors influenced the decision, to better tailor their prompts for improved outcomes.

Furthermore, the cleaning service, facility, and building management want insights into office space usage and predictive analysis of the facility maintenance. Developers naturally seek both decision explanations and feedback to enhance the system [61].

3.3 State of the Art of Routine Task AI Assistants

Section 3.1 illustrates the wide variety of AI assistants. This project concentrates on assistants designed for routine workplace tasks, in collaboration with Planon. To understand existing AI assistant software that aids in such tasks, various applications will be explored regarding their features. A significant number of these applications were discovered through internet searches and subsequently categorized by primary functionalities. Note that this section will exclude AI assistants that make use of generative AI for tasks like creating presentations, summaries, or coding.

3.3.1 Task Scheduling

One of the largest branches in AI assistants for routine tasks is the one that helps with task management and scheduling. One that most came up with the internet search is Motion [2]. This is software that will ”automatically plan your day based on your tasks and priorities.” But of course, there are competitors out there that will claim the same. Some of those competitors are: Trevor AI [3], Timehero [4], BeforeSunset AI [5], and ClickUp [6]. In Figure 3.1 it can be seen how Trevor AI makes suggestions for the remaining tasks that need to be scheduled by placing them in the schedule with a small icon.

They assist in maintaining clarity on task lists and embedding tasks into your timetable. Moreover, when meetings or other events are arranged, your schedule automatically adjusts according to priorities. Beyond prioritization, many of these AI assistants feature task duration prediction for enhanced scheduling efficiency. Additionally, the personalized AI model evolves and modifies to suit each person.

[2]Motion: https://www.usemotion.com/

[3]Trevor AI: https://www.trevorai.com/

[4]Timehero https://www.timehero.com/

[5]BeforeSunset AI: https://www.beforesunset.ai/

[6]ClickUp: https://clickup.com/

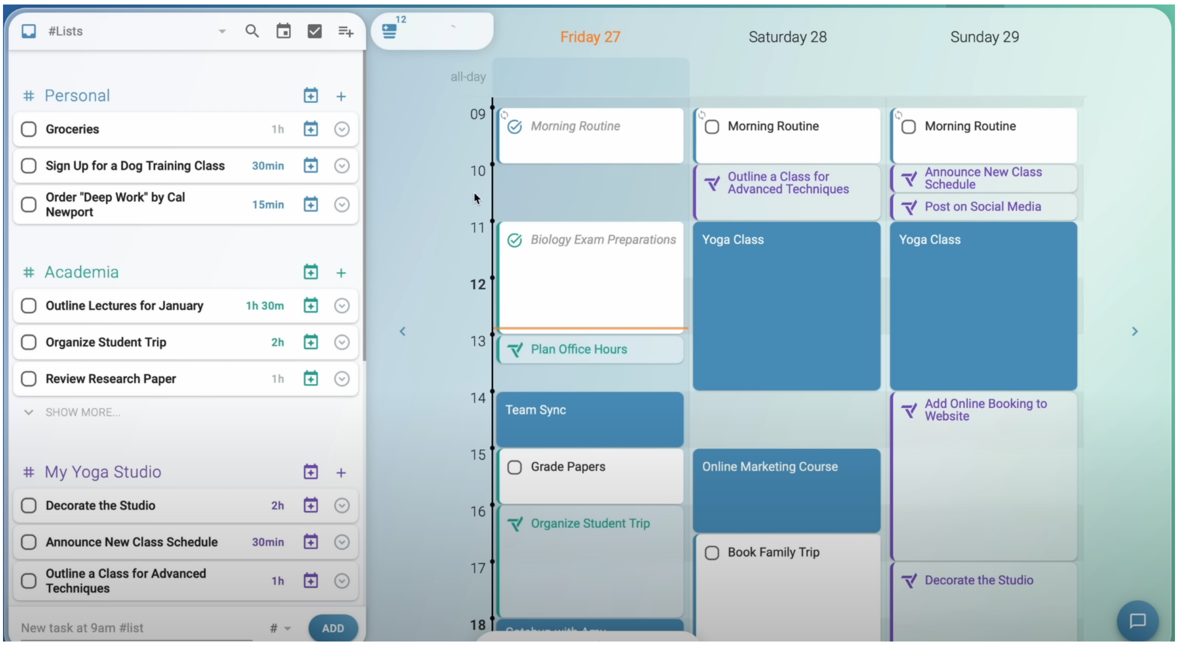


Figure 3.1: Screenshot of Trevor AI

3.3.2 Customer Service

Second, the AI assistants that are mostly set up as chatbots for customer service. One of those software applications is Agentforce [7]. Companies can configure 'agents' for multiple goals like order and delivery management, technical support, financial support, product information, and much more, as can be seen in Figure 3.2. "AI agents are developed through data collection, model training, natural language processing, and reinforcement learning. By utilizing these techniques, developers can create intelligent agents capable of understanding human language, making informed decisions, and taking actions to achieve specific goals."

Some other software applications that are used for customer service AI assistants are Cognigy [8], Intercom [9], Forethought [10], IBM watsonx Assistant [11], and many more. Cognigy's AI assistant goes beyond being just a chatbot; it functions as a phone agent that customers can reach out to for help. "Powered by Generative and Conversational AI, our Voice AI Agents can understand users and deliver the right answer - all in natural language."

[7] Agentforce: https://www.salesforce.com/uk/agentforce/
[8] Cognigy: https://www.cognigy.com/
[9] Intercom: https://www.intercom.com/
[10] Forethought: https://forethought.ai/
[11] IBM: https://www.ibm.com/products/watsonx-assistant

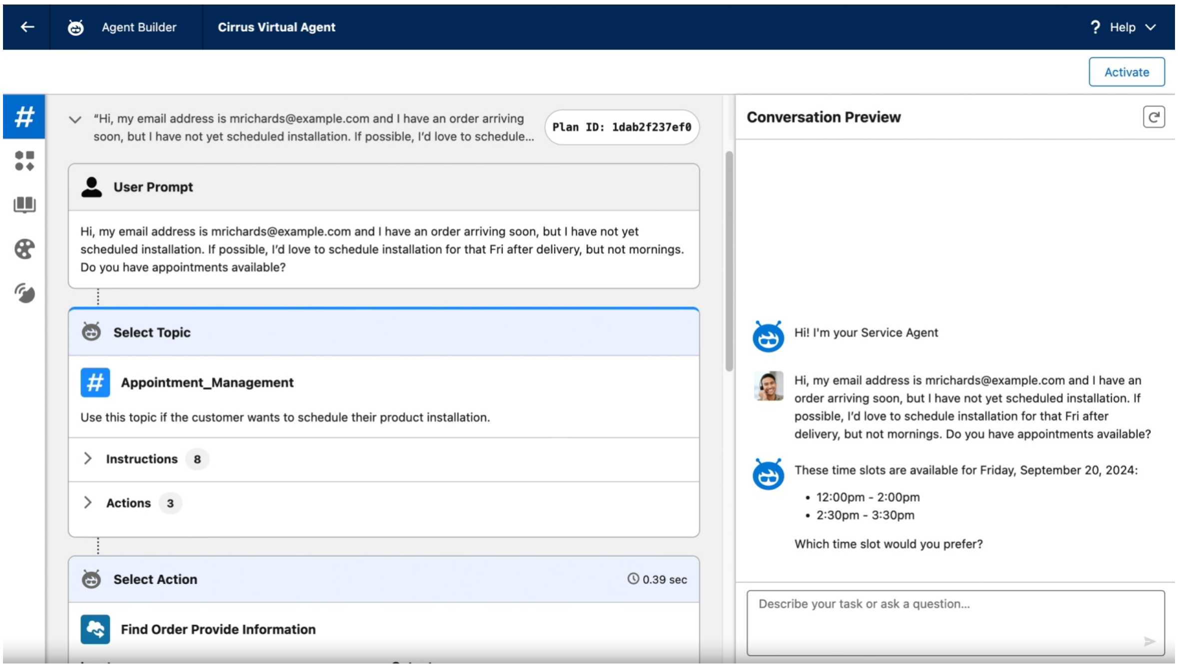


Figure 3.2: Screenshot of Agentforce

3.3.3 Inventory Management

The following section discusses the role of AI assistants in inventory management. These applications help minimize excess inventory, optimize safety stock, and reduce shipping costs. Most of them are also able to have predictive analysis on demand patterns and trends [72].

Some examples of these software applications are Inventory AI by Peak AI [12], Digitshelves [13], SmartStock [14], and C3 AI Inventory Optimization [15].

These applications mostly work in the background, and the users do not interact a lot with them. For example, Inventory AI by Peak AI has the functionalities: dynamic inventory, production planning, rebuy, reorder, and replenishment. All to meet demand and have the optimal stock cover.

3.3.4 Office Space Management

The last selection is for AI assistants that help with office space management, such as booking meeting and desk spaces. These intelligent tools check room availability and factor in personal preferences, using predictive analytics to optimize usage.

[12]Inventory AI: `https://peak.ai/products/inventory-ai/`
[13]Digitshelves: `https://www.digit7.ai/digitshelves/`
[14]SmartStock: `https://smartstock.nl/`
[15]C3 AI: `https://c3.ai/products/c3-ai-inventory-optimization/`

Naturally, these AI assistants will have to make assumptions if detailed prompts are not provided. These assistants are mostly active on a company basis, where workers mostly work hybrid. Hybrid or remote working became more popular after the COVID-19 pandemic [15]. Few assistants are available. An example of these types of assistants is Room Manager AI [16], as can be seen in Figure 3.3 Room Manager AI makes use of a conversational AI. These assistants can "reserve desks, schedule conference rooms, manage visitor invitations, and organize team activities."

Another AI assistant that can help in workplace management is Yarvis from Yarooms [17]. "It can quickly handle multiple reservations, schedule recurrent bookings, locate spaces with specific amenities, find colleagues in the office, and more."

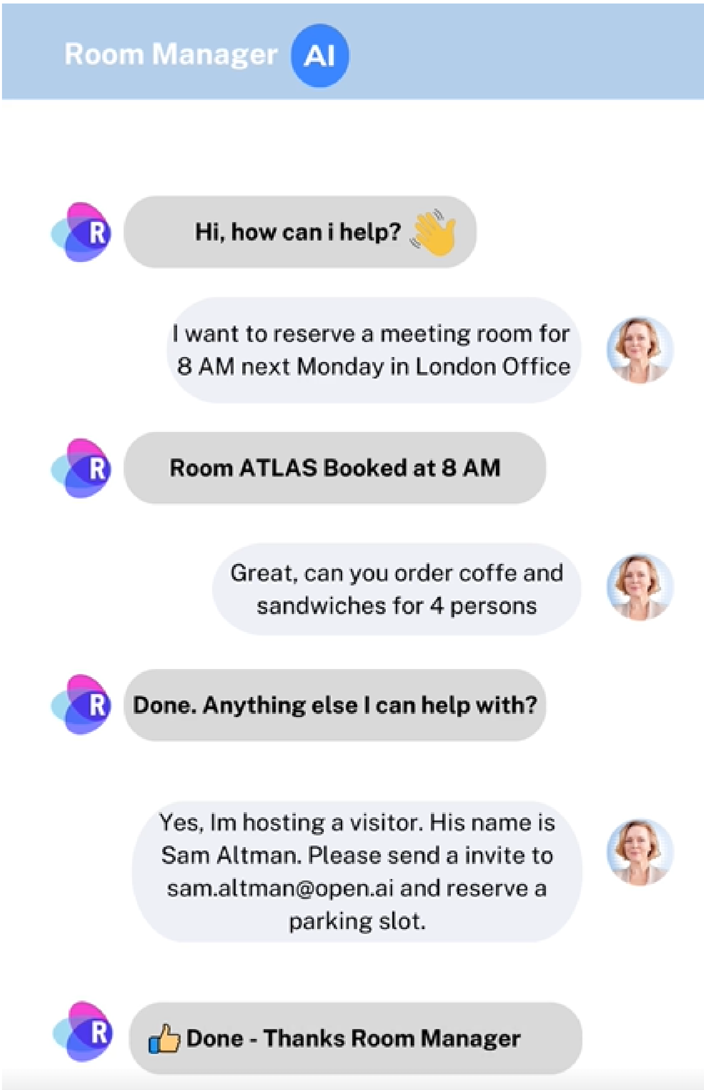


Figure 3.3: Screenshot of Room Manager AI

[16]Room Manager AI: `https://roommanager.com/room-manager-ai/`
[17]Yarooms: `https://www.yarooms.com/product/yarvis-ai-workplace-assistant`

Chapter 4

Explainable AI

This section explores why AI assistants need explainability to enhance user comprehension. Subsequently, it examines various technical and human-centered approaches to achieve this explainability.

4.1 Problem

As mentioned before, AI is becoming more and more involved in our daily lives. AI has already become ubiquitous, and people have become accustomed to AI making decisions for us in our daily lives, from product and movie recommendations on Netflix and Amazon to friend suggestions on Facebook and tailored advertisements on Google search result pages. However, in life-changing decisions such as disease diagnosis, it is important to know the reasons behind such a critical decision [2]. In addition to that, with the automation of routine decisions coupled with more intricate and complex information architecture operating this automation, concerns are increasing about the trustworthiness of these systems [76]. Systems whose decisions cannot be well interpreted are difficult to trust [42]. And for each type of AI assistant, mentioned in Section 3.1 the capability to trust it can differ.

As the core technology of AI, Machine Learning (ML) and its learning process are opaque, and the output of AI-based decisions is not intuitive. For many non-technical users, a ML-based intelligent system is a black box. [83] This black box phenomenon is mentioned by several researchers [2, 43, 76, 83, 42]. They further debate how this phenomenon influences trust in AI systems. If users do not understand the decision-making of the AI assistant, they will not trust the AI assistant, which will lead to not even using the assistant [76]. Increased transparency in algorithmic decision-making enables practitioners to avoid failure modes and build safer models [23].

To solve the black box problem, the field of XAI arose. Because explainability has been described as being a powerful tool for a variety of purposes including "detecting flaws in the model and biases in the data, for verifying predictions, for improving models, and finally for gaining new insights into the problem at hand" [5]. In addition to that, users prefer systems that provide decisions with explanations over systems that provide only decisions. Tasks where explanations provide the most value are those where a user needs to understand the inner workings of how an AI system makes decisions [25].

Moreover, as discussed in Section 3.2, various categories of AI (assistants) need distinct forms of explanation. Additionally, each category's stakeholders require tailored explanations specific to their needs.

Initially considered a niche research area, XAI has evolved into a prominent field due to the growing application of AI [43].

4.1.1 Definition of Explainable AI

To address the black box challenge with XAI, defining XAI is crucial. Adadi [2] points out the lack of a universally accepted definition for explainable AI. Often, XAI refers more to efforts and initiatives aimed at improving AI transparency and trust rather than a strict technical term. Some also argue XAI's aim as making ML-based systems comprehensible to humans [18].

Alongside explaining the technical facets of the models, it is crucial to understand how to effectively communicate these explanations to users. The type of information you provide varies for different stakeholders. As noted in 3.2, varying stakeholders require unique insights into the reasons behind an AI assistant's decisions, reflecting a form of HCXAI [79].

4.2 Typology of Explanations

Understanding the elements of effective explanations is essential for creating high-quality explanations for users. Discussions on explanations have pointed out the limitations of current approaches, the lack of focus upon user needs, the difference between justification (*Why?*) and transparency (*How?*) [74]. Additionally, overly simplistic explanations are ineffective because they lack sufficient detail to inspire confidence or trust in the audience. Overly detailed explanations can also fall short, as they mismatch the recipient's comprehension level [74].

Cabitza et al. [7] together with Miller [46] approach explanations from a more psychological angle to distinguish different types of explanations. They mention that there are explanations that (i) motivate the user to understand, (ii) make users confident they have understood why the model gave that advice, and (iii) convince users to change their mind or anyway confirm the advice of the machine since it was found to be accurate, lawful, and fair.

Tsakalakis et al. [74] developed an explanation typology to facilitate a categorisation method capable of generating a detailed set of explanation requirements, which subsequently aids in identifying the essential components for the computation phase. As discussed in Section 3.2, explanation can be necessary in numerous contexts, depending on the scenario and audience, with diverse explanation requirements. Tsakalakis et al. [74] divide those into nine dimensions:

Source: An explanation requirement may stem from three types of sources: (i) applicable laws; (ii) related authoritative guidance and standards; and (iii) internal compliance or business needs.

Timing: An explanation can be generated either ex ante, i.e. before the object(s) explained within the explanation take(s) place, or ex post, i.e. after the object(s) materialize(s).

Autonomy: The Autonomy dimension distinguishes between explanations that are generated without any input from the recipient (*Proactive*) or explanations that are generated only as a response to specific input from the recipient (*Responsive*).

Trigger: The Trigger dimension expresses the event that triggers the generation of an explanation.

Content: This dimension aims to capture details to be included within the formulation of an explanation. This includes three categories: Sensitivity, Confidentiality, and Minimum content.

Scope: This dimension assesses whether the explanation only applies to a particular case (*Local*) or whether it can be reused in other contexts (*Universal*).

Explainability goal: The Explainability Goal dimension describes the functional objectives that generating a specific explanation aims to fulfil. Explanations can have multiple purposes, contrary to common assumptions. The significance of an explanation, contingent on its language and structure, is closely tied to the particular objective it seeks to accomplish. The different goals can be grouped into two categories: Understandability and Intervenability, see Figures 4.1 & 4.2.

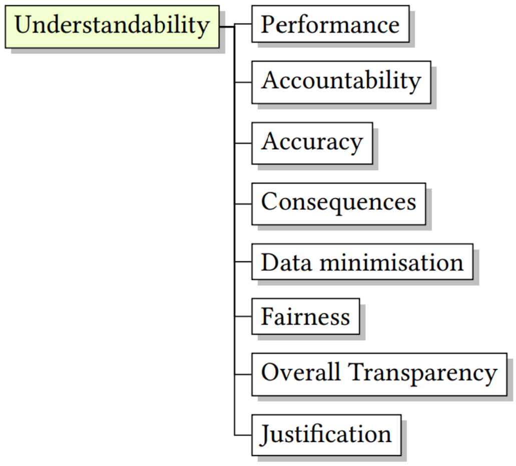


Figure 4.1: The Understandability Goals [74]

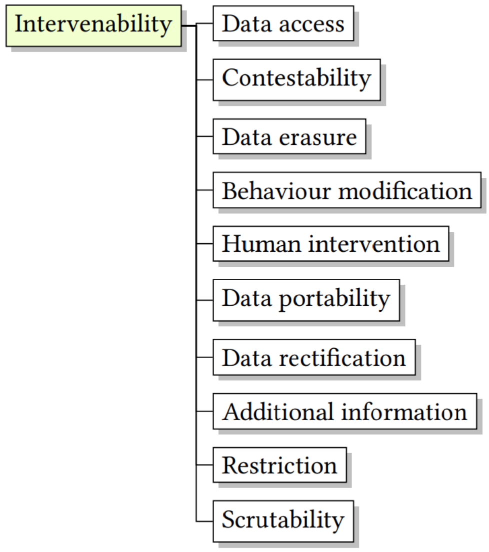


Figure 4.2: The Intervenability Goals [74]

Intended Recipient: This dimension defines the categories of recipients that will be associated with the explanations.

Criticality: This dimension differentiates between mandatory and suggested explanation requirements. An explanation is mandatory when a relevant legal regulation or governance framework demands it as an obligation.

Delving further into the 'explainability goals' within the context of routine tasks reveals that, alongside the previously noted elements of justification and transparency, it is crucial to consider end user control. This is linked to the intervenability objectives highlighted in Figure 4.2, such as *Behaviour modification* and *Human intervention*. By taking these objectives into account during the creation of an explanation, it could empower the end user to adjust their prompt to achieve the desired outcome. This also empowers the statements from Cabitza et al. [7] and Miller [46] from the psychological approach.

Tsakalakis et al. [74] tested the feasibility of its construction for the purpose of computing explanations in the context of two small-scale pilot studies. This should, of course, be extended to multiple domains to see if this typology is domain-agnostic or not. In addition to that, if people want to use the typology, they should map applicable explanation requirements to the nine dimensions of the typology. An essential factor is the necessity of thorough input to produce valuable outcomes. This approach enables people to grasp the decisions, facilitating either their acceptance or the modification of the input as needed.

4.3 Methods of Explainable AI

Given the confirmed necessity of explainability in AI assistants and the criteria for crafting quality explanations, determining the optimal implementation method to enhance understanding is crucial. Numerous approaches have emerged over the past decade.

This section will research the different methods of explainability. Since there are a lot of different methods with many variations, methods will be grouped together on the crucial aspects of interpretability methods. Linardatos [42] has created a taxonomy of methods for explainability, see Figure 4.3. For this thesis, the focus will be on **Local** vs. **Global** and **Model-Specific** vs. **Model-Agnostic**.

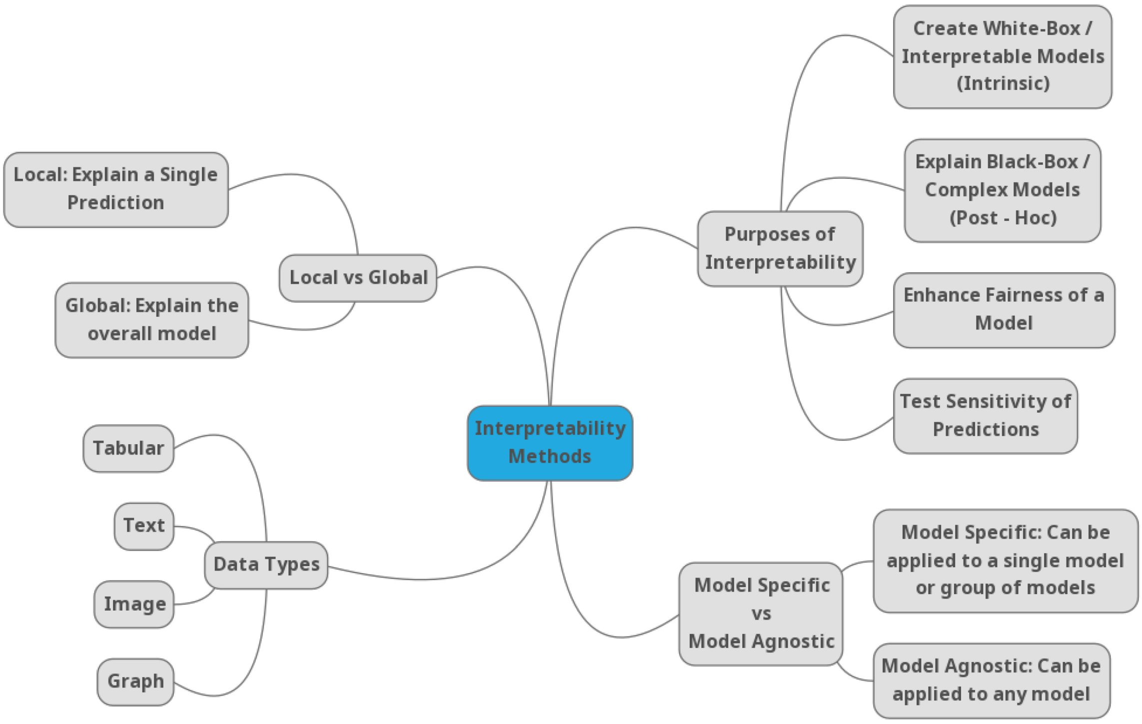


Figure 4.3: Taxonomy Mind-Map of Machine Learning Interpretability Techniques [42]

Local Explanations: The aim of XAI at the local level is to provide insights into why a particular decision was made for a specific input [42, 71, 69]. Local interpretation is important for decision-makers to trust the output or correct the wrong output [13].

Global Explanations: The aim of XAI at a global level is to explain the behaviour of the model across the entire dataset. It gives insights into the main factors influencing the model, and the overall trends and patterns observed [42, 71, 69]. Global interpretation helps decision-makers gain a macro-level understanding of the ML model, including the most influential input features [13].

Model-Agnostic: Methods that can be applied to any ML model, regardless of its internal structure, architecture, or specific learning algorithm [77]. Most post-hoc interpretable ML techniques are model-agnostic [13].

Model-Specific: Methods that are designed to work specifically with the internal structure and architecture of a particular type of model [77].

	Local		**Global**		**Both**
Model-Agnostic	LIME (Local Interpretable Model Agnostic Explanations)	Anchors	Partial Dependence Plot (PDP)	Permutation importance (PIMP)	Shapley Additive explanations (SHAP)
	GraphLIME	Asymmetric Shapley Values (ASV)	Shapley Flow	T-Distributed Stochastic Neighbor Embeddings (T-SNE)	Explain like I'm 5 (ELI5)
	Prediction Difference Analysis (PDA)	Break-Down			
	L2X	Protodash			
	SmoothGrad	ProfWeight			
	LIVE	DLIME			
	RISE	Occlusion sensitivity			
	Counterfactual explanations				
Model-Specific	Layer-wise Relevance Propagation (LRP)	Explainable Graph Neural Networks (XGNN)	Testing with Concept Activation Vectors (TCAV)	Supersparse Linear Integer Models (SLIM)	
	Deep Taylor Decomposition (DTD)	Integrated Gradients	Concept Activation Vectors (CAVs)	Gradient Integrated Relevance Propagation (GIRP)	
	Gradient-weighted Class Activation Mapping (Grad-CAM)	DeepLIFT			
	Contextual Explanation Method (CEM)				

Table 4.1: Comparison of Different Interpretability Methods

In Table 4.1 a selection of XAI methods is mentioned [42, 68, 29]. Since XAI is still very new, new variations and methods are still being invented. The methods that are mostly mentioned in the literature are Local Interpretable Model-agnostic Explanations (LIME), Layer-wise Relevance Propagation (LRP), and SHapley Additive exPlanations (SHAP). These methods will be further explained in the sections below.

4.3.1 LRP

LRP is a propagation-based explanation method; it requires access to the model's internals [29]. LRP does not explain the prediction of a deep neural network in one step, as model-agnostic methods would do, but exploits the network structure and redistributes the explanatory factors layer by layer. With back-propagating from the model's output onto the input variables [29]. Propagation rules can be implemented efficiently and modularly in most modern neural network software [67]. Next to that, LRP has high computational efficiency, and with the possibility of changing the parameters of the LRP rules, it can provide high explanation quality, which makes this a method that is applicable in many different practical scenarios [29, 67].

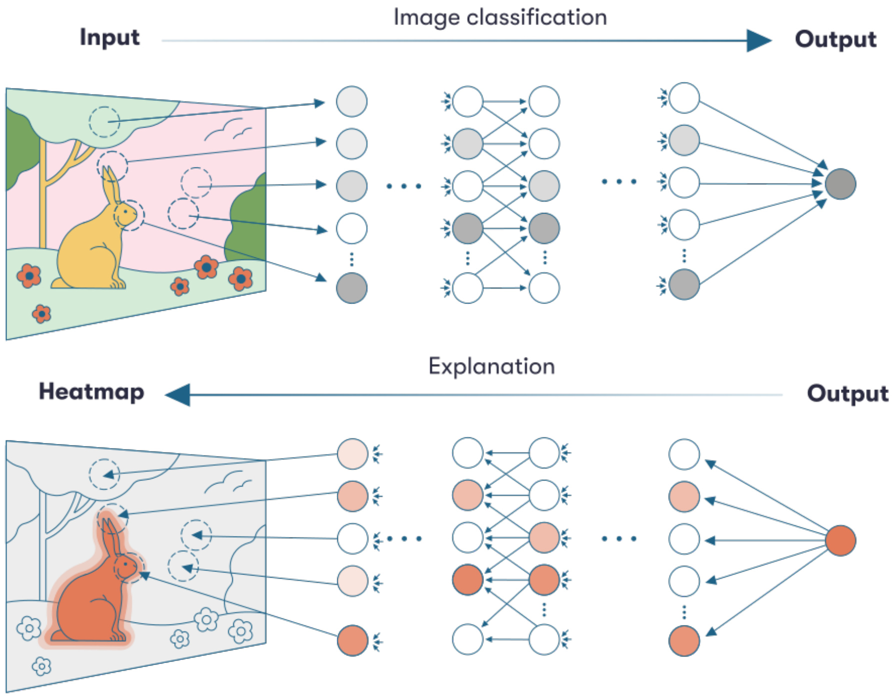


Figure 4.4: Visual Representation of LRP [47]

4.3.2 LIME

LIME is an algorithm that explains the predictions of any classifier or regressor in a faithful way [64], by fitting a local surrogate model, whose predictions are easy to explain [29]. The LIME method samples in the neighbourhood of the input of interest, evaluates the neural network at these points, and tries to fit the surrogate function such that it approximates the function of interest [67]. The method begins by selecting a range of various possible inputs close to the specific input of interest. This means it looks at small variations or perturbations of the input to see how they affect the output. For each of these sampled points, the method evaluates the neural network. This allows the algorithm to observe how slight changes in input affect the network's predictions. With the data from the neighbouring evaluations, the algorithm then creates a simpler model, called a surrogate model, that aims to approximate the behaviour of the neural network around the input point in question. This surrogate function is easier to interpret compared to the complex neural network.

LIME is model-agnostic; it can be applied to any classifier, even without knowing its internals, like its architecture or weights of a neural network classifier [67, 29, 64].

LIME has been applied successfully in many instances; however, it also has some limitations. A key drawback is the high computational resources required for calculations [29, 67].

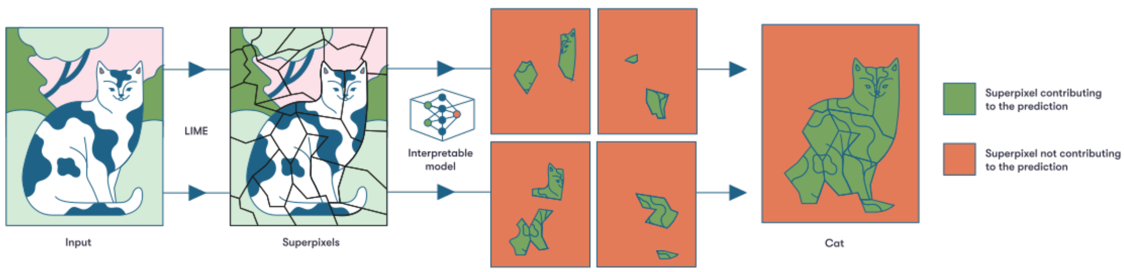


Figure 4.5: Visual Representation of LIME [47]

4.3.3 SHAP

SHAP is a post-hoc model-agnostic method that can be applied to any ML model [66]. And as the name already suggests, the SHAP method is based on the Shapley values. Those values originate from game theory and it is a method where a reward is assigned to game players based on their total gain contribution [54].

SHAP faces a major challenge in terms of computational intensity. For contemporary models, including deep neural networks with high-dimensional data, calculating Shapley values precisely is impractical [29].

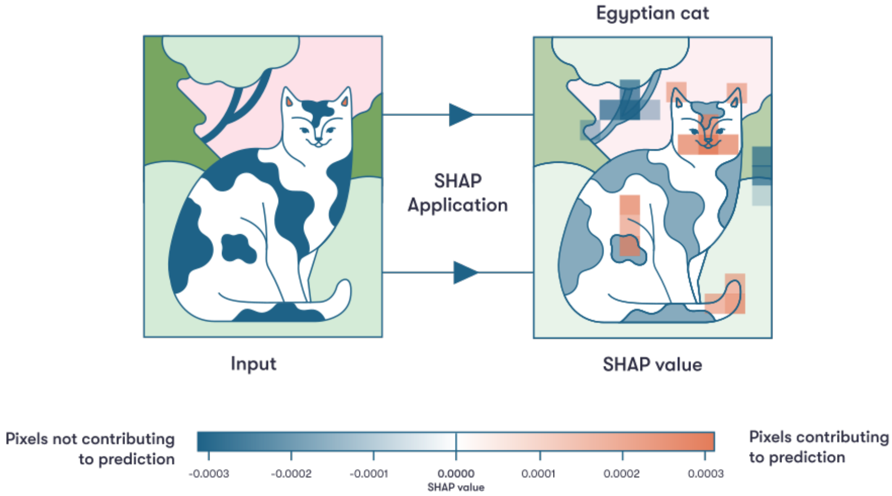


Figure 4.6: Visual Representation of SHAP [47]

4.4 Human-Centred Explainable AI

Having delved into the technical aspects of XAI, it's crucial to deliver clear explanations to users. Research [51] revealed that more than 30% of users struggled to comprehend XAI explanations enough to utilize them in basic tasks. Recently, the Human-Computer Interaction (HCI) field has intensified efforts to assess human requirements for AI explanations. These insights feed into XUI, also known as Explainable Interfaces (EI). It provides both naive and expert users with different user experiences depending on their skill level, with the focus on UX design aspect of XAI (e.g., structure, design, format, design, and content of the explanations) [55, 4]. The primary concern of XAI algorithms is to generate *what to explain*, XUI research is about *how to explain* in a way that is effective for specific user groups [55]. Most of the conducted research concentrates on the technical aspects of XAI.

4.4.1 Design Requirements

The research by Jung et al. [34] delves into recommendations for designing human-centred explanations, or XUI. Their analysis spans across general systems and the medical field; however, this thesis will only address findings related to general systems. Due to the number of articles reviewed by Jung et al. [34], summarising all the evidence is unfeasible. Over 50% of the summarised recommendations are derived from just one or two articles, possibly suggesting that the field is both diverse and still developing. Additionally, the number of articles represents the level of interest in the topic rather than the strength of evidence supporting a recommendation. Nevertheless, they suggested basic principles to keep in mind when designing an XUI. This can be seen in the mind map in Figure 4.7; the unfolded mind map can be found in Appendix A.

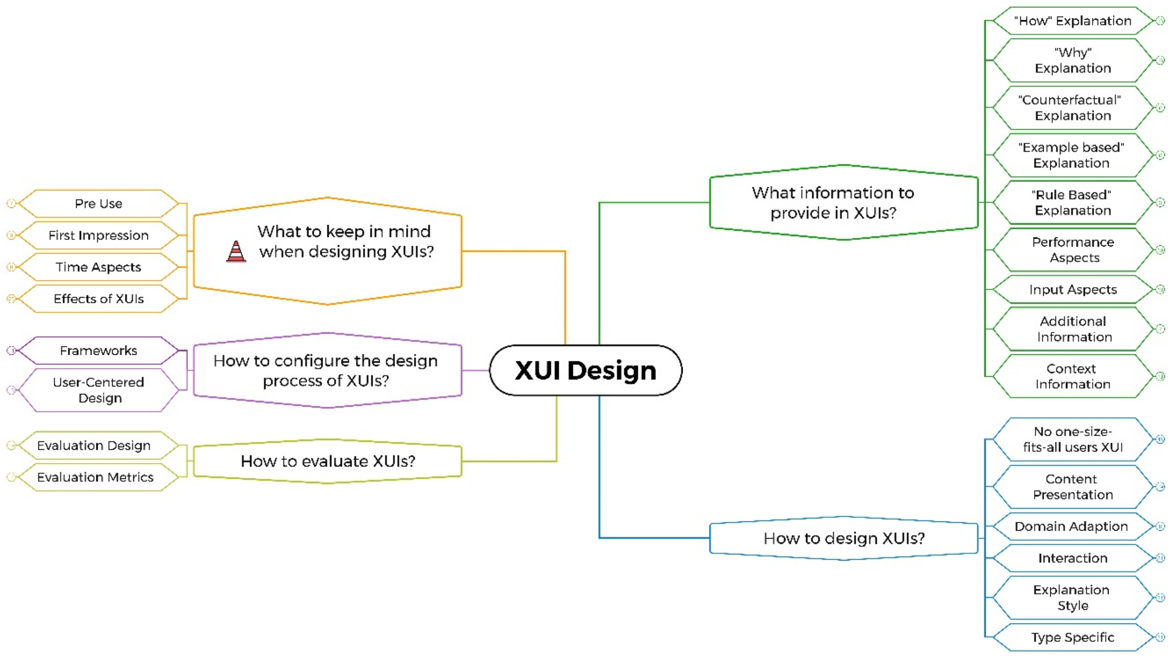


Figure 4.7: Mind-Map of Orientation Nodes [34]

The various elements highlighted in the mind map are extensively discussed in this thesis. Nonetheless, a few minor yet crucial aspects in XUI design could be missed. For instance, users' initial feelings about the XUI during use, and its impact on use duration, are notable. Also, explanations might not adequately mitigate the influence of initial impressions of system performance on user perception [34].

Nguyen et al. [55] extended their research beyond this mind map to examine design requirements, classifying their findings into visual hierarchy and necessary features as illustrated in Figure 4.8. Visual hierarchy pertains to "arranging design elements in a manner that directs the viewer's attention according to their priority." Required features are functionalities essential for the final XUI, significantly influencing its design space.

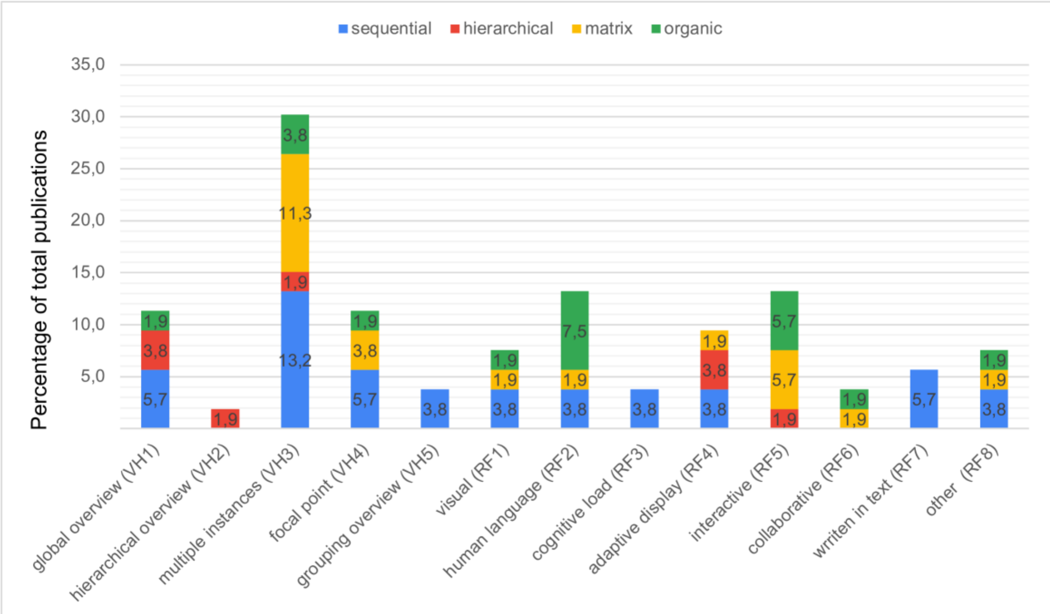


Figure 4.8: Design Requirements Including Visual Hierarchy and Required Features [55]

In conclusion, to create an XUI tailored to a specific domain, one should utilize the core features and principles outlined by Jung et al. [34] and Nguyen et al. [55] to determine the specific design needs of a domain-specific AI integrated system such as AI assistants.

4.4.2 XAI-Interaction Types

Next to the design requirements, users have various interaction methods with the XUI. Interaction can occur in various ways and can have different goals [30]. These types of interaction are:

Interaction as (Information) Transmission (1): The goal of this interaction centres around presenting users with one complete explanation. In this way, transparency of the AI's decisions occurs [14, 30].

Interaction as Dialogue (2): The goal of this interaction is to facilitate natural and interactive conversation with the AI system. This is done in order to create transparency and comprehensibility [14, 30].

Interaction as Control (3): The goal of this interaction is to support rapid convergence towards desired AI behaviour to enhance the effectiveness [14, 30].

Interaction as Experience (4): The goal of this interaction is to emphasize managing the expectations and preferences of users about the AI. It centres around the explanatory goals of trust, satisfaction, and persuasiveness [14, 30].

Interaction as Optimal Behaviour (5): The goal of this interaction is to adjust human behaviour despite limitations of fully understanding the AI behaviour. This interaction is also to enhance the effectiveness [14, 30].

Interaction as Tool Use (6): The goal of this interaction is to facilitate learning from AI behaviour about a given domain to also enhance the effectiveness. [14, 30].

Interaction as Embodied Action (7): The goal of this interaction is to establish a joint understanding with the AI for effective collaboration in a given domain with the explanatory goal of effectiveness [14, 30].

For different end goals, different interaction types should be considered to have the best option. According to Williams [79], interactive explanations are a promising approach for communicating explanations to users of AI systems.

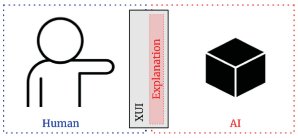


Figure 4.9: (1)

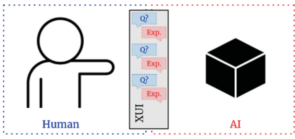


Figure 4.10: (2)

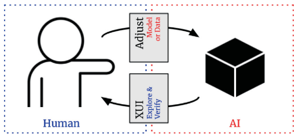


Figure 4.11: (3)

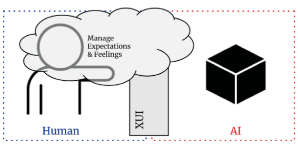


Figure 4.12: (4)

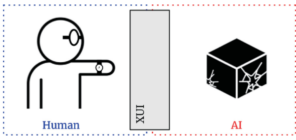


Figure 4.13: (5)

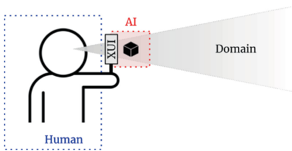


Figure 4.14: (6)

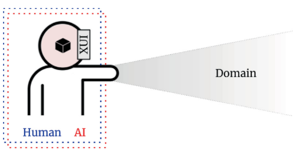


Figure 4.15: (7)

Figure 4.16: XAI-Interaction Types [14]

4.4.3 End User Control

As can be seen in Figure 4.7 and in the study of Chromik and Butz [14], the interaction types of control and tool use are also brought to light. This, with the intervenability goals in Figure 4.2 also indicates end user control, resulting in a good extra topic next to the usual questions of *how?* and *why?*. Particularly in the domain of routine tasks in the workplace, the end user should always have control, also when they know if the AI assistant has it wrong.

4.5 Evaluating Explainable AI

After creating the XAI / XUI for an AI system, it is important to evaluate the explanation. If it is really helpful for the stakeholder in question. There are various ways in which an XAI / XUI can be evaluated. Those evaluation methods can be split into two main ways of evaluating: **Objective** and **Subjective**.

Objective (that usually doesn't involve users) evaluations: includes objective metrics and automated approaches to evaluate methods for explainability [75].

Subjective (that usually involves users) evaluations: contains those evaluating methods for explainability with a human-in-the-loop approach by involving end-users and exploiting their feedback or informed judgment [75].

Considering Figure 4.7, in the mind map is a branch on "How to evaluate XUIs?" showing considerations for evaluating design and evaluation metrics to select for evaluation.

An approach is proposed, by Nauta et al. [52], since different aspects regarding explanation quality can be evaluated, it can therefore be argued that explainability is a non-binary characteristic that can be measured by evaluating to what degree certain properties are satisfied. That is why they composed twelve properties of an explanation, see Table 4.2, that can be evaluated in order to see how good the XAI is.

Nauta et al. [52] analysed over 300 papers that introduce a method of explaining a ML model and how they were evaluated. The following results came to light: **33%** only evaluated with anecdotal evidence, **58%** applied quantitative evaluation, and **22%** were assessed using human subjects in a user study; of those evaluations with human subjects, almost a quarter was with domain experts. These methods are linked to the different Co-12 properties, see the third column of Table 4.2. The majority of XAI evaluation focused on evaluating Coherence, Completeness, Compactness, or Correctness.

Co-12 Property	Description	Methods
Content	**Correctness** Describes how faithful the explanation is w.r.t. the black box. *Key idea:* Nothing but the truth	**Model Parameter Randomization Check:** Alter the components of the predictive model at random and ensure the explanation varies.
		Explanation Randomization Check: Introduce random fluctuations into the explanation embedded within the predictive model and observe whether the predictive model's output alters.
		White Box Check: Utilize the explanation technique on an interpretable white-box model to verify how well the explanation aligns with the white-box's reasoning process.
		Controlled Synthetic Data Check: Construct a synthetic dataset designed so that the predictive model adheres to a predetermined reasoning (important: verify this assumption by reporting near-perfect accuracy, for instance). Evaluate whether the explanation aligns with the reasoning used in the data generation process.
		Single Deletion: Alter, mask, or remove one feature in the input data and assess the impact on the output of the prediction model. Evaluate the correlation with the importance score of the explanation.
		Incremental Deletion: Gradually modify the input by sequentially deleting, altering, or incorporating features as dictated by the explanation's order. For every modified input, assess the resultant variation in the predictive model's output. Document the mean alteration in the log-odds score, AUC, curve steepness, or the count of features required to alter a decision. Conduct comparisons with random ranking or alternative baseline metrics.
	Completeness Describes how much of the black box behaviour is described in the explanation. *Key idea:* The whole truth	**Preservation Check:** Providing the explanation (or data derived from it) to the predictive model is expected to yield a decision consistent with that of the original, complete input sample.
		Deletion Check: Providing input without explanation's relevant features should result in a different decision by the predictive model than the decision for the original.
		Fidelity: Evaluate the consistency between the predictive model's output and the corresponding explanation when using the same input samples.
		Predictive Performance: Predictive performance of the interpretable model or predictive explanation with respect to the ground-truth data.
	Consistency Describes how deterministic and implementation-invariant the explanation method is. *Key idea:* Identical inputs should have identical explanations	**Implementation Invariance:** Assess if the explanation method remains consistent across different implementations of the predictive model by checking if two implementations that pro- duce identical outputs for a given input also yield the same explanation.
	Continuity Describes how continuous and generalizable the explanation function is. *Key idea:* Similar inputs should have similar explanations	**Stability for Slight Variations:** Assess the resemblance between explanations for two slightly different samples. Small variations in the input, for which the model response is nearly identical, should not lead to significant alterations in the explanation.
		Fidelity for Slight Variations: Evaluate the concordance between interpretable predictions for original and slightly different samples: an explanation for original input x should accurately predict the model's output for a slightly different sample x.
		Connectedness: Assess the degree to which a counterfactual explanation is linked to the examples in the training dataset: the ideal scenario is that the counterfactual is not an outlier and there exists a continuous trajectory connecting a generated counterfactual to a training example.
	Contrastivity Describes how discriminative the explanation is w.r.t. other events or targets. *Key idea:* Answers "why not?" or "what if?" questions	**Target Sensitivity:** The explanation for a specific target or model output should differ from that for other targets.
		Target Discriminativeness: The explanation needs to be target-discriminative, enabling another model to accurately infer the correct target from the explanation, whether via supervised or unsupervised learning methods.
		Data Randomization Check: Alter the labels randomly in a copy of the training dataset, train a model using this altered dataset, and verify that the explanations for this model on a test dataset differ from those of a model trained using the original dataset.
	Covariate complexity Describes how complex the (interactions of) features in the explanation are. *Key idea:* Human-understandable concepts in the explanation	**Covariate Homogeneity:** Assess the reliability with which a covariate in an explanation represents a specified concept that can be interpreted by humans.
		Covariate Regularity: To assess the regularity of an explanation, calculate its Shannon entropy, which quantifies the level of noise in the explanation and indicates the ease with which the explanation can be memorized.
Presentation	**Compactness** Describes the size of the explanation. *Key idea:* Less is more	**Size:** The total size (absolute) or sparsity (relative) of the explanation.
		Redundancy: Calculate the redundancy or overlap between different parts of the explanation.
		Counterfactual Compactness: Given a counterfactual explanation showing what needs to be changed in the input to change the predicted output of the predictive model, measure how much needs to be changed.
	Composition Describes the presentation format and organization of the explanation. *Key idea:* How something is explained	**Perceptual Realism:** Measure how realistic a generated explanation is compared to real, original samples.
	Confidence Describes the presence and accuracy of probability information in the explanation. *Key idea:* Confidence measure of the explanation or model output	**Confidence Accuracy:** Evaluate the precision of confidence or uncertainty estimates if they are included in the explanation.

User	**Context** Describes how relevant the explanation is to the user and their needs. *Key idea:* How much does the explanation matter in practice	**Pragmatism:** The expense or level of difficulty a user encounters when implementing recommendations from a counterfactual explanation, which details the changes needed for the user to achieve a specific outcome as predicted by the model.
		Simulated User Study: Produce an artificial dataset to enable the automatic assessment of explanation utility for tasks pertinent to users.
	Coherence Describes how accurate the explanation is with prior knowledge and beliefs. *Key idea:* Plausibility or reasonableness to users	**Alignment with Domain Knowledge:** Compare the generated explanation with a “ground- truth” expected explanation based on domain knowledge.
		XAI Methods Agreement: Quantitatively compare explanations from different XAI methods and evaluate their agreement.
	Controllability Describes how interactive or controllable an explanation is for a user. *Key idea:* Can the user influence the explanation?	**Human Feedback Impact:** Measure the improvement of explanation quality after human feedback, where the user is seen as a system component.

Table 4.2: Description of Co-12 Properties [52]

There are numerous approaches to assess an XAI model broadly. Most of these methods are applicable without involving users. Mohseni et al. [48] highlights user-centric evaluation techniques, such as *interviews* or *Likert-scale questionnaires*, to measure users’ comprehension of the model and their satisfaction. Some of the standardised questionaries are the Explanation Satisfaction Scale (ESS) and the Explanation Goodness Checklist (EGC) [28].

Chapter 5

The Company Planon

This chapter provides an overview of Planon, as the thesis is conducted in partnership with them. Understanding Planon's potential contributions is crucial. The chapter includes an introduction to the company and its software, case studies on firms using Planon, and discusses Planon's future and how this chapter ties into the background research.

5.1 Introduction

Planon is a Dutch software company founded in 1982. They are a provider of smart and sustainable building management software solutions. Planon is an international software vendor that produces Facility Management and Real Estate management software for organisations; examples of those software solutions can be read in Section 5.2. Next to having an office in the Netherlands, it has offices all over the world. Some of those locations include Germany, Canada, Belgium, Australia, France, Hong Kong, India, the United States of America, Sweden, and a few more. Planon has over 1000+ employees across those locations and more than 3 million users across the world. Planon is recognized as a 'leader' for integrated real estate and facility management software in the Verdantix Green Quadrant®: Connected Portfolio Intelligence Platforms (CPIP) and Integrated Workplace Management System (IWMS) 2025 [1].

5.2 Software

Planon has developed a range of diverse software applications over time. They claim that: "Our advanced capabilities help optimise workplace performance across all industries. They simplify business processes and reduce costs throughout the real estate life cycle, which provides benefits for professionals in many roles." An examination of Planon's software applications is conducted to identify the most beneficial one for this project.

5.2.1 IWMS & Campus Management Solution

Firstly, there are the IWMS and Campus management systems; these software platforms resemble one another, each comprising various customizable modules tailored to campus or workplace requirements. They encompass Real Estate Management, Space & Workplace Services Management, Asset & Maintenance Management, and Energy & Sustainability Management.

The Real Estate Management part claims to provide "strategic information and streamlined processes for optimising any real estate portfolio. And enabling optimal support throughout the real estate lifecycle facilitates the primary processes and reduction of accommodation and operational costs." They also have a cloud-based integration solution for SAP S/4HANA® of the Real Estate Management part.

Space & Workplace management is a tool to plan, manage and operate different accommodation and workplace concepts. This can "help optimise space utilisation, improve the occupancy of meeting spaces and workplaces, and increase the service level for employees."

The Asset & Maintenance Management part of the software claims to ensure "optimal asset performance for buildings, installations, and equipment, whilst enabling cost-efficient maintenance."

And finally, the Energy & Sustainability Management measures and monitors the sustainability profile of buildings and processes in a continuous cycle and enables improvements.

5.2.2 Facility Services Business Solution

The Facility Services Business Solution offers tools and processes in four key areas that claim to make service offering, planning, execution, monitoring, and billing more efficient, scalable, and transparent with extensive process automation and seamless integration. The Facility Services Business Solution consists of four modules: Customer Management for managing all the communication with the customers. The Operations Command Centre claims to reduce project on-boarding efforts. The Mobile Field Services can help with the use of mobile devices and tablets to improve the productivity of field engineers, increase the speed of processing, and ensure compliance. The last one, Revenue Optimisation, can help with increasing the speed and completeness of billing.

5.2.3 Lease Accounting Solution

The Lease Accounting Solution ensures timely reporting readiness and eliminates the risk of non-compliance regarding the new accounting standards, ASC 842 and IFRS 16, which became effective for public companies in the fiscal year 2019.

5.3 Case Studies

To further investigate the application of the software in the real world, three of the released case studies of companies that use the software of Planon are analysed. All these case studies focus on the IWMS software of Planon, since this includes the Space & Workplace management module.

This first case study is from the company **BayWa**[1]. The BayWa Group is active worldwide in the energy, agriculture, and construction sectors. Next to the 20-story headquarters, the company has 10 other locations that are predominantly used as offices. BayWa is using the Planon Workplace App to provide professional workplace management in its office buildings. BayWa has already used Planon's IWMS for over a decade, and after the COVID-19 pandemic, it needed a tool for hybrid working and flexible workplace management.

The main focus of implementing was the workspace booking system; there are multiple ways of booking a workspace: via the workspace booking site looking at an overview, with a 'simple' QR code taking them to that site, a 'smart' QR code booking the desk that they are standing at, or via the app that gives full details of each workspace. Compared to the Computer-Aided Facility Management (CAFM) front end, the Planon Workplace App offers employees a faster, easier, and more visually appealing solution for booking their workspaces, according to the Head of Facility Management of BayWa. BayWa is now rolling out the system to more of their offices.

The second case study is about the company **Tegel Projekt GmbH**[2]. The company's goal is to transform the old airport site that closed in 2021 into a real-life testing ground for developing solutions that will shape the city of the future. The airport had never used CAFM software, but for this project, the company saw potential in Planon's IWMS. The IWMS has been the backbone of the project to work efficiently and in full compliance with the law with just seven employees.

For this project, IoT sensor technology is vitally important, as it enables the team to easily digitalise and monitor existing buildings. The sensors that were placed could measure: desk occupancy, CO_2 and temperature, and motion sensors for

[1] https://planon.showpad.com/share/I4nv41WvGEAcdpTHpY25z

[2] https://planon.showpad.com/share/ugCbScbEexP2CmSvxxQsk

meeting rooms and common spaces. With all this information, Tegel Projekt GmbH could adhere to the German labour laws, reduce cleaning costs for spaces that are not used, and achieve more efficient space and occupancy management with the use of desk booking.

The last case study is about the implementation of the Planon software at the **University of Eindhoven**[3]. The problem was that the space was not used efficiently, so they wanted to improve the current situation with a standardised solution. With the integrated solution of Planon, they could connect it to their roster scheduling system to have it work together.

The university implemented, next to the software, kiosks to let the workers and students have better insights into the available rooms. And with the Planon app, it became easier to book a room or study location. After already three months, there were improvements in the space usage.

5.4 Connection to the Preliminary Research

With a broad understanding of Planon established, it's crucial to explore its connection to Chapters 3 & 4.

As described in this chapter, Planon provides various software solutions that can automate routine tasks for employees. Moreover, Planon plans to integrate AI to enhance their software's efficiency, effectively transforming them into "AI assistants" just like the assistants mentioned in Chapter 3. This shift is driven by the growing significance of AI and the need to outpace competitors exploring similar technologies.

Planon seeks guidance on implementing changes clearly so customers understand and remain satisfied with the software. Chapter 4 addresses this necessity. A dual-focused explanation, encompassing both technical and human aspects, is vital for implementing AI. The technical details support both Planon developers and the client's IT team using the software. Meanwhile, HCXAI offers users insights into the AI assistant's decision-making process.

[3] https://planon.showpad.com/share/qEmI8Ve7mkg7wX46xXSoe

Chapter 6

Ideation

This chapter covers the project's ideation phase, detailing the idea-generation process. It features outcomes from the initial survey and interviews used to understand user requirements for an XUI. The chapter concludes with the ideas developed during this phase.

6.1 Method

A comprehensive set of guidelines from the literature is provided for designing an XUI. These guidelines are used during brainstorming to capture initial ideas. To narrow down these options, a survey will be conducted among Planon employees via email to identify their preferences for AI and XUI utilization. The complete list of questions of the survey is available in Appendix B. Following the survey, a brief interview will be held to gain further insights and verify survey responses. Participants consented to future contact by providing their email addresses at the survey's end, separate from their answers. With the answers from the participants, new ideas and alterations of previous ideas are made. After the drawings on paper, the majority of the ideas are digitalized in Figma [1] with the corporate identity of Planon to have a better visualization and grasp of how it could look. Those ideas are then grouped together based on type of explanation and features.

6.2 Results of the survey and interview

The survey received 61 responses from a diverse group at Planon Nijmegen, encompassing varied specialisations and expertise. Some respondents noted in the "other remarks" section that certain questions or topics were outside their expertise and could not provide well-informed answers. These responses will be excluded.

[1] Figma:https://www.figma.com/

Four individuals participated in interviews, and their feedback largely aligned with the survey results, offering mainly clarifications or minor additions.

The survey covers various subjects, starting with the overall AI experience. Analysing the responses reveals that every participant has interacted with an AI at least once, predominantly evoking positive emotions such as: Helped, Surprised, Excited, Curious, and Happy. Of the negative emotions, Disappointed was mentioned the most, more than Confused and Frustration. It was further supported by the interview that most participants found the AI assistant helpful when it performed tasks such as summarizing text, drafting email templates, or coding. However, disappointment arose when it provided incorrect information or failed to understand the question's intent. The survey reveals that most respondents view the use of AI in software beneficial and employ it in their work. Additionally, they do not perceive AI as a threat, nor do they become frustrated when its reasoning is unclear. Participants generally trust AI to deliver satisfactory results but still prefer to verify these results independently at other sources.

From these responses, it's evident that people favor the use of AI in applications. However, the key challenge lies in preventing or reducing dissatisfaction with AI's results.

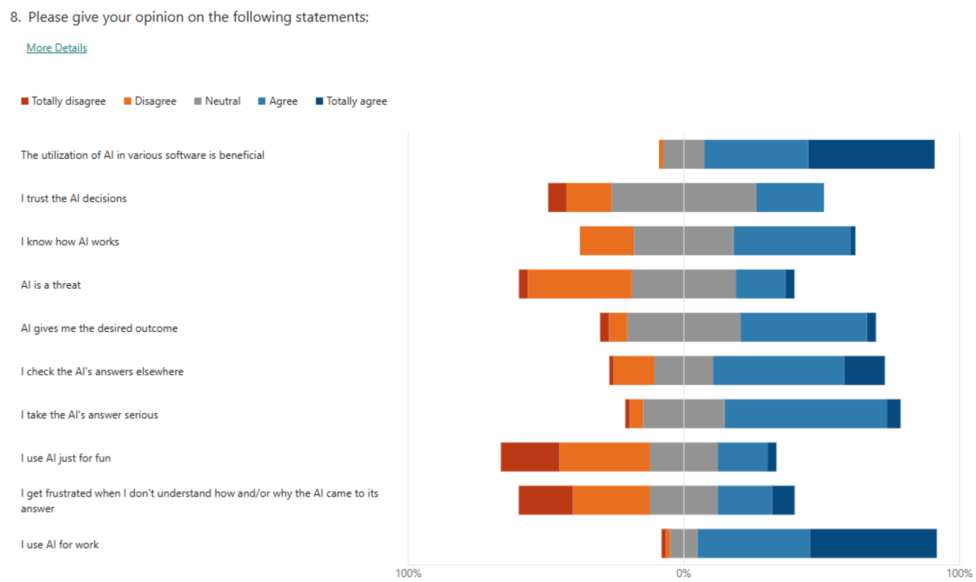


Figure 6.1: Various statements on AI

Section 4 on Explainable AI provided valuable insights into the explanation style as well as the interaction types and features of the XUI. In terms of the style of the explanation, the participants are most interested in the *"why"* explanation. Followed by example-based explanations and performance aspects, respectively. From the interview, it came to light that the type of explanation that helps people the most is the "why" explanation; the example-based explanation is mostly preferred beforehand as a guide on how to start prompting. Nonetheless, individuals may revisit it due to uncertainty about why their prompt may be incorrect.

Participants favoured dialogue interaction with the AI the most. Subsequently, they preferred the XUI used primarily for (information) transmission to users, as can be seen in Figure 6.2. In contrast to the literature, the control options received minimal support in the survey responses.

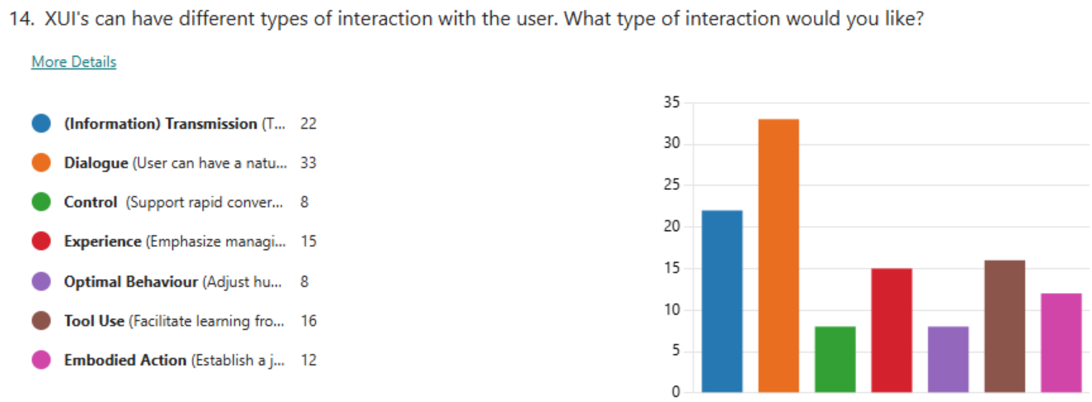


Figure 6.2: Type of interaction with the XUI

The findings indicate a strong preference for the "why" explanation style. Since the survey didn't specifically address global and local explanations—potentially due to their abstractness—no conclusions can be drawn in that area. Participants were drawn to familiar interaction types: dialogue and information transmission.

Delving deeper into the functionalities of the XUI. The timing of XUI availability is a crucial factor. 21% of the participants want the XUI always available, compared to the 77% that want it only after they have asked for it, see Figure 6.3.Participants were also required to prioritize additional potential features for the XUI, in descending order of significance, as illustrated in Figure 6.4.

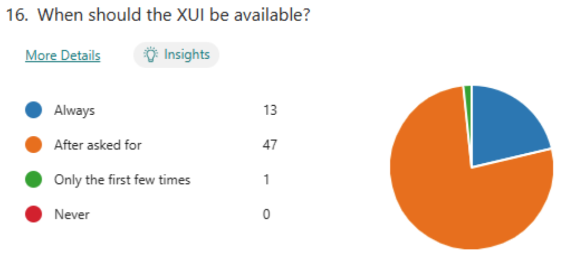


Figure 6.3: When XUIavailable

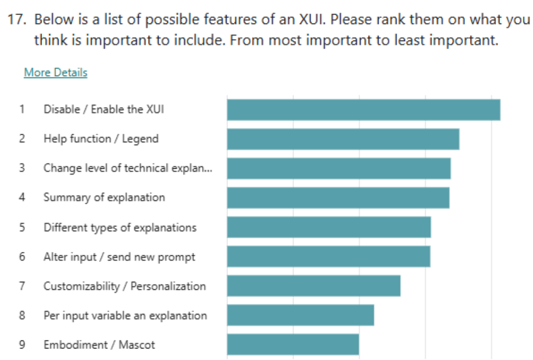


Figure 6.4: Ranked functionalities

These findings highlight the specific smaller features that users favor in the XUI. Though minor, these features significantly enhance the overall UX. As the example-based explanation and help function score highly in their respective surveys, they will be integrated into the prototypes.

Finally, there was a section for people who rely on others to reserve rooms. Approximately 40% of the respondents occasionally have someone else handle the booking. Over 20% of the time, these rooms fell short of meeting requirements, predominantly due to lacking facilities.

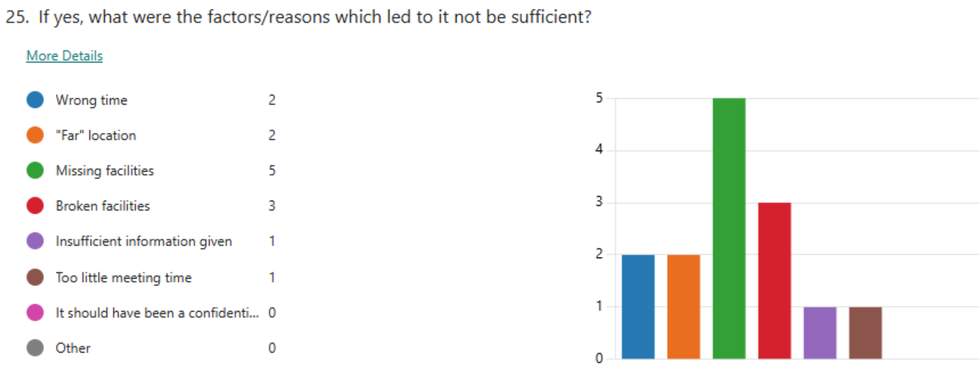


Figure 6.5: Reasons why the room did not suffice

From this, it can be seen that AI assistant can probably reduce that percentage by a lot.

Reflecting on the broader context of an XUI, it's crucial to understand user objectives with the XUI. Interviews revealed that users employ the XUI both to accept results and modify prompts. Users appreciate having example-based explanations either in advance or when results deviate significantly from expectations. Additionally, users tend to accept explanations tied to their outcomes, believing these insights subconsciously guide their future prompts. Consequently, the next section's ideas address both objectives to accommodate all user preferences.

6.3 Ideas

This section summarizes various concepts generated from different brainstorming sessions. Initially, brainstorming is guided by existing literature. Then, a following session incorporates insights from survey and interview responses. Ultimately, these sessions contribute to creating a design space that combines all the ideas.

6.3.1 Brainstorming Ideas

To generate numerous ideas, quick brainstorming is effective, as there are various methods to explain AI decisions. In the context of room booking, several clear (input) parameters of the AI can be derived, such as location, time, required facilities, and attendees. Initially, rough sketches of potential ideas are created, which are later digitized in Figma.

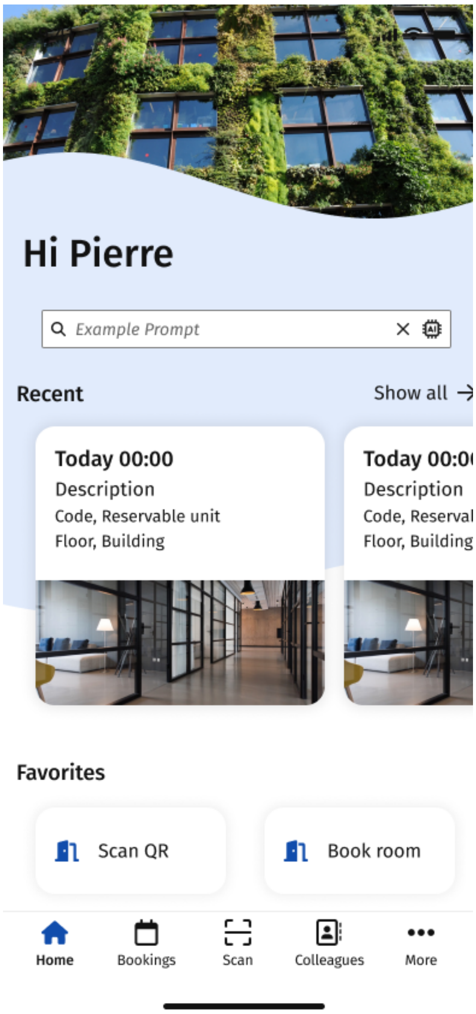


Figure 6.6: Prompt input-bar

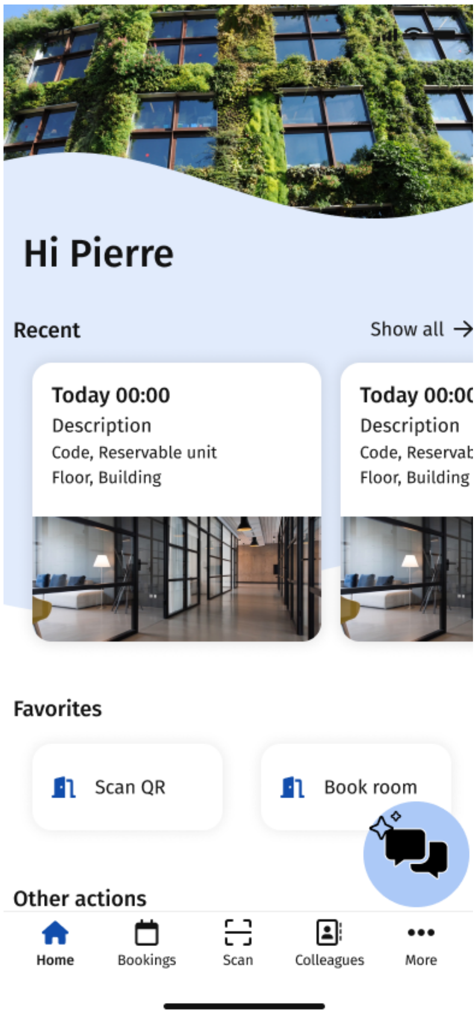


Figure 6.7: Chat overlay

To consider how an explanation might be conveyed, it's crucial to understand the overall presentation of the AI assistant. How do users access this assistant? Will it resemble a chatbot? Initial brainstorming explores its appearance on the workplace app's main page, including user access methods or immediate prompt entries. Figures 6.6 and 6.7 illustrate two concepts. Subsequent brainstorming divides explanations into global and local categories. Global explanations provide overviews on the "results" page, as shown in Figures 6.8 and 6.9, while local explanations focus on specific input variables.

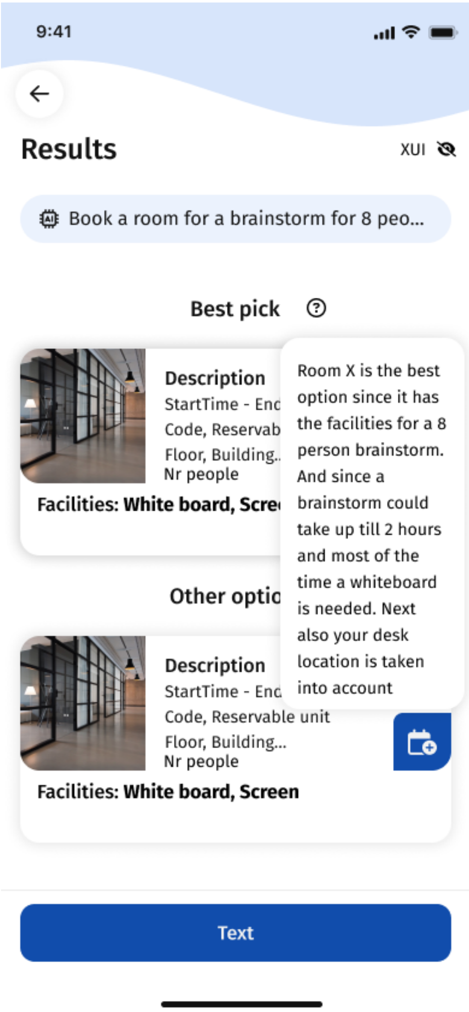


Figure 6.8: Summary explanation of why this is the best room

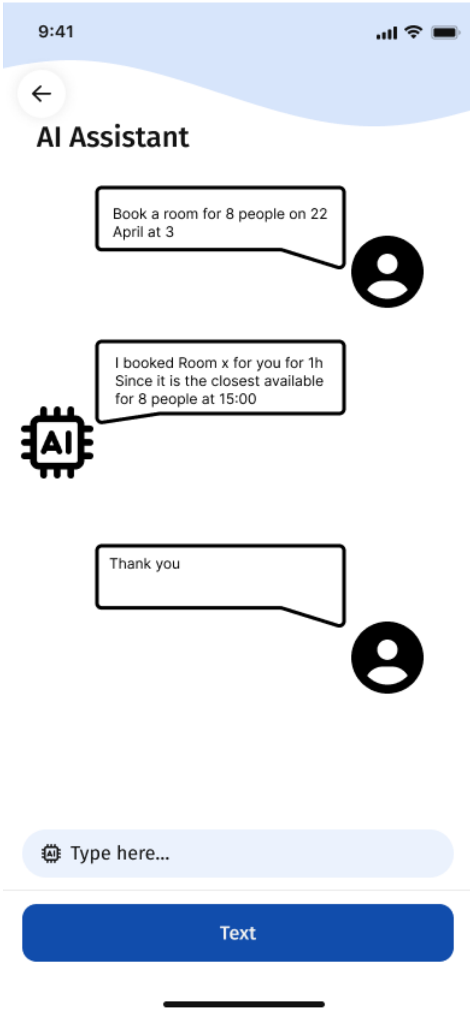


Figure 6.9: Chatbot variation with explanation

Global explanations can be combined with local explanations to give a deeper understanding. Figures 6.10 and 6.11 provide examples of location and facilities explanations. The meeting location is reasoned by proximity to attendees, and facilities are evaluated to ensure they meet all specified criteria for the meeting.

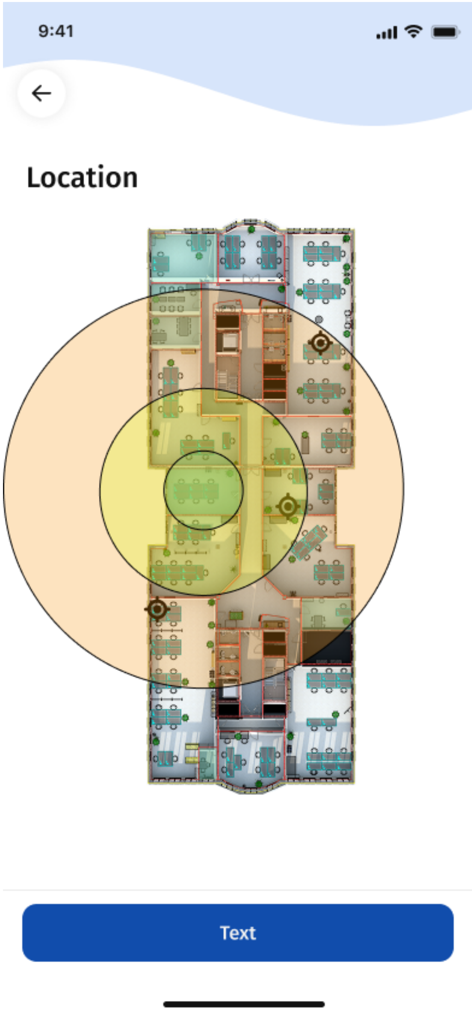


Figure 6.10: Location indication in comparison to the the participants

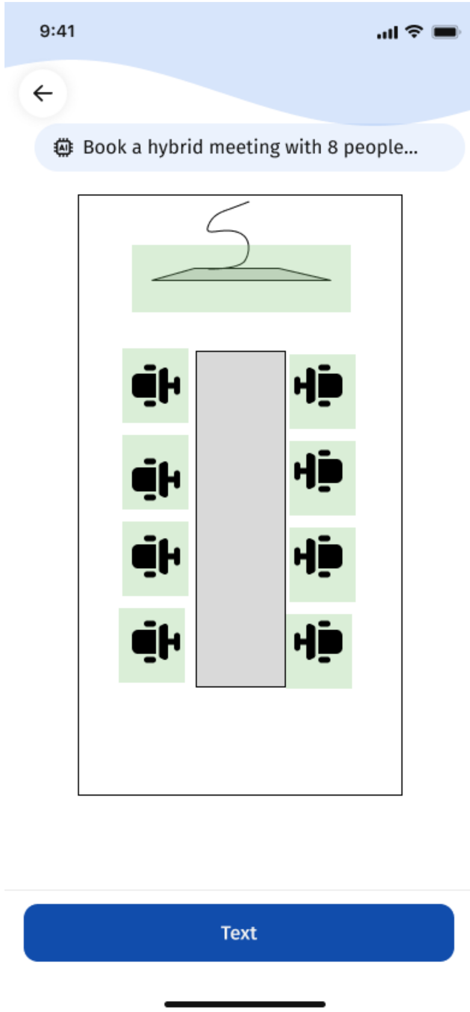


Figure 6.11: Available facilities of the room

6.3.2 Survey Influenced ideas

Based on the survey and interview outcomes, additional concepts and alterations are made. These centre on XUI functionalities, such as toggling explanations, a preference for summaries over individual input variables, and a help feature for XUI understanding. Since most users prefer the XUI upon request, some ideas address this preference.

Figure 6.12 provides a global explanation through a legend that clarifies the highlights. Meanwhile, Figure 6.13 introduces a dialogue variation where explanations are offered upon request.

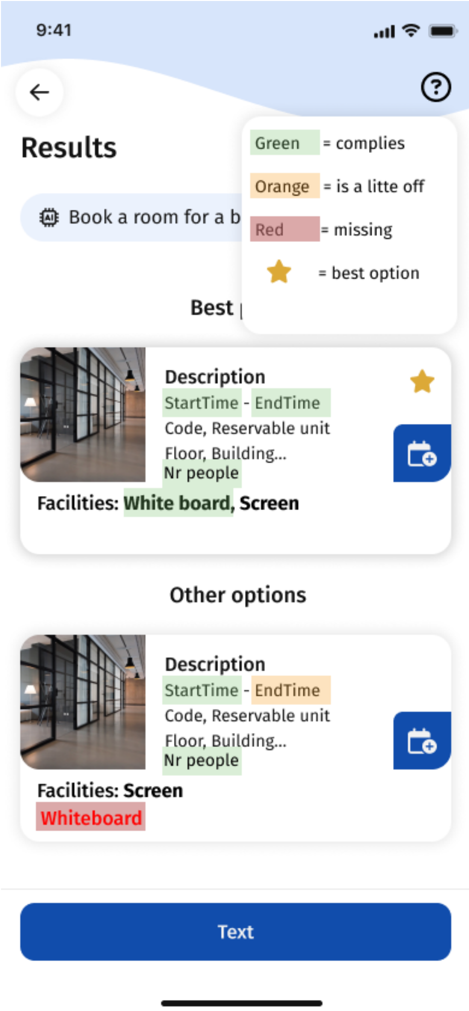


Figure 6.12: Highlights on parameters with legend

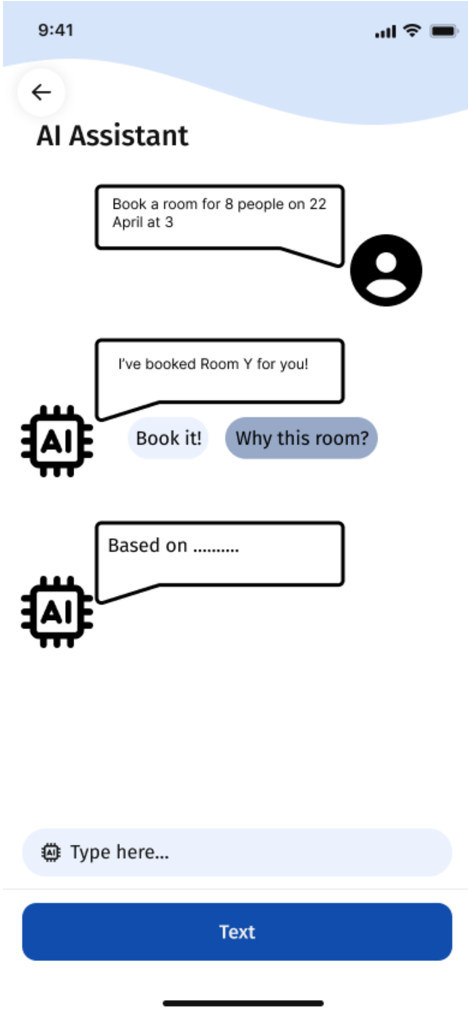


Figure 6.13: Alternative dialogue version

Another method of explanation that has high responses is the example-based explanation. In this case the XUI gives examples of different kinds of prompts to see what the outcome will be. In Figure 6.14 a menu can be seen where users can navigate to pages that give different kinds of explanations to help the user understand. In Figure 6.15 an example of example-based explanations can be seen.

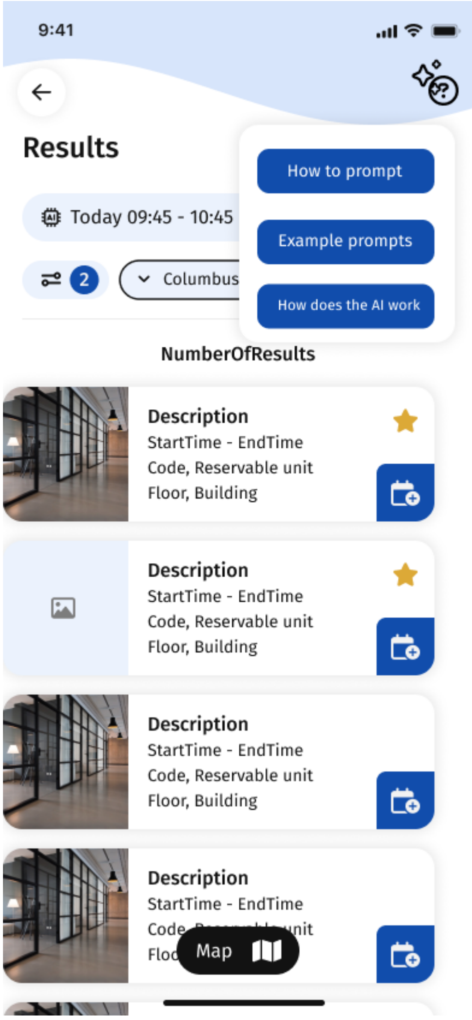


Figure 6.14: Menu to different explanations

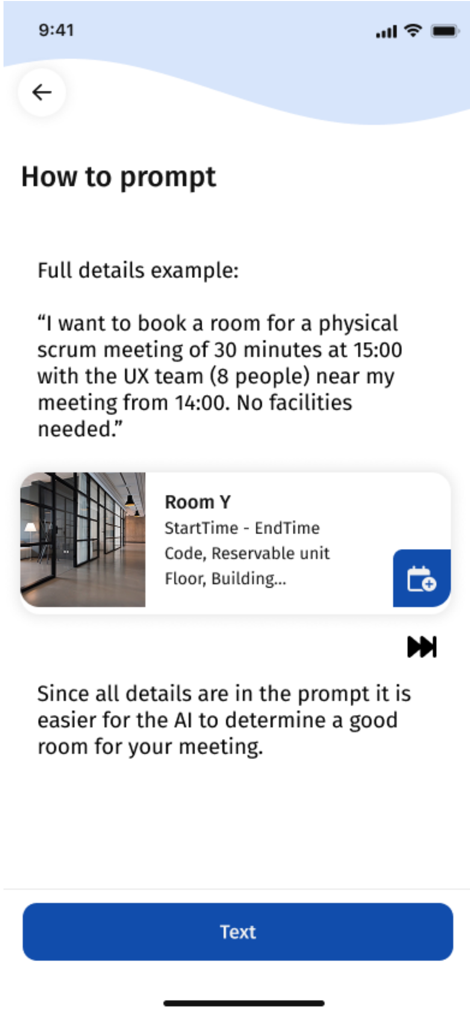


Figure 6.15: Example-based explanation

6.3.3 Design space

Once ideas are generated, they are organized into one design space. The created groups include: *Front page, Dialogue, Local explanation, Global explanation, input variable, Satisfaction, Help function, Settings*. Although some Front page concepts might fit into another category like Satisfaction, they are grouped together because the Front page serves as the entry to the XUI / AI assistant. The distinct groups are illustrated in Appendix C.

As shown in Figure 6.16, each idea is marked with a cross, check mark, or question mark to assist in selecting ideas for prototyping. Multiple ideas are combined to develop four distinct prototypes for evaluation.

Figure 6.16: Design space of the ideation phase

Chapter 7

Designed Prototypes

The upcoming section delves into the creation of initial prototypes. Drawing upon literature guidelines, brainstorming sessions, and survey and interview feedback, various prototypes will be developed, merging and enhancing concepts from the ideation phase.

7.1 Design Method

To create the prototypes of the XUI, the software Figma[1] is used. The software enables users to create application interfaces and includes a prototyping feature for creating interactive mock-ups. Users can link actions to specific buttons to create a flow through different screens of the application prototype. Figma is an excellent tool for testing and experimenting with design variations to visualize the final outcome.

The prototype's design space concepts are being refined for improved visual appeal. Then, multiple ideas are combined to create a cohesive and comprehensive explanation.

For this user test, four prototypes are made to make different combinations between interaction style and type of explanation possible. The combinations that will be made are:

- *Dialogue interaction - Global explanation*
- *Dialogue interaction - Local explanation*
- *(Information) Transmission interaction - Global explanation*
- *(Information) Transmission interaction - Local explanation*

[1] Figma:https://www.figma.com/

For the dialogue interactions, extra variables are made in Figma to hide and show certain components of the prototype. In Section 7.2.3 and 7.2.4 all the components that are hidden at first are shown.

In these prototypes, the AI considers potential parameters to inform its decision. These encompass *time, number of participants, location*, and *required facilities*. This approach yields both an optimal choice and alternative options for the user's final selection.

7.2 Visuals

Diverse prototypes and their ultimate designs are discussed in this section. Certain prototype components remain identical, such as the example pages, which appear consistently across all prototypes. Refer to Figure 7.2 for the example pages' layout. Additionally, the homepages are uniform across interaction types. For details on these designs, see Figure 7.1.

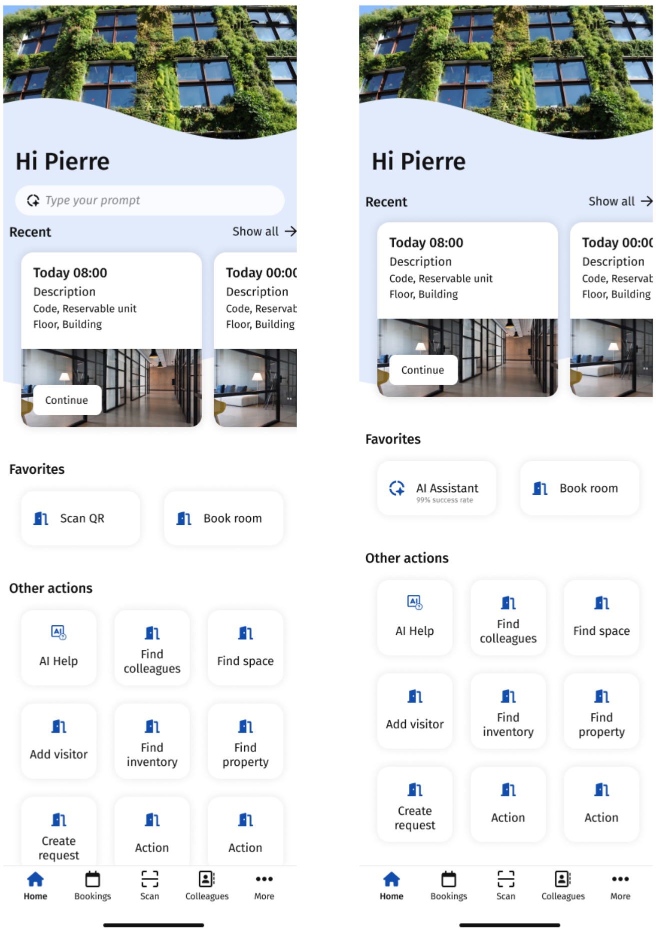


Figure 7.1: Homepage variation used - **Left**: Transmission **Right**: Dialogue

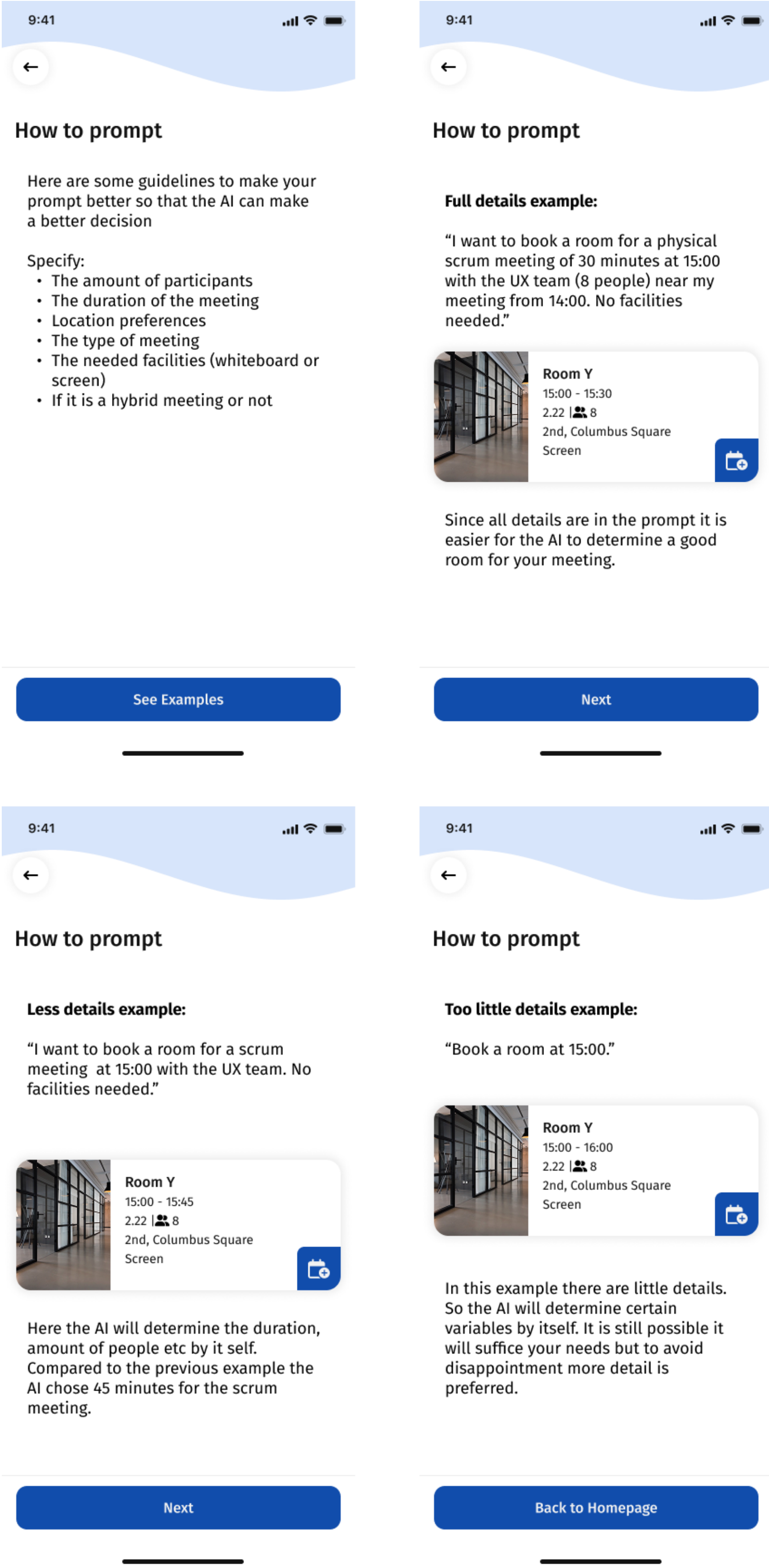


Figure 7.2: Examples pages with three different examples

To integrate the various pages within the prototype, flows are established. Illustrated in Figure 7.3 is the flow of the Transmission Global prototype, demonstrating how the components are interconnected. The flow and all prototype screens of the other prototypes can be found in Appendix D.

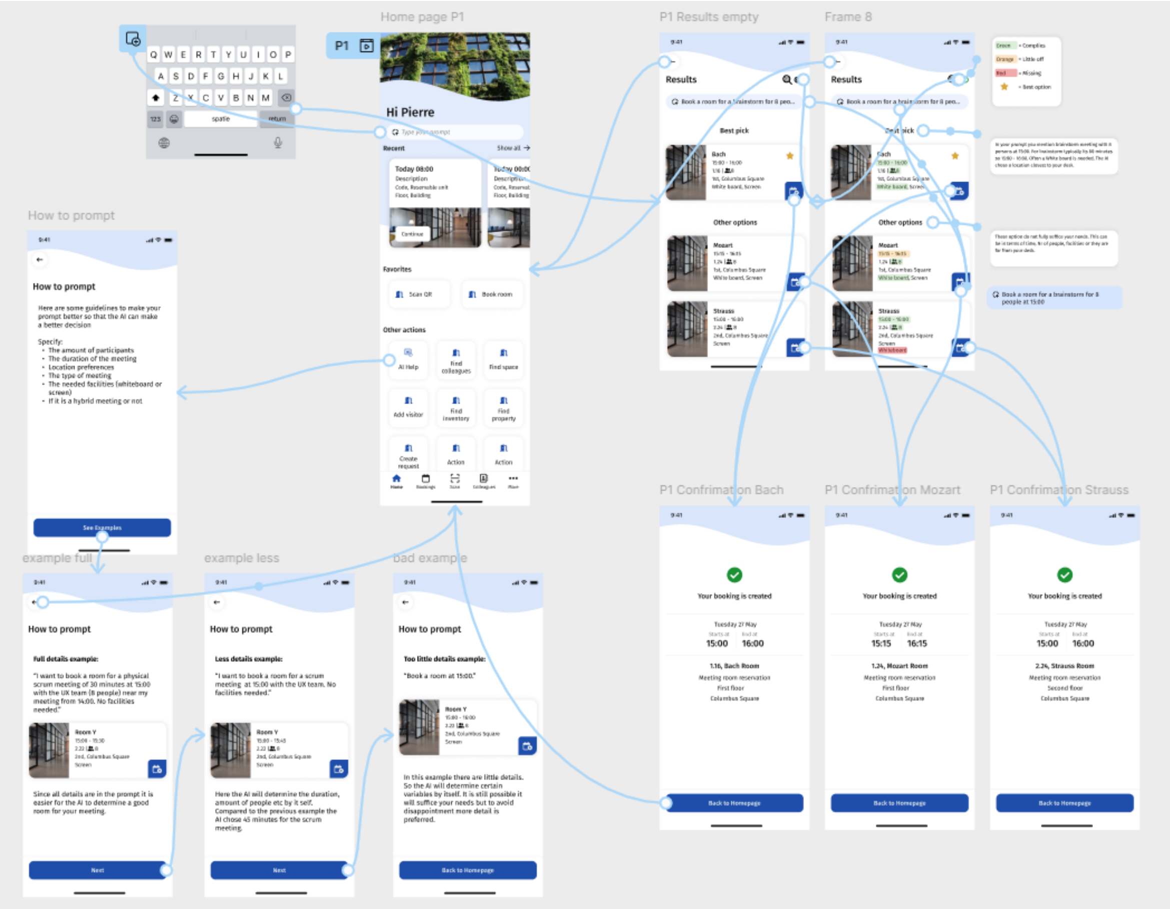


Figure 7.3: Flow of Transmission Global Prototype

7.2.1 Transmission Global

A Transmission Global explanation prototype primarily features various color highlights and a summarized explanation of the best pick and the other options. Figure 7.4 displays the XUI activated with its overlays next to it.

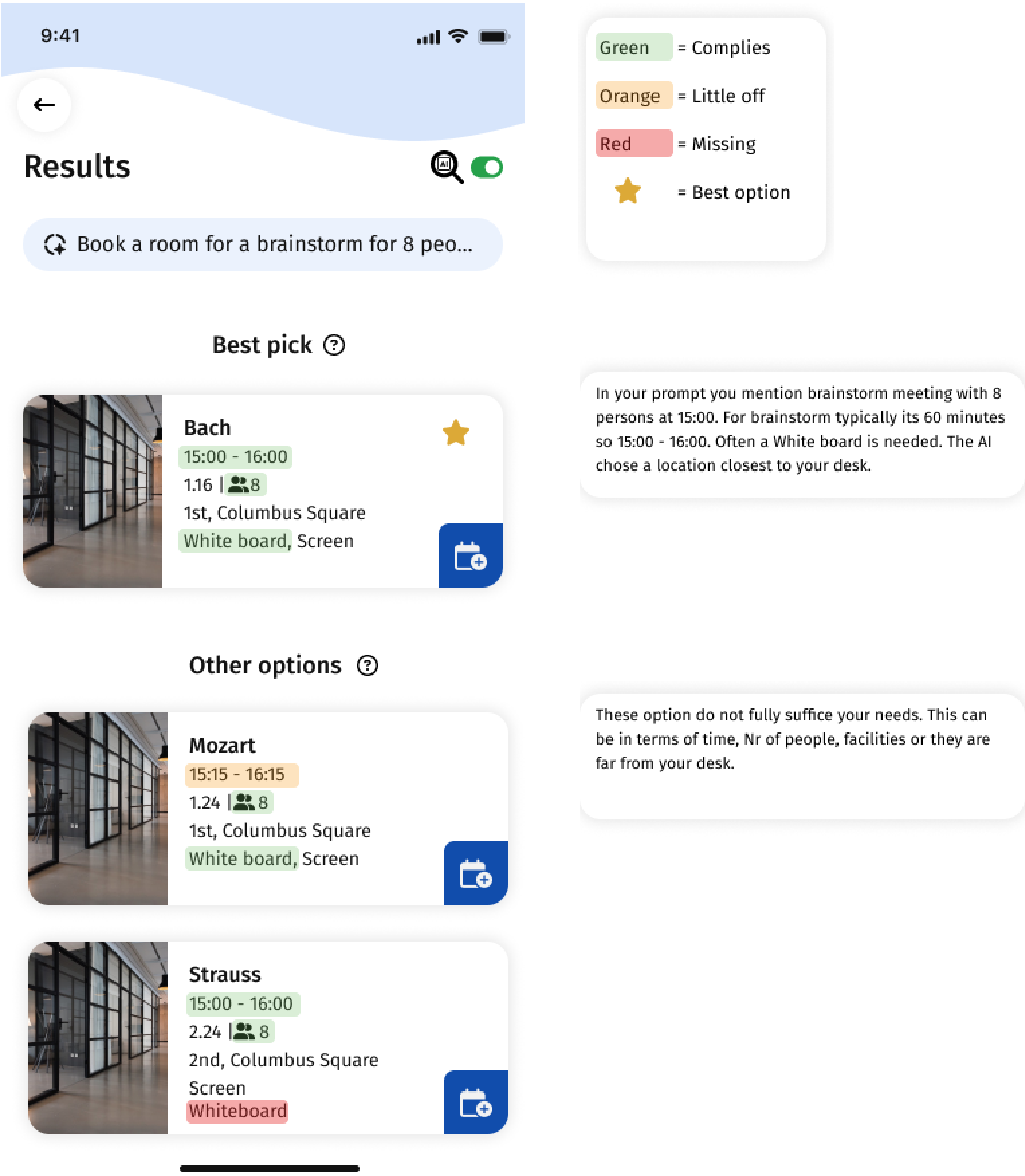


Figure 7.4: Transmission Global prototype with highlights and summary explanation

7.2.2 Transmission Local

The Transmission Local explanation prototype prominently includes additional pages detailing location and facilities, as well as the number of individuals involved. Figure 7.5 displays the XUI activated, showing the location explanation page.

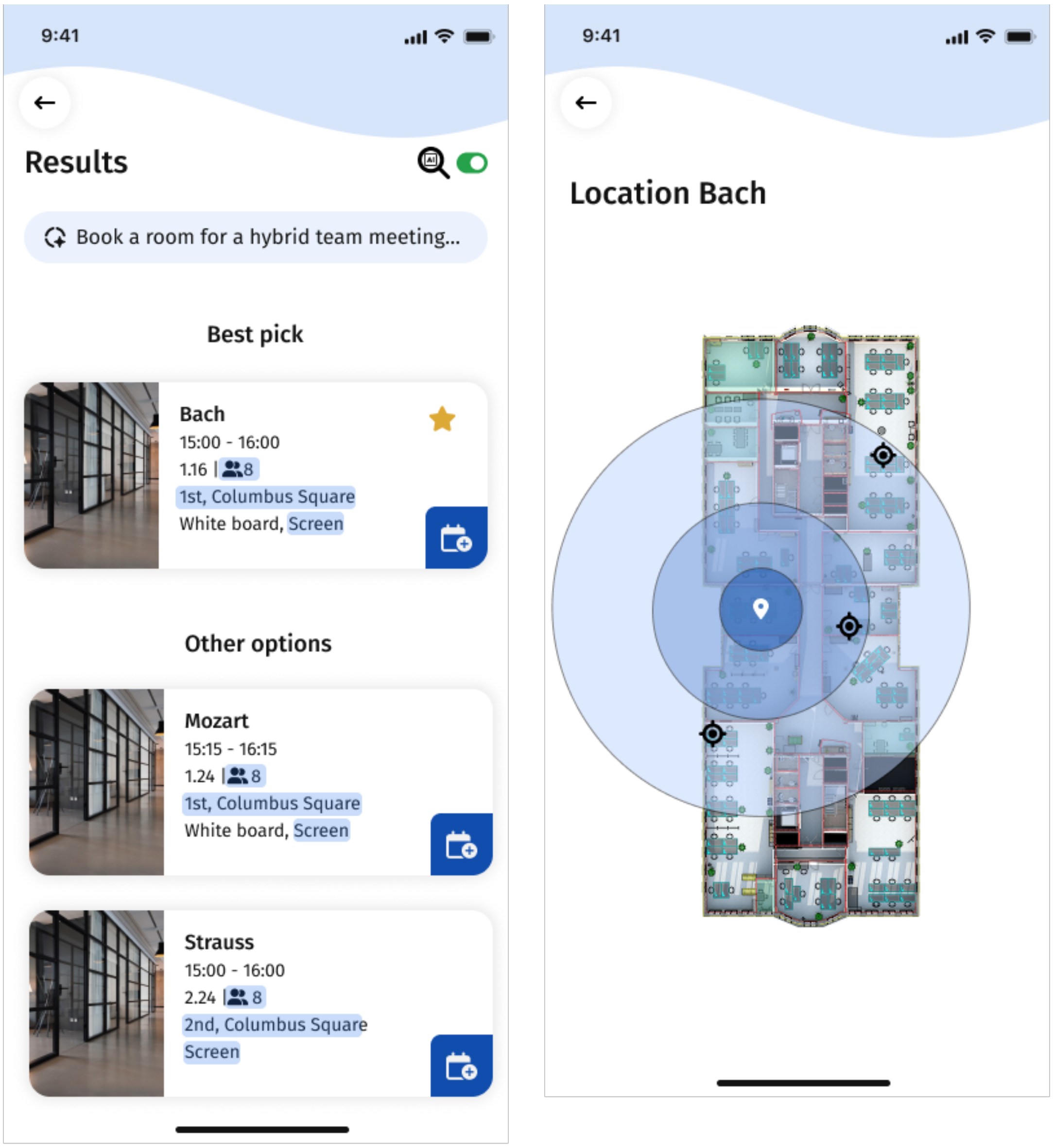


Figure 7.5: Transmission Local prototype with highlights to more in-depth explanations

7.2.3 Dialogue Global

The Dialogue Global explanation prototype functions similarly to a chatbot, responding to the questions from the user. In this mock-up, it addresses the question *"Why this room?"* by providing a concise explanation. Refer to Figure 7.6.

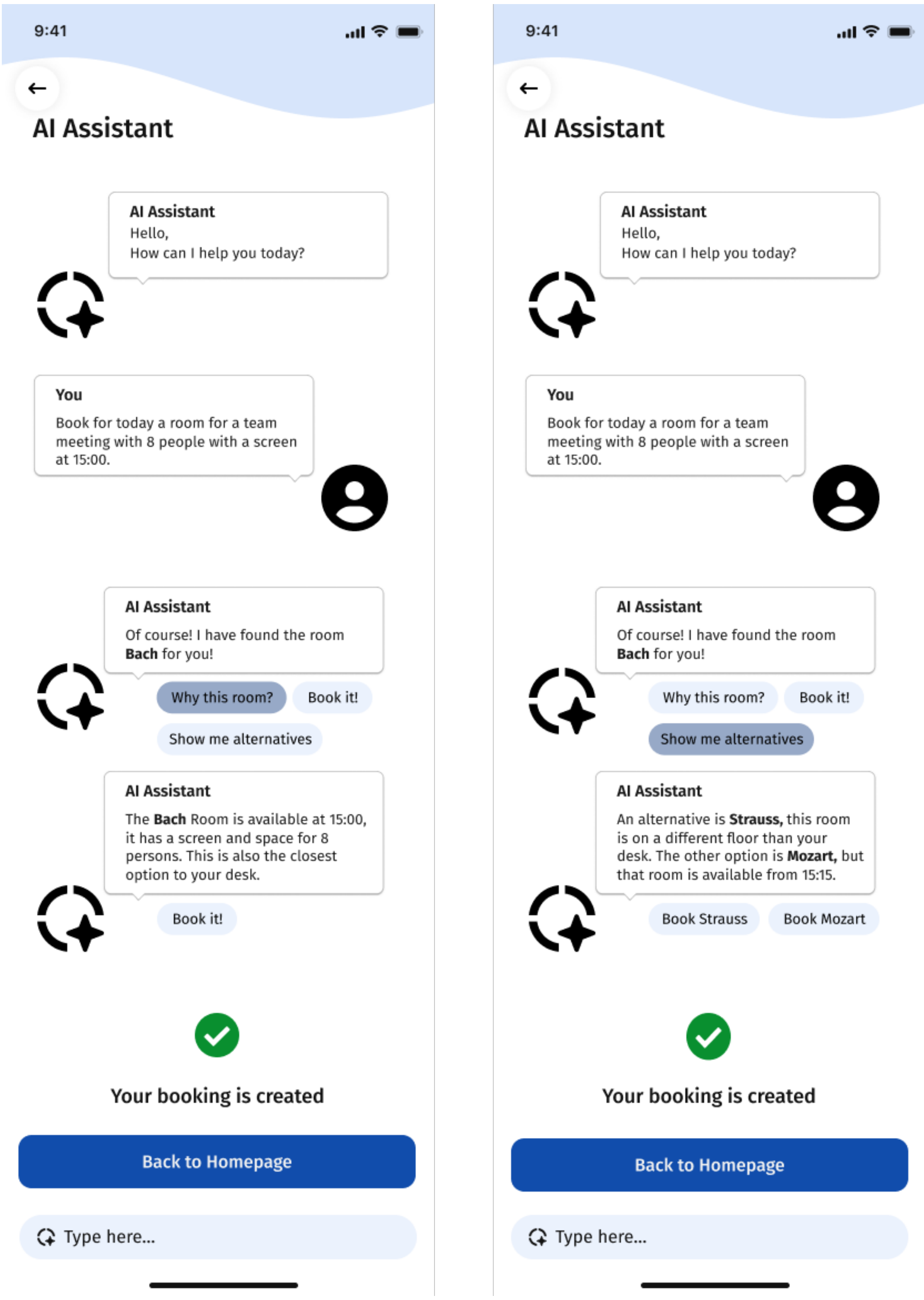


Figure 7.6: Dialogue Global prototype with conversations shown and possible reply answers

7.2.4 Dialogue Local

Instead of providing a summary, the Dialogue Local explanations prototype offers options to access local pages concerning location, facilities, and capacity. Refer to Figure 7.7. These local explanation pages are similarly utilized in Transmission Local explanation.

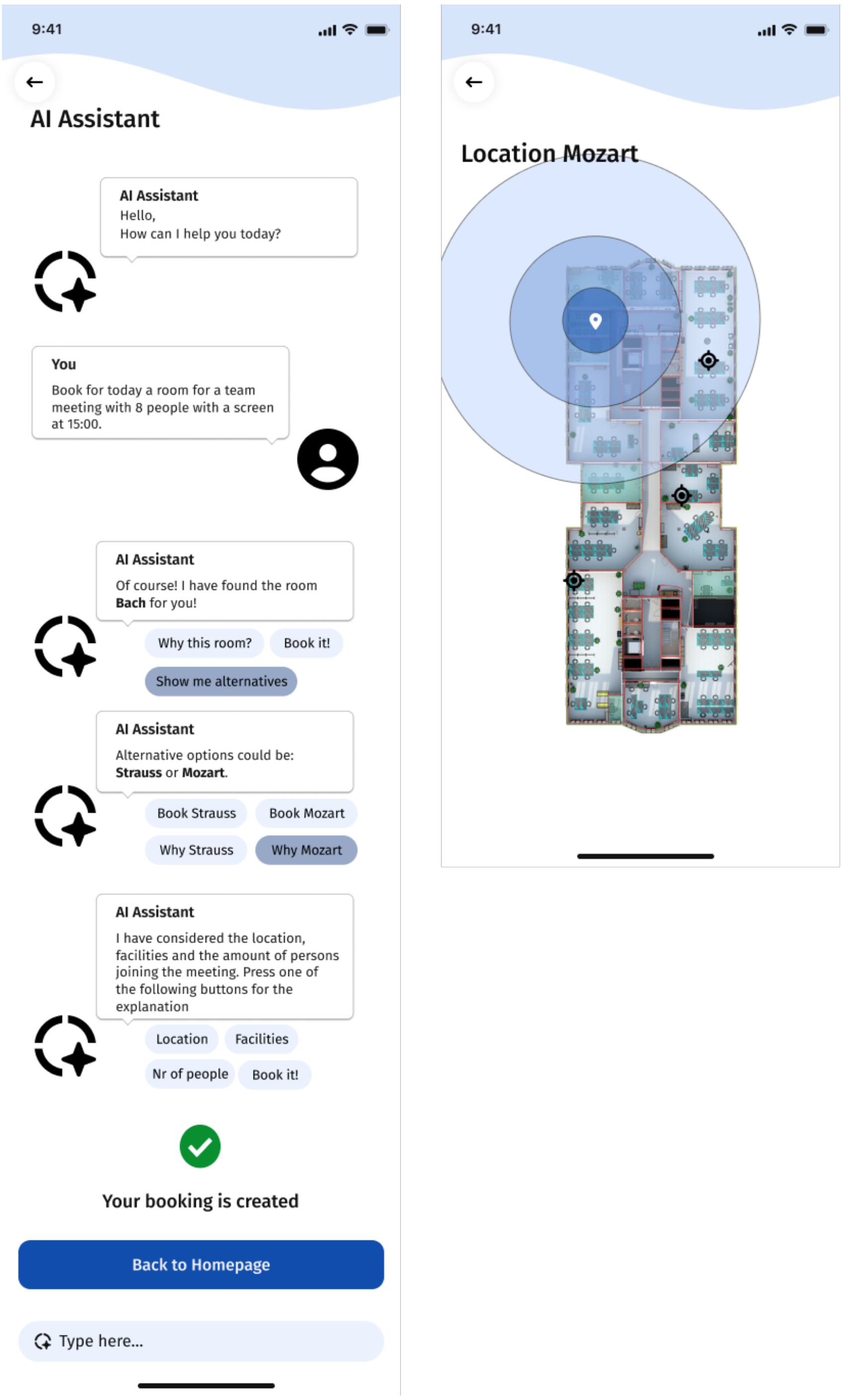


Figure 7.7: Dialogue Local prototype with conversation leading to a more in-depth explanation of the (in this case) location

Chapter 8

Evaluation

This chapter focuses on evaluating the prototypes. Initially, the objectives are defined, and an evaluation plan is developed. Following this, a quantitative and qualitative analysis of the outcomes is conducted, culminating in a conclusion.

8.1 Objectives

The main objective of the evaluation is *to understand why the explanation works well or not*. Reflecting on the main objective of this step, *"What are important design requirements and considerations to take into account when designing an XUI for AI assistants for routine tasks in the workplace?"*, and in combination with the answers of the survey and interview, some important evaluation points can be derived.

- *What are good design options for these prototypes?*
- *Would people prefer a Global or Local explanation?*
- *Would people prefer a dialogue or a (information) transmission interaction?*
- *What kind of features should be available in the XUI?*

8.2 Evaluation Plan

To establish an effective study, various techniques and standardised questionnaires are examined. A within-subject evaluation [11] is ideal, but with four prototypes, the user test would be too lengthy. To address this, participants evaluate two prototypes each, leading to twelve possible scenarios that the participants can follow. When the participant loads the user test, a web server on Google's Apps Script [1] retrieves the

[1] https://script.google.com/home

next scenario number from a spreadsheet. The code for this script can be found in Appendix E. This is done via a spreadsheet because, if participants withdraw from the user test, you can add the loaded scenario number back into the spreadsheet to ensure enough responses for that particular scenario.

A standardised questionnaire, with extra self-made questions, is used to evaluate the prototypes. The standardised questionnaire is the ESS for Explainable AI. Explanation Satisfaction is defined as the degree to which users feel that they sufficiently understand the AI system or process being explained to them [28]. This differs from the EGC. The Goodness Checklist focuses more on the systems designer's perspective and the Satisfaction Scale more on the user's perspective [28]. The self-made questions focus on the different features in the prototypes. The full list of questions can be found in Appendix F.

To reach enough participants, the user test will be carried out through a survey. This way, the user test can be sent to multiple different locations of Planon. The prototypes are embedded into the survey. The participants get a few tasks to perform with the prototype and answer the questions. The list of tasks can be found in Appendix F.

Before distributing the user test to participants, a pilot test will ensure functionality and check for any necessary question revisions. Once complete, the user test is dispatched to Planon employees in Nijmegen, The Netherlands, and Hyderabad, India.

After distributing the survey, responses are frequently reviewed to ensure balanced testing of all scenarios. As participants may open the survey for more details yet not continue due to limited time, insufficient topic knowledge, or unfamiliarity with Figma prototypes, the scenario assignment spreadsheet is adjusted to equally assess each scenario.

8.3 Results

In total, 36 participants were recruited for the user test. Each of the scenarios was tested 3 times; this results in that every prototype was evaluated 18 times. Half of the time as the first seen prototype and the other half as the second seen prototype. As there are just 3 participants for each scenario, with insufficient responses, analysis on the effect of viewing a certain prototype first cannot be conducted.

In the ESS, participants gave answers to nine questions on a 5-point Likert scale from ”I agree strongly” to ”I disagree strongly”. Table 8.1 presents the mean and standard deviation derived from the ESS's findings, while Figure 8.1 illustrates these in a boxplot. The answer options ”I agree strongly” to ”I disagree strongly” are transformed to a rating of 1.00 to 5.00.

Prototype	DG	DL	TG	TL
Mean	**1,98**	**2,33**	**2,38**	**2,42**
Std. Deviation	1,02	1,18	1,21	1,16

Table 8.1: ESS evaulation analysis

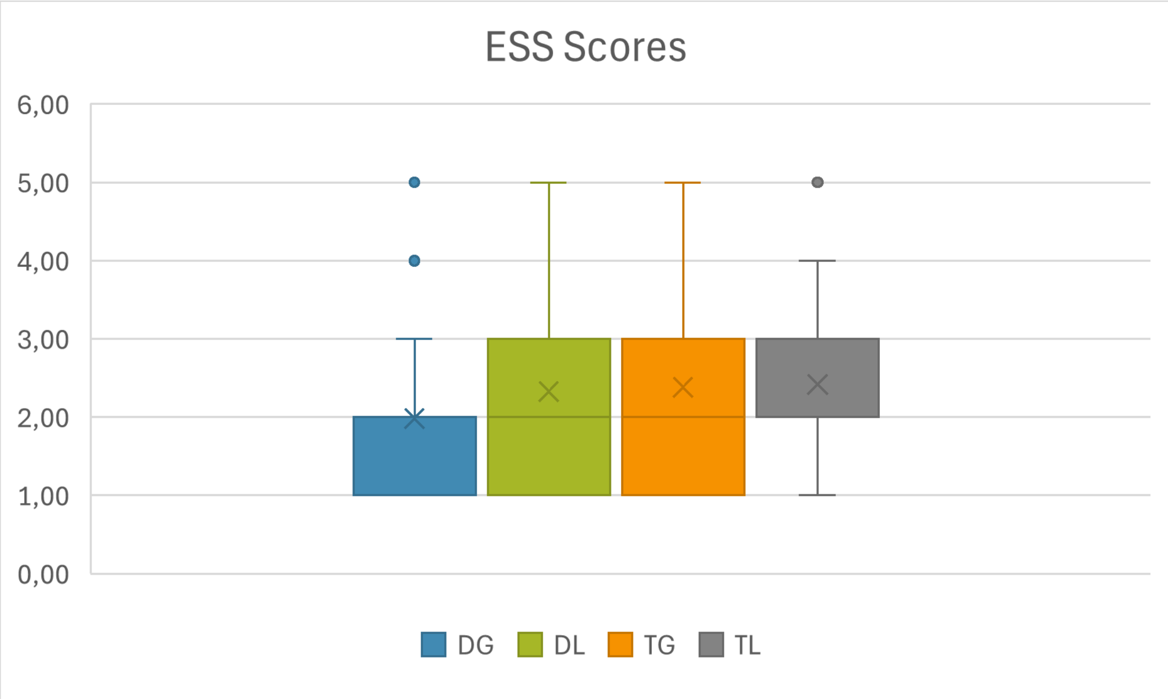


Figure 8.1: Boxplot of the ESS results for each prototype

What can be seen from these results is that Dialogue Global has the lowest mean with a score just under the 2 ("I agree somewhat"). The same is true for the standard deviation; Dialogue Global has the lowest value. That means that it has the most positive responses with the least variety in the answers. Transmission Local has the most negative responses which result in the highest mean. Question 9 of the ESS, "This explanation helps me know when I should trust and not trust the AI," shows an overall mean of 2.86 representing a neutral agreement with this statement across all prototypes. A lot of participants voted neutral about it. This might be explained by the fact that trust is something that probably cannot develop with approximately working 15 minutes with a prototype.

Before checking if there is a significant difference in satisfaction score, it is crucial to see if there is an internal consistency across all items of the ESS. This is done with a reliability analysis. The measure of Cronbach's Alpha is taken into account to determine the consistency. The overall reliability score across all the prototypes can be found in Figure 8.2. The Cronbach's Alpha value of 0.920, which is higher than

0.70, indicates that there is an internal consistency between the responses on the questions of the ESS. The reliability analysis output per prototype can be found in Appendix G; those values also indicate internal consistency.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.920	.922	9

Figure 8.2: Reliability analysis output across all prototypes (rounded to 3 decimals)

To determine if there is a significant difference in satisfaction scores across all the prototypes, an ANOVA analysis is done with a 95% confidence interval, see Figure 8.3. All the scores of all the questions of the ESS are taken into account for this analysis. The null hypothesis (H_0) suggests there's no significant difference, while the alternative hypothesis (H_1) argues otherwise. With a significance value of 0.003, which is below 0.05, thereby indicating a significant difference is evident, justifying the rejection of H_0.

ANOVA

ESS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	19.667	3	6.556	4.796	.003
Within Groups	880.333	644	1.367		
Total	900.000	647			

Figure 8.3: ANOVA analisis between all prototypes (rounded to 3 decimals)

A follow-up analysis will employ a two-tailed t-test with a 95% confidence interval to determine if a significant difference in satisfaction score exists between Dialogue or Transmission interaction and the Global and Local explanation style. See Figure 8.4 for these findings. For the first pair, the scores for all Dialogue prototypes are combined, and similarly, the scores for Transmission prototypes are also combined. And for the second pair, all the scores of the Global prototypes are combined and the same for the Local prototypes. This resulted in a significance value of $<$0.001 for both evaluations, which is below 0.05, thereby indicating a significant difference.

Upon analysing the results, Dialogue interaction exhibits an average score that exceeds Transmission interaction by 0.250 points. And Global exceeds Local with 0.191 points.

Paired Samples Test

		Paired Differences							Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	One-Sided p	Two-Sided p
Pair 1	Dialogue - Transmission	-.24691	.65422	.03635	-.31842	-.17541	-6.793	323	<,001	<,001
Pair 2	Global - Local	-.19136	.60987	.03388	-.25801	-.12470	-5.648	323	<,001	<,001

Figure 8.4: T-test analysis between interaction style (rounded to 3 decimals)

Lastly, a post-hoc Tukey analysis, on the previously done ANOVA analysis, with a 95% confidence interval will be held to determine if a significant difference in satisfaction score exists between each of the four prototypes. Those results can be seen in Figure 8.5. Comparing the Dialogue Global with the other prototypes, all the significant values are below 0.05, thereby indicating a significant difference. It is evident that the mean differences between Dialogue Global and the others span from 0.35 to 0.44 points, expressing greater satisfaction with Dialogue Global. There is no significant difference between the other prototypes (Dialogue Local - Transmission Global, Dialogue Local - Transmission Local, and Transmission Global - Transmission Local).

Multiple Comparisons

Dependent Variable: ESS

Tukey HSD

(I) Prototypes	(J) Prototypes	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
DG	DL	-.34568*	.12991	.040	-.6803	-.0111
	TG	-.40123*	.12991	.011	-.7358	-.0666
	TL	-.43827*	.12991	.004	-.7729	-.1037
DL	DG	.34568*	.12991	.040	.0111	.6803
	TG	-.05556	.12991	.974	-.3902	.2791
	TL	-.09259	.12991	.892	-.4272	.2420
TG	DG	.40123*	.12991	.011	.0666	.7358
	DL	.05556	.12991	.974	-.2791	.3902
	TL	-.03704	.12991	.992	-.3716	.2976
TL	DG	.43827*	.12991	.004	.1037	.7729
	DL	.09259	.12991	.892	-.2420	.4272
	TG	.03704	.12991	.992	-.2976	.3716

*. The mean difference is significant at the 0.05 level.

Figure 8.5: Tukey output ESS

From the quantitative analysis, it can be seen that Dialogue Global has the highest satisfaction score. Later on in the survey, the participants were asked about their preference for a prototype. 58% of the participants indicated a preference for a Dialogue interaction, and there is no difference in Local or Global explanation preference, indicating that a more in-depth investigation is needed.

In addition to this analysis, the open-answer questions responses indicate as well a clear divide among participants regarding their preference for the level of detail in the explanation. For some, the small summary was enough, but others liked more details and even wanted more detailed explanations. From these answers and the ranked functionalities from Figure 6.3, it can be concluded that users should be able to change the level of technical detail in the XUI.

Exploring further into prototype features, a frequently noted useful aspect was the highlights in the transmission interaction and the easy prompting via the buttons in the dialogue interaction. Additionally, the example page is universally cited as beneficial across all prototypes. 80% of the participants gave positive answers.

For features deemed redundant, this largely stemmed from a preference for global explanations over local ones. Those participants indicated the local explanation was redundant. Other feedback focused mainly on enhancing specific features.

An aspect that might have impacted the prototypes' evaluation is if it was clear to the participant where to find the explanation. This was again asked with a 5-point Likert scale from "I agree strongly" to "I disagree strongly". To investigate if there is a significant difference between the prototypes, an ANOVA analysis is done with a 95% confidence interval. In Figure 8.6, it can be seen that there is a significant difference, with a value of 0.001, in the ability to find the explanation. Additionally, in Figure 8.7 a boxplot of the scores is visualised.

ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	24.264	3	8.088	6.003	.001
Within Groups	91.611	68	1.347		
Total	115.875	71			

Figure 8.6: ANOVA analysis on "If the explanation was clear to find"

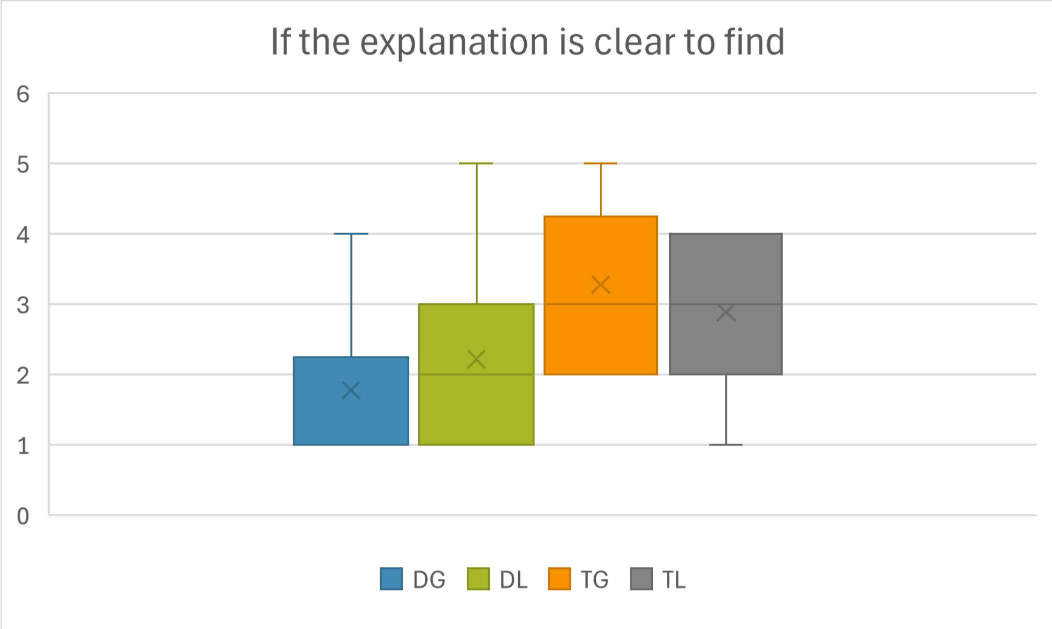


Figure 8.7: Boxplot of the "If the explanation was clear to find" scores

A two-tailed t-test with a 95% confidence interval to determine if a significant difference in "clear to find" score exists between Dialogue or Transmission interaction and the Global or Local style of explanation. It can be seen in Figure 8.8, that there is also a significant difference, with a value of $<$0.001, for Dialogue and Transmission. But there is no significant difference between Global and local explanations. Upon analysing the results, Dialogue interaction exhibits an average score that exceeds Transmission interaction by 1.083 points. And Global exceeds Local with only 0.028 points.

Paired Samples Test

		Paired Differences							Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	One-Sided p	Two-Sided p
Pair 1	Dialogue - Transmission	-1.08333	1.10518	.18420	-1.45727	-.70939	-5.881	35	<,001	<,001
Pair 2	Global - Local	-.02778	.94070	.15678	-.34607	.29051	-.177	35	.430	.860

Figure 8.8: T-Test analysis between Dialogue - Transmission and Global- Local

Delving further into the prototypes, a post-hoc Tukey analysis, on the previously done ANOVA analysis, with a 95% confidence interval, will be held to determine if a significant difference in satisfaction score exists between each of the four prototypes. From the results in Figure 8.9, it can be seen that there is a significant difference between: DG-TG, DG-TL, and DL-TG. With a mean difference of 1.500, 1.111, 1.056 points, respectively.

Multiple Comparisons

Dependent Variable: Score
Tukey HSD

(I) Prototype	(J) Prototype	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
DG	DL	-.44444	.38690	.661	-1.4634	.5745
	TG	-1.50000*	.38690	.001	-2.5190	-.4810
	TL	-1.11111*	.38690	.027	-2.1301	-.0921
DL	DG	.44444	.38690	.661	-.5745	1.4634
	TG	-1.05556*	.38690	.039	-2.0745	-.0366
	TL	-.66667	.38690	.320	-1.6857	.3523
TG	DG	1.50000*	.38690	.001	.4810	2.5190
	DL	1.05556*	.38690	.039	.0366	2.0745
	TL	.38889	.38690	.747	-.6301	1.4079
TL	DG	1.11111*	.38690	.027	.0921	2.1301
	DL	.66667	.38690	.320	-.3523	1.6857
	TG	-.38889	.38690	.747	-1.4079	.6301

*. The mean difference is significant at the 0.05 level.

Figure 8.9: Tukey output ”If the explanation was clear to find”

Moreover, respondents noted in the comparison questions that although dialogue interaction was preferred for its intuitiveness, users also appreciated the transmission interaction style and may adapt to it over time.

Furthermore, participants offered numerous suggestions to improve the prototype, including new features and enhancements for existing features. Notably, they frequently commented that it was confusing that the AI help page could not be accessed directly from the AI page. Additional recommendations included more prompt suggestions based on past ones and a ”book my favourite desk” suggestion. Participants also indicated that they would like to have the option to add options for the room like temperature. A participant recommended offering a tutorial for initial use, and voice command is also mentioned multiple times. Specifically for the dialogue interaction, they suggest that follow-up questions should appear as their own individual messages in the dialogue. And that there should be an obvious colour difference between messages sent by the AI and those sent by the user. Others added that the AI-generated output should be more distinctly coloured to indicate the use of AI. It was advised that local explanations should incorporate photos of the rooms instead of drawings, accompanied by a summary of available facilities, such as the number of chairs, to eliminate the need for user counting as required by the current design.

Considering the broader perspective on integrating AI into routine tasks, several participants expressed approval of its implementation. Some, however, were against using AI for these tasks and favoured the current method of manual filling in a form.

Those who supported AI usage highlighted its speed, which increases app usage likelihood.

The collected data strongly supports the claim that Dialogue interaction and Global explanation styles are significantly more effective in enhancing user satisfaction, usability, and discoverability of AI explanations. However, the need for customization, trust-building mechanisms, and clarity in interface design remains central to achieving the broader goal of creating explainable user interfaces that truly serve users in real-world tasks.

Chapter 9

Final Reflections

This chapter reflects back on the study done and will conclude its findings. It will formulate the contribution to the academic field. And it will discuss the findings and limitations to make recommendations for future research.

9.1 Conclusions

This study explored the different ways to create an XUI for AI assistants in routine tasks in the workplace. The main goal of this thesis is to find answers to the main research question: ***"How can an XUI be designed to enhance the usability of AI assistants used for routine tasks in the workplace?"***. Drawing from the literature, Chapters 3 & 4, a framework is developed to categorise user interactions and explanation styles, forming a foundation for crafting an XUI. Moreover, from the survey, Chapter 6, findings and user tests, Chapter 8, I can derive more specific guidelines and recommendations for designing an XUI to improve AI assistants' workplace usability for routine tasks.

Firstly, it was found that the Dialogue interaction has the most preference, significantly more satisfaction than the others, and it was also significantly clearer to find the explanation. That said, this applies to the dialogue and transmission application style that I present in this thesis; maybe other ways of transmission and dialogue interaction will act differently.

Secondly, in multiple instances, there was no clear difference in preference for a Global or Local explanation. Although Global's satisfaction score is significantly better, users indicated they want an option that the explanation style can change based on their preference. How this exactly should be incorporated requires future work.

Thirdly, utilizing examples on how to work with the AI proved effective for all participants, making it an excellent method for achieving user success.

Next, the XUI with the explanation should be provided once the user has requested it. Since not all users might want to see the explanation each time they have to book a room in this case.

And finally, participants indicated that they would like all outputs from AI to be clearly identified, for example, with colours. It is nowadays a commonly proposed strategy for reducing the risks of (generative) AI [80].

In addition to these guidelines, smaller recommendations that can be taken into account when designing an XUI were found to enhance the UX but should be applied on a case-by-case basis. Nevertheless, certain limitations exist, and further studies are suggested to enhance the validity of these guidelines, as mentioned in Section 9.3.

All in all, with these guidelines in mind, designs possess more tangible guidance compared to general guidelines.

9.2 Contributions

The field of XUI is still in its early stages, creating many opportunities for this and future work to explore. As can be read in the background, the domain of routine tasks in the workplace is especially not researched a lot. That is why this study can contribute a lot to this academic field.

First, a foundational contribution lies in the development of a conceptual framework that categorises potential user interactions and explanation styles tailored to the domain of routine tasks. This framework was constructed through a literature review combined with qualitative insights gathered from representative end users. The resulting taxonomy offers a structured lens through which designers and researchers can analyze and create interactive systems that require interpretability and user engagement.

Second, this framework was used to translate the identified interaction and explanation styles into design examples. These examples demonstrate how these abstract guidelines can be visualised into user interface solutions. Every design selects particular components from the framework, demonstrating how the proposed ideas can be applied and adjusted across various scenarios in the field.

Finally, the interfaces that were created went through user tests with end users. This evaluation phase not only validated the satisfaction of the explanation of the proposed designs but also generated a set of refined design recommendations and considerations for future work.

Together, these results enhance the expanding research on human-centred explanation design, serving as a valuable foundation for future studies. This thesis delivers a comprehensive contribution from theoretical groundwork to empirical evidence furthering academic insights and practical strategies for designing explainable, user-centric interfaces for AI assistants in routine tasks in the workplace.

9.3 Discussion and Future research

Upon reviewing the procedures and results of this study, quite a few findings suggest knowledge gaps that can be addressed in future research.

Starting with the usability of the design. To effectively assess usability and ensure user satisfaction with the XUI design, a long-term evaluation is necessary. This approach will allow us to measure trust in the software, as initial evaluations showed most participants had a neutral stance on this matter. Trust inherently develops over extended periods.

Alongside long-term assessments, the prototypes were exclusively evaluated by Planon employees; this should be extended to the Planon clients to get their input since they are also end users. With more participants, more data can be analysed to validate the choices made.

Secondly, the preference in explanation style. Although the results from the ESS show significant results in satisfaction between prototypes, with the best being Dialogue Global, the answers from a later question in the evaluation show no difference in preference between Dialogue Local and Dialogue Global. And from the open questions and answers from the initial survey, it can be seen that people are divided about the Global or Local explanation. Future research should investigate an option for users to change the explanation style as a kind of personalisation in order to create a suitable solution for everyone. The ontology from Chari et al. [10] has already incorporated the Explanation Modality where user preference is taken into account. Other ontologies that look at explanation models are the XMLPO [82] and the ontology of Nakagawa et al. [50]. In addition to that, Overeem [58] investigated both expert and non-expert versions. Despite no significant differences in results, it remains worth examining.

Furthermore, as previously mentioned, these findings relate to the design strategies outlined in this thesis. Additional adaptations of Dialogue and Transmission interactions should be explored to determine if the observed differences persist.

Furthermore, by recruiting more participants in the future, it will be possible to evaluate the impact of which prototype is viewed first.

Prototypes were developed and assessed using Figma, which inherently limits user interactions. Specifically, users couldn't type their own prompts, and follow-up prompts required button inputs. While participants reacted positively, further testing is necessary to determine the need for these buttons once free text input is enabled.

This study centres on room booking; however, it's crucial to investigate other routine workplace tasks to determine if the same findings apply to the other tasks.

Bibliography

[1] URL: https://planonsoftware.com/.

[2] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160.

[3] Wael Alharbi. "AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools". In: *Education Research International* 2023.1 (2023), p. 4253331.

[4] Benjamin Alt et al. "Human-AI Interaction in Industrial Robotics: Design and Empirical Evaluation of a User Interface for Explainable AI-Based Robot Program Optimization". In: *arXiv preprint arXiv:2404.19349* (2024).

[5] LR Biggers et al. "'Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models". In: *Empirical Softw. Eng* 19.3 (2014).

[6] Keshav Bimbraw. "Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology". In: *2015 12th international conference on informatics in control, automation and robotics (ICINCO)*. Vol. 1. IEEE. 2015, pp. 191–198.

[7] Federico Cabitza et al. "Quod erat demonstrandum?-Towards a typology of the concept of explanation for the design of explainable AI". In: *Expert systems with Applications* 213 (2023), p. 118888.

[8] Eva Cetinic and James She. "Understanding and creating art with AI: Review and outlook". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18.2 (2022), pp. 1–22.

[9] Zhanxiang Chai, Tianxin Nie, and Jan Becker. *Autonomous driving changes the future*. Springer, 2021.

[10] Shruthi Chari et al. "Explanation ontology: a model of explanations for user-centered AI". In: *International semantic web conference*. Springer. 2020, pp. 228–243.

[11] Gary Charness, Uri Gneezy, and Michael A Kuhn. “Experimental methods: Between-subject and within-subject design”. In: *Journal of economic behavior & organization* 81.1 (2012), pp. 1–8.

[12] Chi-Fen Emily Chen and Wei-Yuan Cheng. “The Use of a Computer-Based Writing Program: Facilitation or Frustration?.” In: *Online Submission* (2006).

[13] Zhe Chen et al. “Interpretable machine learning for building energy management: A state-of-the-art review”. In: *Advances in Applied Energy* 9 (2023), p. 100123.

[14] Michael Chromik and Andreas Butz. “Human-XAI interaction: a review and design principles for explanation user interfaces”. In: *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18*. Springer. 2021, pp. 619–640.

[15] Randa Diab-Bahman and Abrar Al-Enzi. “The impact of COVID-19 pandemic on conventional work settings”. In: *International Journal of Sociology and Social Policy* 40.9/10 (2020), pp. 909–927.

[16] George Dickie. “Defining art”. In: *American philosophical quarterly* 6.3 (1969), pp. 253–256.

[17] Linqian Ding and Di Zou. “Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence”. In: *Education and Information Technologies* (2024), pp. 1–53.

[18] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[19] Upol Ehsan et al. “Human-centered explainable AI (HCXAI): coming of age”. In: *Extended abstracts of the 2023 CHI conference on human factors in computing systems*. 2023, pp. 1–7.

[20] Ahmed Elgamal and Marian Mazzone. “Artists, artificial intelligence and machine-based creativity in playform”. In: *Artnodes: revista de arte, ciencia y tecnología* 26 (2020), p. 12.

[21] Emilio Ferrara. “Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies”. In: *Sci* 6.1 (2023), p. 3.

[22] Adam Fitriawijaya and Taysheng Jeng. “Integrating multimodal generative AI and blockchain for enhancing generative design in the early phase of architectural design process”. In: *Buildings* 14.8 (2024), p. 2533.

[23] Christopher Frye, Colin Rowat, and Ilya Feige. "Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability". In: *Advances in neural information processing systems* 33 (2020), pp. 1229–1239.

[24] Erick Guerra. "Planning for cars that drive themselves: Metropolitan planning organizations, regional transportation plans, and autonomous vehicles". In: *Journal of Planning Education and Research* 36.2 (2016), pp. 210–224.

[25] David Gunning et al. "DARPA's explainable AI (XAI) program: A retrospective". In: *Authorea Preprints* (2021).

[26] Khan Mohammad Habibullah et al. "Explainable AI: A Diverse Stakeholder Perspective". In: *2024 IEEE 32nd International Requirements Engineering Conference (RE)*. IEEE. 2024, pp. 494–495.

[27] Jamil Hamadneh, Szabolcs Duleba, and Domokos Esztergár-Kiss. "Stakeholder viewpoints analysis of the autonomous vehicle industry by using multi-actors multi-criteria analysis". In: *Transport Policy* 126 (2022), pp. 65–84.

[28] Robert R Hoffman et al. "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance". In: *Frontiers in Computer Science* 5 (2023), p. 1096257.

[29] Andreas Holzinger et al. "Explainable AI methods-a brief overview". In: *International workshop on extending explainable AI beyond deep models and classifiers*. Springer. 2020, pp. 13–38.

[30] Kasper Hornbæk and Antti Oulasvirta. "What is interaction?" In: *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2017, pp. 5040–5052.

[31] Mahipal Jadeja and Neelanshi Varia. "Perspectives for evaluating conversational AI". In: *arXiv preprint arXiv:1709.04734* (2017).

[32] Shilpi Jain et al. "Impact of irritation and negative emotions on the performance of voice assistants: Netting dissatisfied customers' perspectives". In: *International Journal of Information Management* 72 (2023), p. 102662.

[33] Tingting Jiang et al. "Human-AI interaction research agenda: A user-centered perspective". In: *Data and Information Management* 8.4 (2024), p. 100078.

[34] Ian-C Jung et al. "Overview of basic design recommendations for user-centered explanation interfaces for AI-based clinical decision support systems: A scoping review". In: *Digital Health* 11 (2025), p. 20552076241308298.

[35] Markus Kattnig et al. "Assessing trustworthy AI: Technical and legal perspectives of fairness in AI". In: *Computer Law & Security Review* 55 (2024), p. 106053. ISSN: 0267-3649. DOI: https://doi.org/10.1016/j.clsr.2024.106053. URL: https://www.sciencedirect.com/science/article/pii/S0267364924001195.

[36] Karl Kempf et al. "Issues in the Design of AI-Based Schedulers: A Workshop Report". In: *AI Magazine* 11.4 (Dec. 1990), p. 37. DOI: 10.1609/aimag.v11i4.868. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/868.

[37] Marzieh Khakifirooz, Michel Fathi, and Alexandre Dolgui. "Theory of AI-driven scheduling (TAIS): a service-oriented scheduling framework by integrating theory of constraints and AI". In: *International Journal of Production Research* (2024), pp. 1–35.

[38] Mohamed Khalifa and Mona Albadawy. "Using artificial intelligence in academic writing and research: An essential productivity tool". In: *Computer Methods and Programs in Biomedicine Update* (2024), p. 100145.

[39] Ruben Koole. "Research topics: Enhancing AI Assistants for Routine Tasks in the Workplace: The Role of Explainable AI in Technical and Human-Centred Way". Unpublished report. Mar. 2025.

[40] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. "To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis". In: *Organization science* 33.1 (2022), pp. 126–148.

[41] Siyu Catherine Li. "AI assistants in workplaces: implementation, workers, agency, and organizational policies a qualitative study". PhD thesis. The University of Texas at Austin, 2024.

[42] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai: A review of machine learning interpretability methods". In: *Entropy* 23.1 (2020), p. 18.

[43] Luca Longo et al. "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions". In: *Information Fusion* 106 (2024), p. 102301.

[44] Sherry Yancheng Ma. "Exploring ambiguity in generative AI images and its impact on collaborative design ideation". PhD thesis. Industrial Engineering and Innovation Sciences, Eindhoven University of . . ., 2024.

[45] Shiva Mayahi and Marko Vidrih. "The impact of generative ai on the future of visual content marketing". In: *arXiv preprint arXiv:2211.12660* (2022).

[46] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267 (2019), pp. 1–38.

[47] MinnaLearn. *Types of explainable AI*. URL: https://courses.minnalearn.com/en/courses/trustworthy-ai/preview/explainability/types-of-explainable-ai/.

[48] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. "A multidisciplinary survey and framework for design and evaluation of explainable AI systems". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11.3-4 (2021), pp. 1–45.

[49] Blake Montgomery. "Mother says AI chatbot led her son to kill himself in lawsuit against its maker". en. In: *The Guardian* (Oct. 2024). URL: https://www.theguardian.com/technology/2024/oct/23/character-ai-chatbot-sewell-setzer-death.

[50] Patricia Inoue Nakagawa et al. "Semantic description of explainable machine learning workflows for improving trust". In: *Applied Sciences* 11.22 (2021), p. 10804.

[51] Menaka Narayanan et al. "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation". In: *arXiv preprint arXiv:1802.00682* (2018).

[52] Meike Nauta et al. "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai". In: *ACM Computing Surveys* 55.13s (2023), pp. 1–42.

[53] Ljubica Nedelkoska and Glenda Quintini. *Automation, skills use and training*. Vol. 202. OECD Publishing Paris, 2018.

[54] Hung Truong Thanh Nguyen et al. "Evaluation of explainable artificial intelligence: Shap, lime, and cam". In: *Proceedings of the FPT AI Conference*. 2021, pp. 1–6.

[55] Thu Nguyen, Alessandro Canossa, and Jichen Zhu. "How Human-Centered Explainable AI Interface Are Designed and Evaluated: A Systematic Survey". In: *arXiv preprint arXiv:2403.14496* (2024).

[56] Ruth O'Neill and Alex MT Russell. "Grammarly: Help or hindrance? Academic learning advisors' perceptions of an online grammar checker". In: *Journal of Academic Language and Learning* 13.1 (2019), A88–A107.

[57] Daniel Omeiza et al. "Explanations in autonomous driving: A survey". In: *IEEE Transactions on Intelligent Transportation Systems* 23.8 (2021), pp. 10142–10162.

[58] Jeroen Overeem. “Measuring the effect of non-expert language on explanation satisfaction and user trust in Conversational XAI systems”. MA thesis. University of Twente, 2024.

[59] Gustavo Pinto et al. “Developer Experiences with a Contextualized AI Coding Assistant: Usability, Expectations, and Outcomes”. In: *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*. 2024, pp. 81–91.

[60] Gustavo Pinto et al. “Lessons from building stackspot ai: A contextualized ai coding assistant”. In: *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*. 2024, pp. 408–417.

[61] Alun Preece et al. “Stakeholders in explainable AI”. In: *arXiv preprint arXiv:1810.00184* (2018).

[62] Aleksandra Przegalinska et al. “Collaborative AI in the workplace: Enhancing organizational performance through resource-based and task-technology fit perspectives”. In: *International Journal of Information Management* 81 (2025), p. 102853.

[63] Arun Rai. “Explainable AI: From black box to glass box”. In: *Journal of the Academy of Marketing Science* 48 (2020), pp. 137–141.

[64] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “” Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[65] Erik Rice. “Design Thinking: A Process for Developing and Implementing Lasting District Reform. Knowledge Brief.” In: *Stanford Center for Opportunity Policy in Education* (2011).

[66] Ahmed M Salih et al. “A perspective on explainable artificial intelligence methods: SHAP and LIME”. In: *Advanced Intelligent Systems* (2024), p. 2400304.

[67] Wojciech Samek et al. *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. Springer Nature, 2019.

[68] A Saranya and R Subhashini. “A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends”. In: *Decision analytics journal* 7 (2023), p. 100230.

[69] Jessica Schrouff et al. “Best of both worlds: local and global explanations with human-understandable concepts”. In: *arXiv preprint arXiv:2106.08641* (2021).

[70] Agnia Sergeyuk et al. “Using AI-based coding assistants in practice: State of affairs, perceptions, and ways forward”. In: *Information and Software Technology* 178 (2025), p. 107610. ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2024.107610. URL: https://www.sciencedirect.com/science/article/pii/S0950584924002155.

[71] Mattia Setzu et al. “Glocalx-from local to global explanations of black box ai models”. In: *Artificial Intelligence* 294 (2021), p. 103457.

[72] Navdeep Singh and Daisy Adhikari. “AI in inventory management: Applications, Challenges, and opportunities”. In: *International Journal for Research in Applied Science and Engineering Technology* 11.11 (2023), pp. 2049–2053.

[73] Cleidson de Souza et al. “Lessons from Building CodeBuddy: A Contextualized AI Coding Assistant”. In: *arXiv preprint arXiv:2311.18450* (2023).

[74] Niko Tsakalakis et al. “A Typology of Explanations for Explainability-by-Design”. In: *ACM Journal on Responsible Computing* (2024).

[75] Giulia Vilone and Luca Longo. “Notions of explainability and evaluation approaches for explainable artificial intelligence”. In: *Information Fusion* 76 (2021), pp. 89–106.

[76] Warren J Von Eschenbach. “Transparency and the black box problem: Why we do not trust AI”. In: *Philosophy & Technology* 34.4 (2021), pp. 1607–1622.

[77] Yifei Wang. “A comparative analysis of model agnostic techniques for explainable artificial intelligence”. In: *Research Reports on Computer Science* (2024), pp. 25–33.

[78] Stefan Wellsandt, Karl Hribernik, and Klaus-Dieter Thoben. “Anatomy of a digital assistant”. In: *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems: IFIP WG 5.7 International Conference, APMS 2021, Nantes, France, September 5–9, 2021, Proceedings, Part IV*. Springer. 2021, pp. 321–330.

[79] Oyindamola Williams. “Towards human-centred explainable ai: A systematic literature review”. In: *Master’s Thesis* (2021).

[80] Chloe Wittenberg et al. “Labeling AI-generated content: promises, perils, and future directions”. In: (2024).

[81] Tianyu Wu et al. “A brief overview of ChatGPT: The history, status quo and potential future development”. In: *IEEE/CAA Journal of Automatica Sinica* 10.5 (2023), pp. 1122–1136.

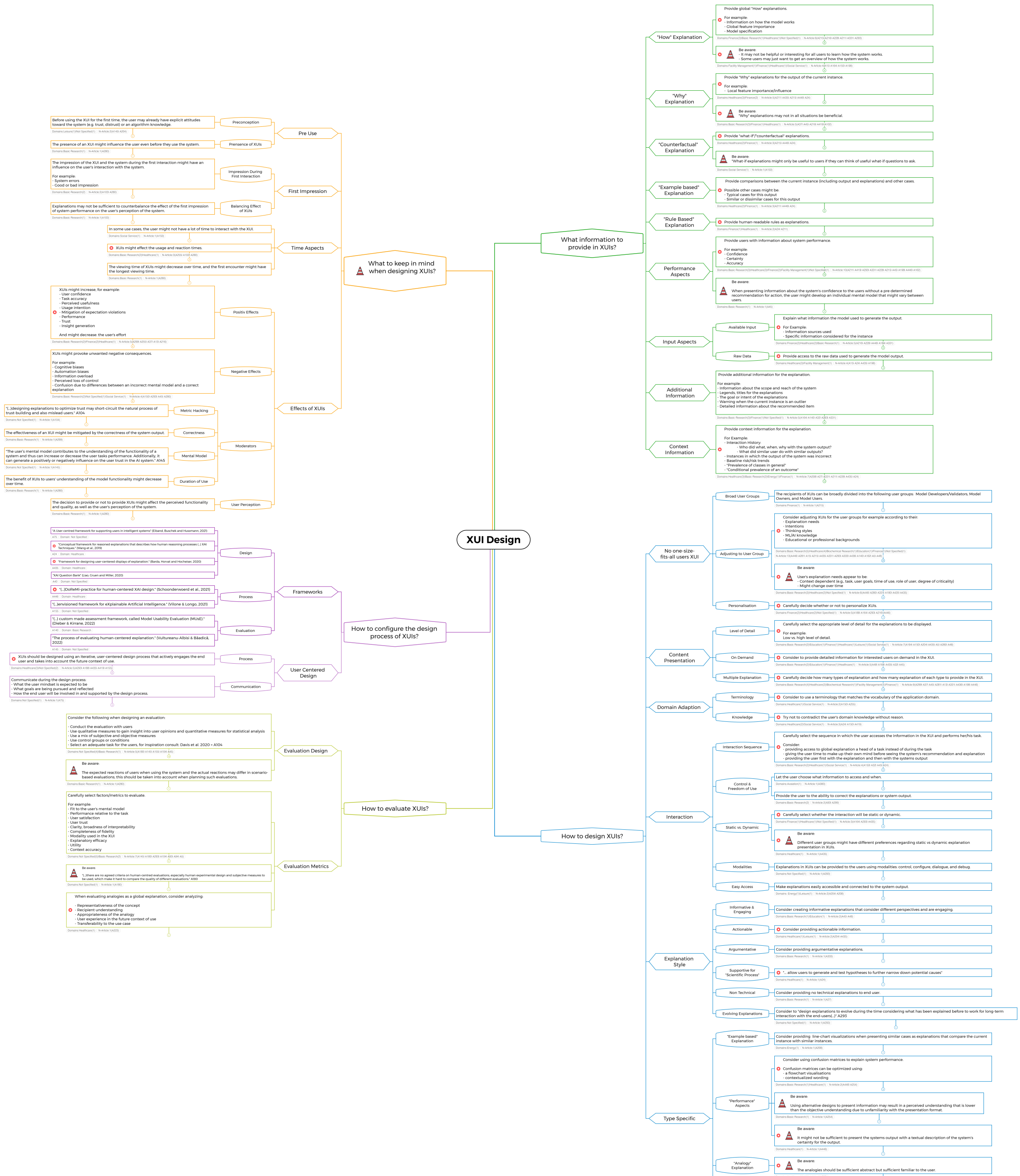
[82] Donika Xhani et al. “XMLPO: An Ontology for Explainable Machine Learning Pipeline”. In: *Formal Ontology in Information Systems*. IOS Press. 2024, pp. 193–206.

[83] Wei Xu. “Toward human-centered AI: a perspective from human-computer interaction”. In: *interactions* 26.4 (2019), pp. 42–46.

[84] Kai-Qing Zhou and Hatem Nabus. “The ethical implications of DALL-E: Opportunities and challenges”. In: *Mesopotamian Journal of Computer Science* 2023 (2023), pp. 16–21.

Appendix A

XUI Mind Map Unfolded



Appendix B

Questions Survey Ideation Phase

This is the full list of questions asked in the survey during the ideation phase.

B.1 General AI Experience

- Have you ever used an AI assistant? (e.g. ChatGPT, Copilot, Motion, DALL-E)
- Which of the following emotions have you felt when using AI assistants?

  (Anger, Confused, Curious, Disappointed, Excited, Frustration, Happy, Helped, Safe, Scared, Surprised)
- Can you elaborate why you felt that way?
- Please give your opinion on the following statements: (5 point Likert scale, Totally disagree to Totally agree)

  (The utilization of AI in various software is beneficial, I trust the AI decisions, I know how AI works, AI is a threat, AI gives me the desired outcome, I check the AI's answers elsewhere, I take the AI's answer serious, I use AI just for fun, I get frustrated when I don't understand how and/or why the AI came to its answer, I use AI for work)
- What makes you trust/distrust an AI?
- Do you have further remarks about your AI experience?

B.2 Explainable Artificial Intelligence

- Have you ever heard of Explainable User Interfaces (XUI) before this survey?

- What kind of explanation/aspects would you prefer in a XUI?

  *(**How** (Information on how the model works, on a global aspect), **Why** ('Why this specific prediction is given"), **Counterfactual** (or What-if explanation, how can we alter the output), **Example based** (provide typical examples for a certain output), **Rule based** (Provide (human-readable) rules as explanations), **Performance aspects** (Provide information about e.g. confidence, certainty, accuracy of the answer), **Input aspects** (Explain what information the model used to generate the output))*

- In which style would you prefer the explanation? (Low or High Level)

  (Technical Terms, Details, Amount of Explanation)

- XUI's can have different types of interaction with the user. What type of interaction would you like?

  *(**(Information) Transmission** (The XUI solely provides information), **Dialogue** (User can have a natural/interactive conversation with the XUI), **Control** (Support rapid convergence towards desired AI behaviour), **Experience** (Emphasize managing the expectations and preferences of users about the AI), **Optimal Behaviour** (Adjust human behaviour despite limitations of fully understanding the AI behaviour.), **Tool Use** (Facilitate learning from AI behaviour), **Embodied Action** (Establish a joint understanding with the AI for effective collaboration between user and XUI))*

- Would you prefer: textual, visual or a combination of both as explanation?

  (Textual, Visual, Combination of both)

- When should the XUI be available?

  (Always, After asked for, Only the first few times, Never)

- Below is a list of possible features of an XUI. Please rank them on what you think is important to include. From most important to least important.

  (Change level of technical explanation, Disable / Enable the XUI, Embodiment / Mascot, Help function / Legend, Customizability / Personalization, Different types of explanations, Alter input / send new prompt, Per input variable an explanation, Summary of explanation)

- Do you have further remarks about the implementation of XUI?

B.3 Usage of the Workplace App

- Do you use or have you used the Planon Workplace app?
- If yes, how often you use the app?

  (Daily, Every other day, Weekly, Monthly, Rarely, Not any more)
- Please give your opinion on the following statements:(5 point Likert scale, Totally disagree to Totally agree)

  (Booking a room is a lengthy process, I just walk to the room and see if its available, I care strongly about which room is booked, I book rooms near my desk, I book rooms near each other (when having multiple different meetings))
- Do you have further remarks on the Workplace App?

B.4 Human Assistance

- Do you let someone else book a (meeting) room for you?

  If Yes is answered the person goes to the following questions:
- Was there an instance when the room not suffice your needs when someone else booked the room?
- If yes, what were the factors/reasons which led to it not be sufficient?

  (Wrong Time, "Far" Location, Missing facilities, Broken facilities, Insufficient information given, Too little meeting time, It should have been a confidential/private meeting people passing by not allowed/prefer to see what is happening)
- Do you prefer letting someone else book rooms for you?
- Do you have further remarks about Human assistance?

Appendix C

Ideation Images

This appendix includes extra images of the ideation phase.

C.1 Settings

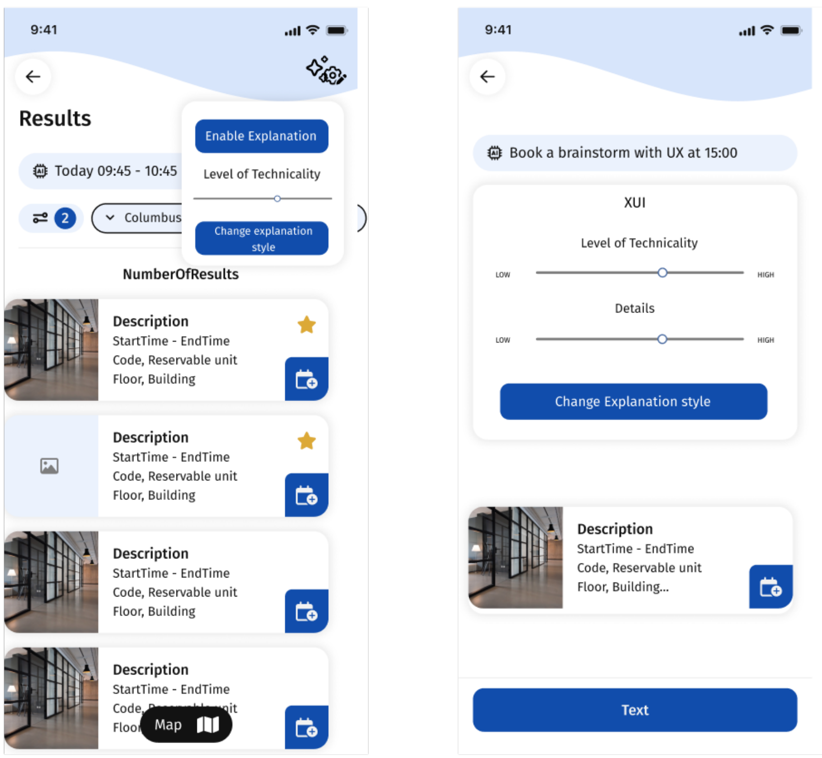


Figure C.1: Setting people could change in the explanations

C.2 Global explanation

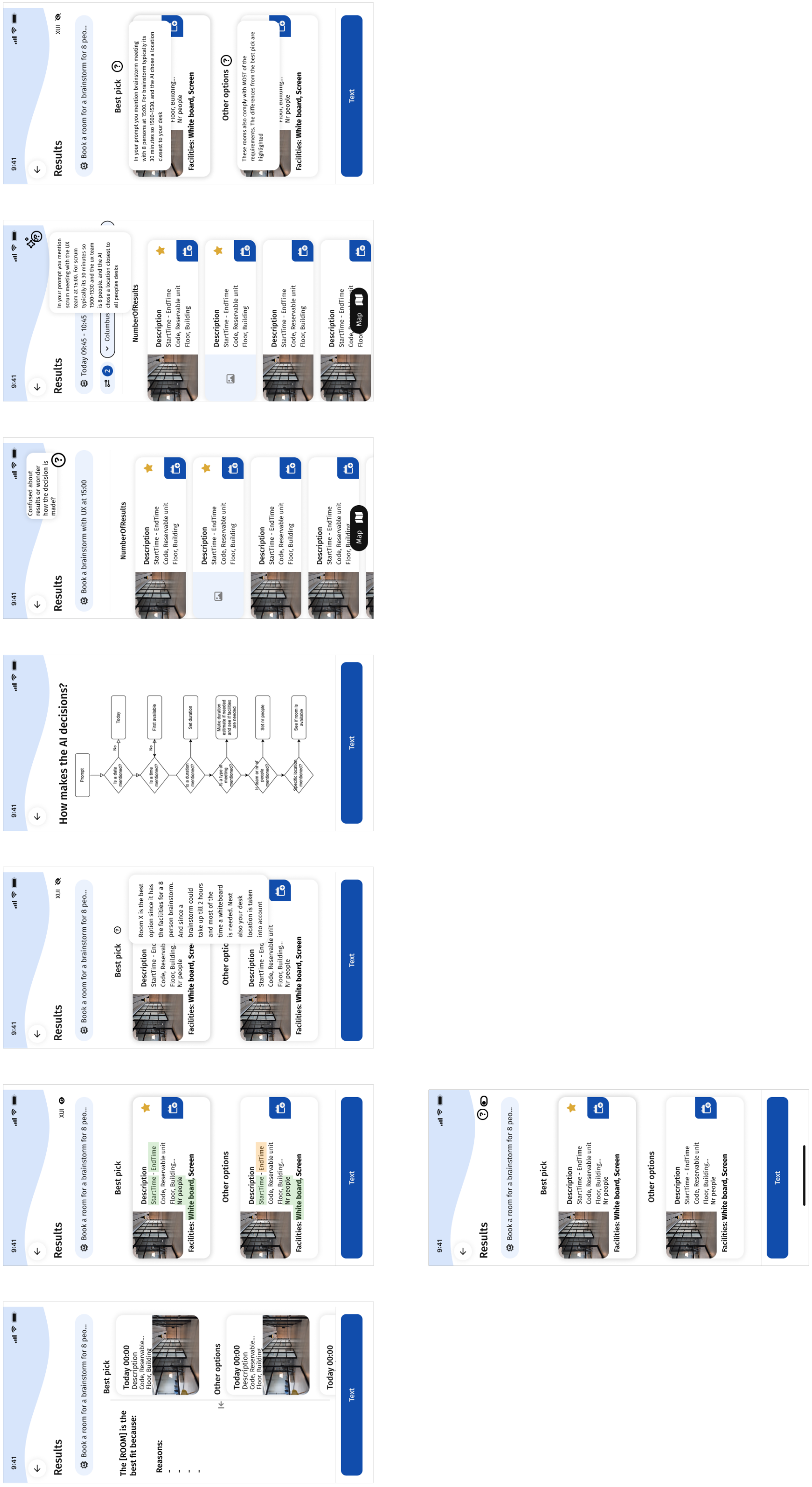


Figure C.2: Global explanations of the Ideation Phase

C.3 Local explanation

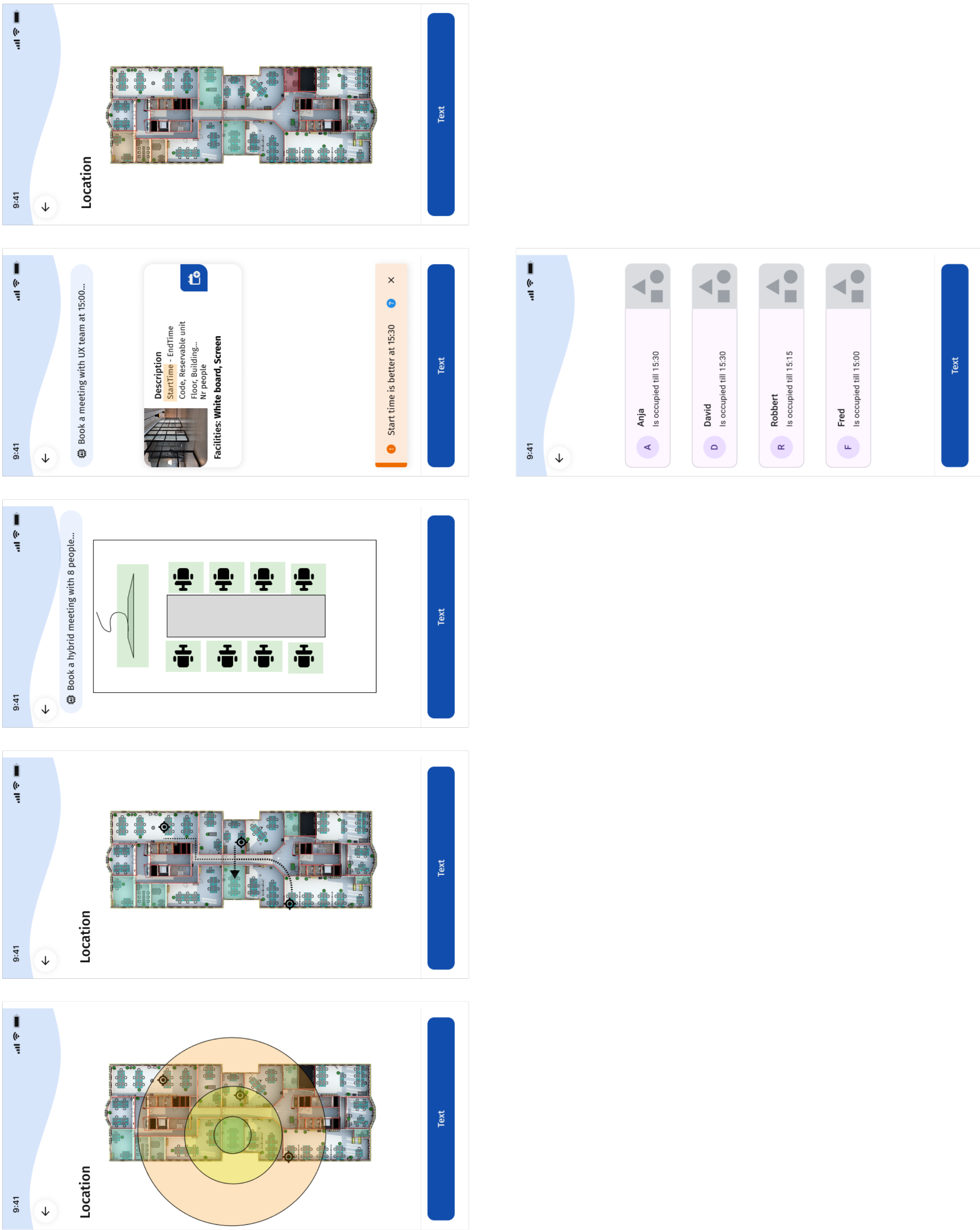


Figure C.3: Local explanations of the Ideation Phase

C.4 Input variable

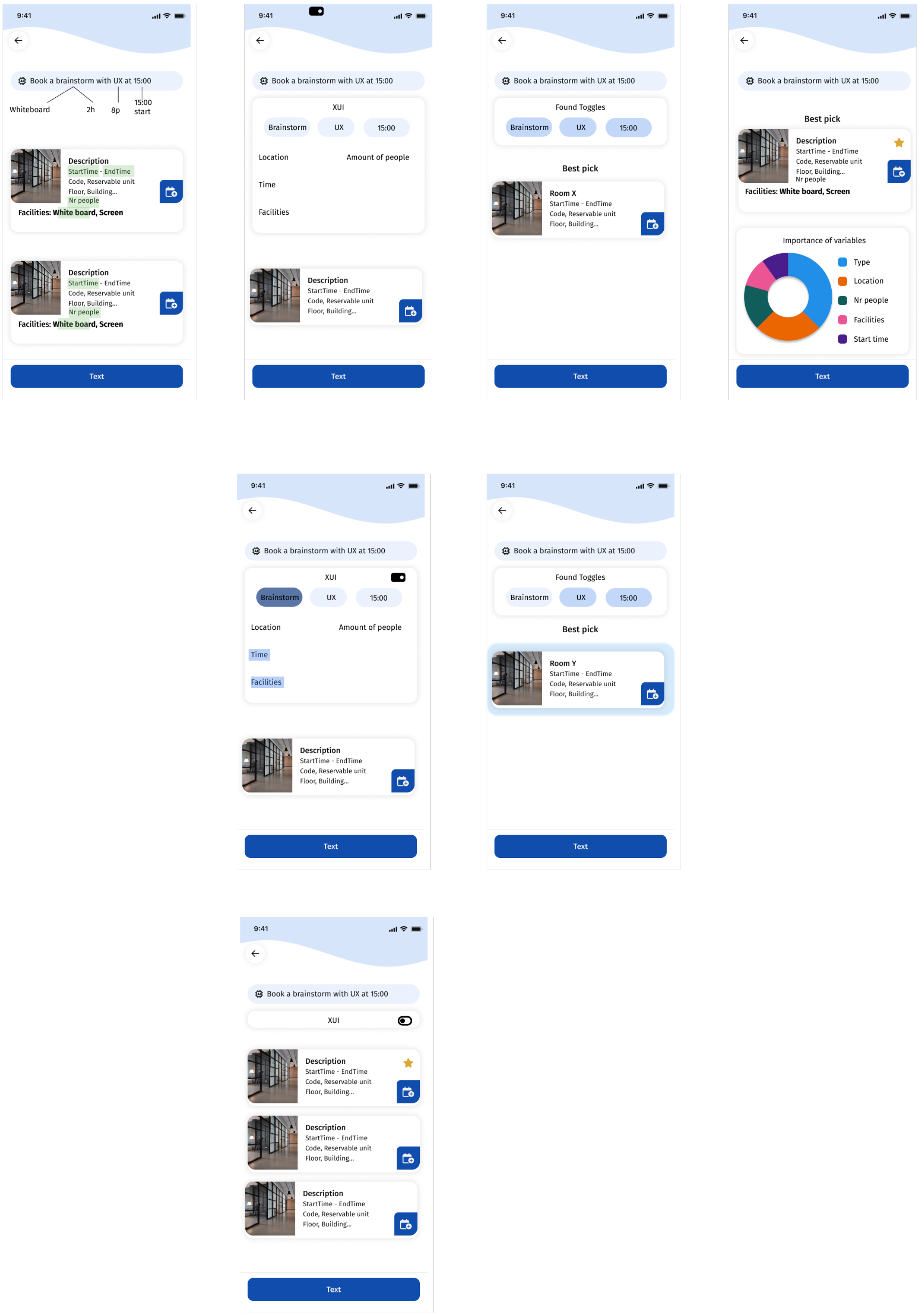


Figure C.4: Explanations based on input variable

C.5 Satisfaction

Figure C.5: Explanations showing satisfaction

C.6 Help function

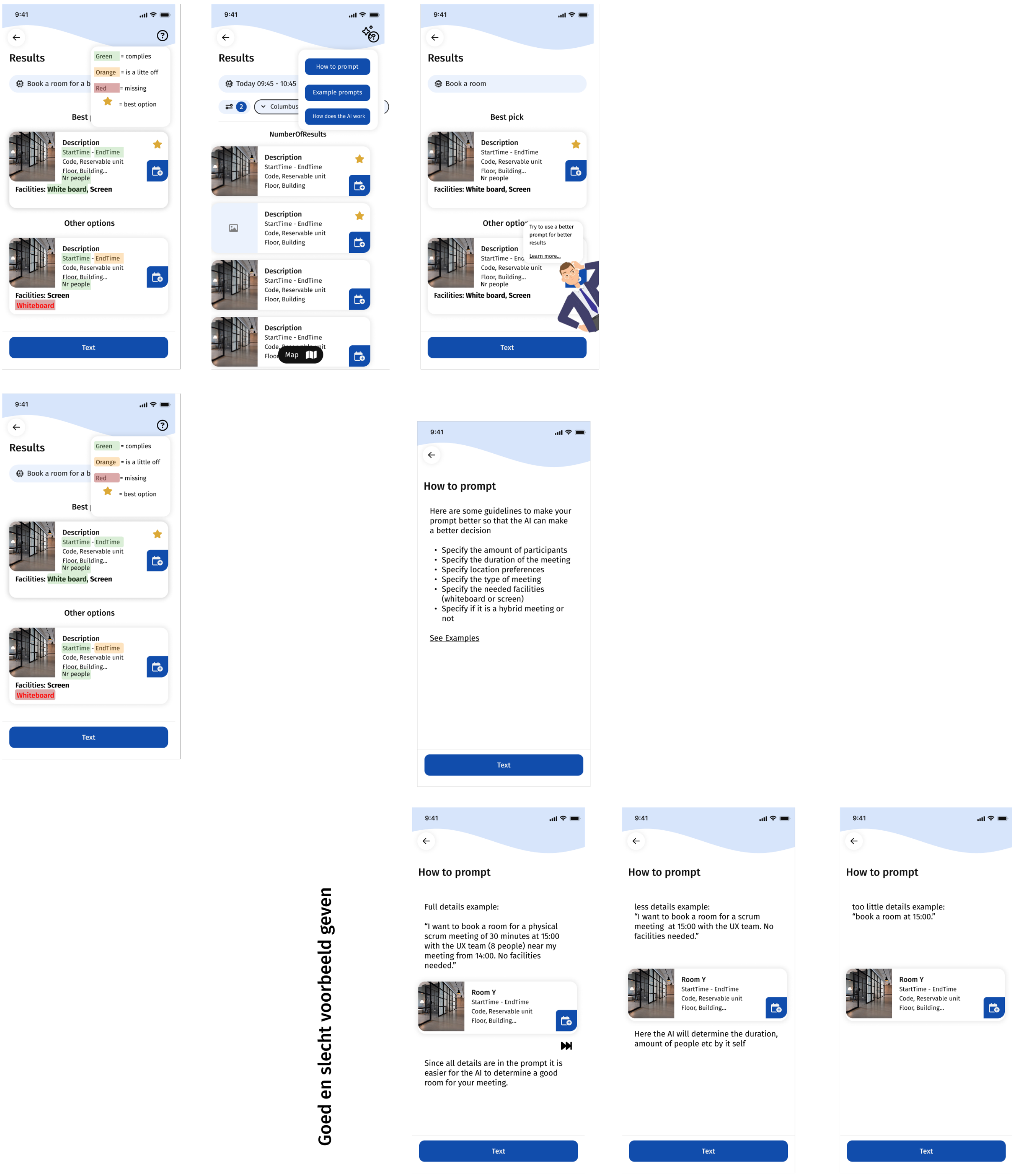


Figure C.6: Explanations with help function and/or legend

C.7 Dialogue Explanations

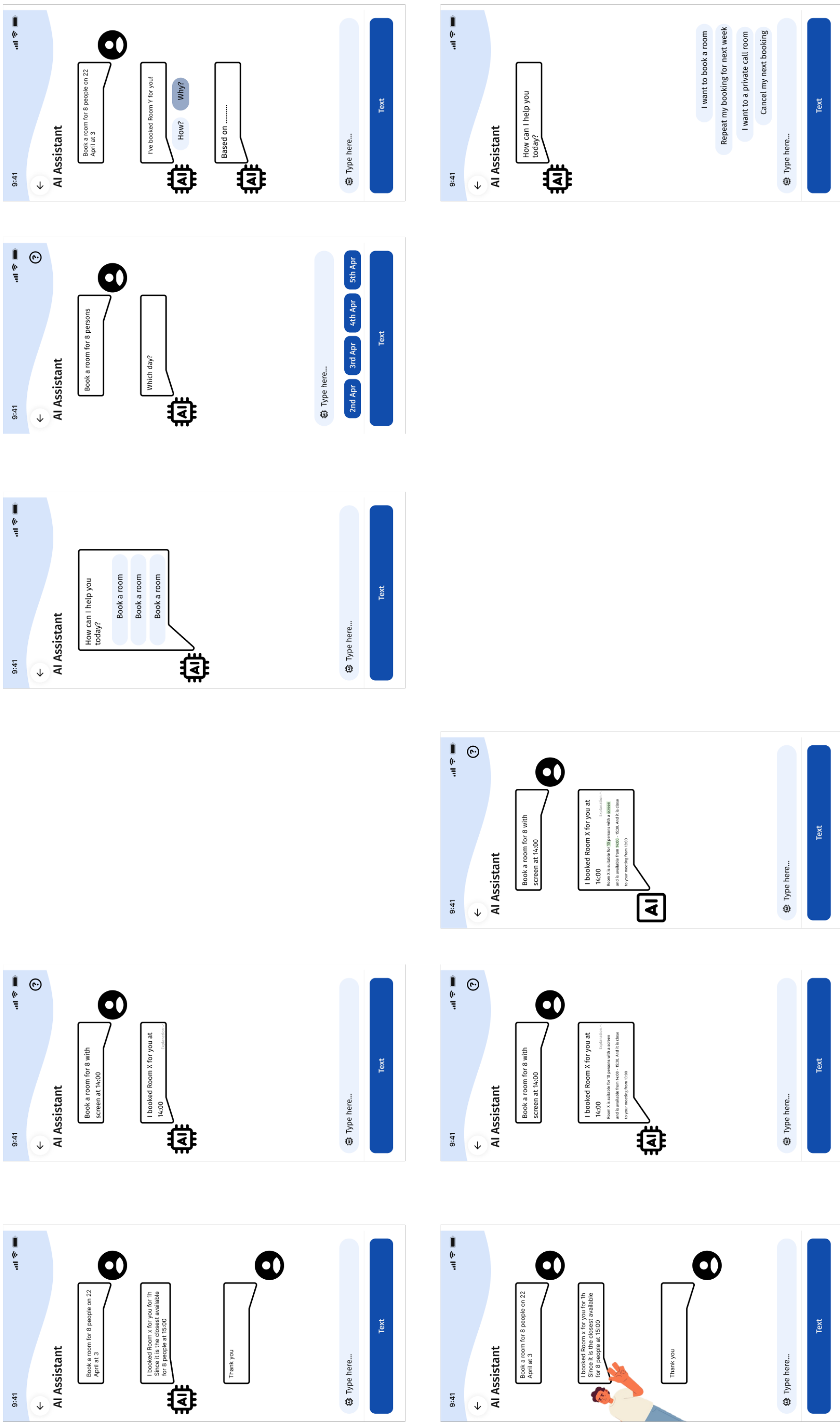


Figure C.7: Explanation with a Dialogue interaction

C.8 Homepage

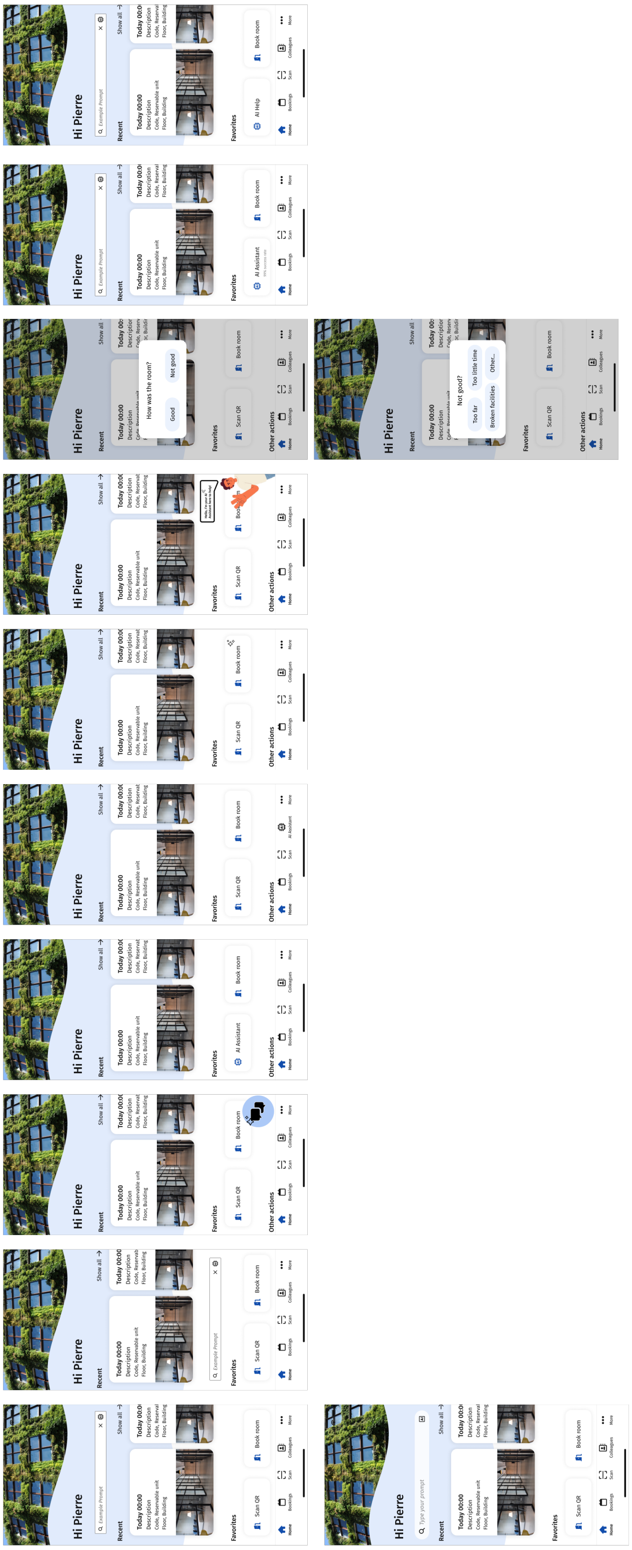


Figure C.8: Homepage variations

Appendix D

Prototype Images

This appendix includes extra images of the prototyping phase.

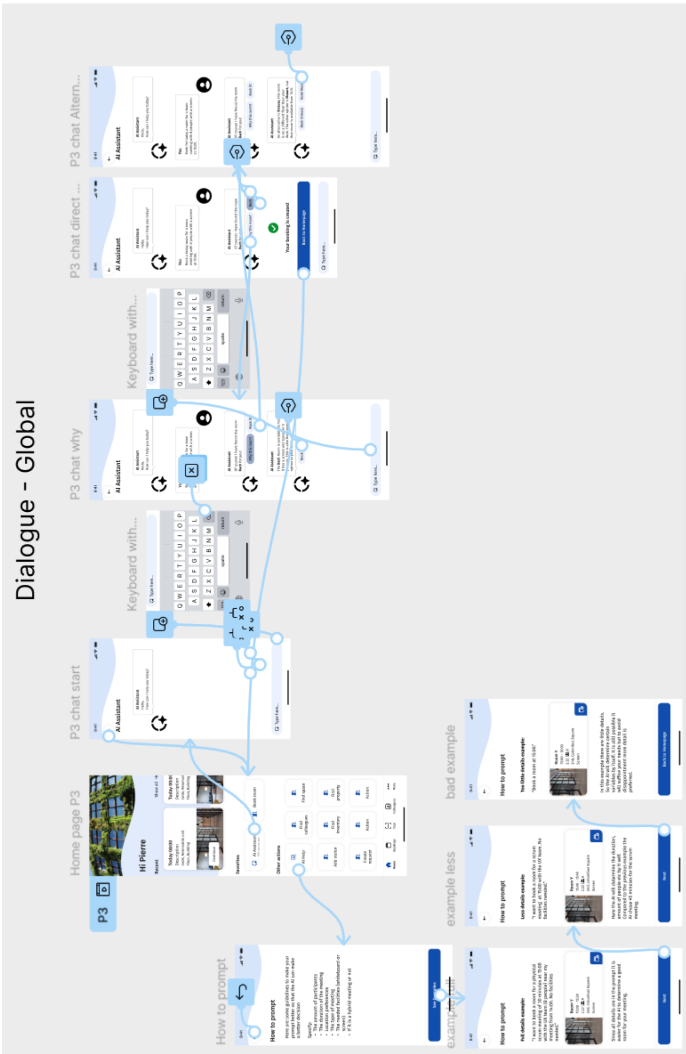


Figure D.1: Flow Dialogue Global

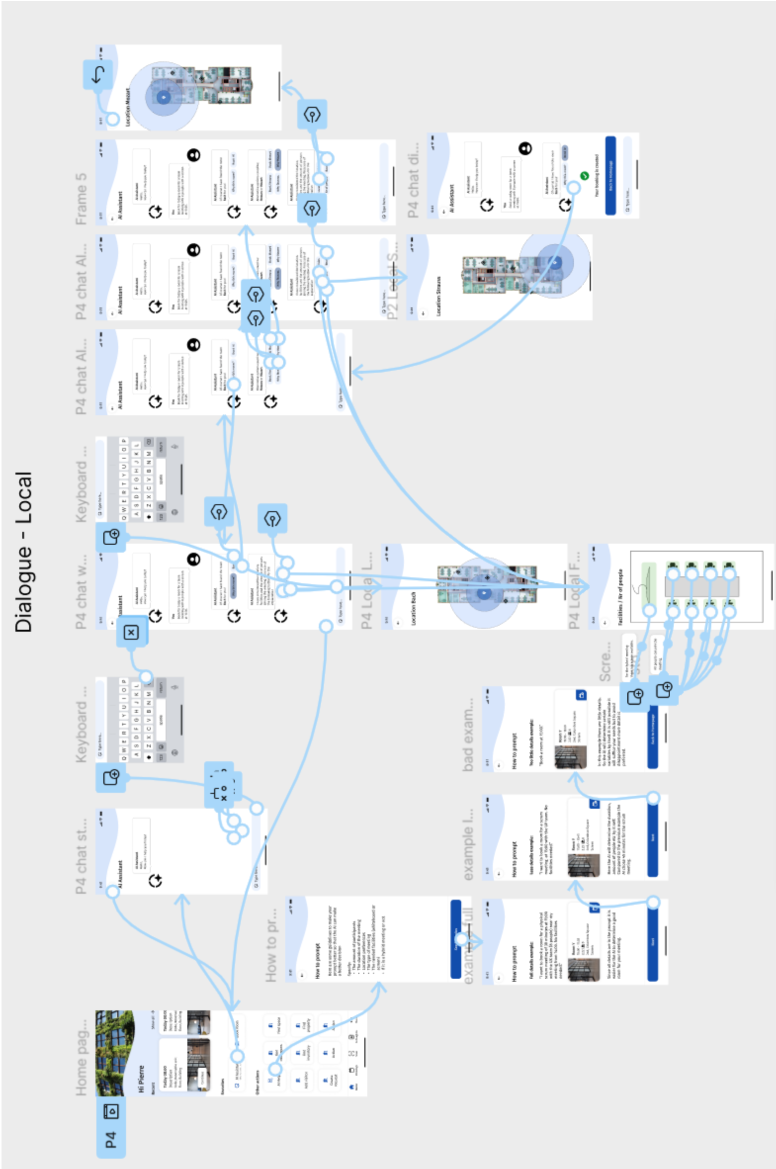


Figure D.2: Flow Dialogue Local

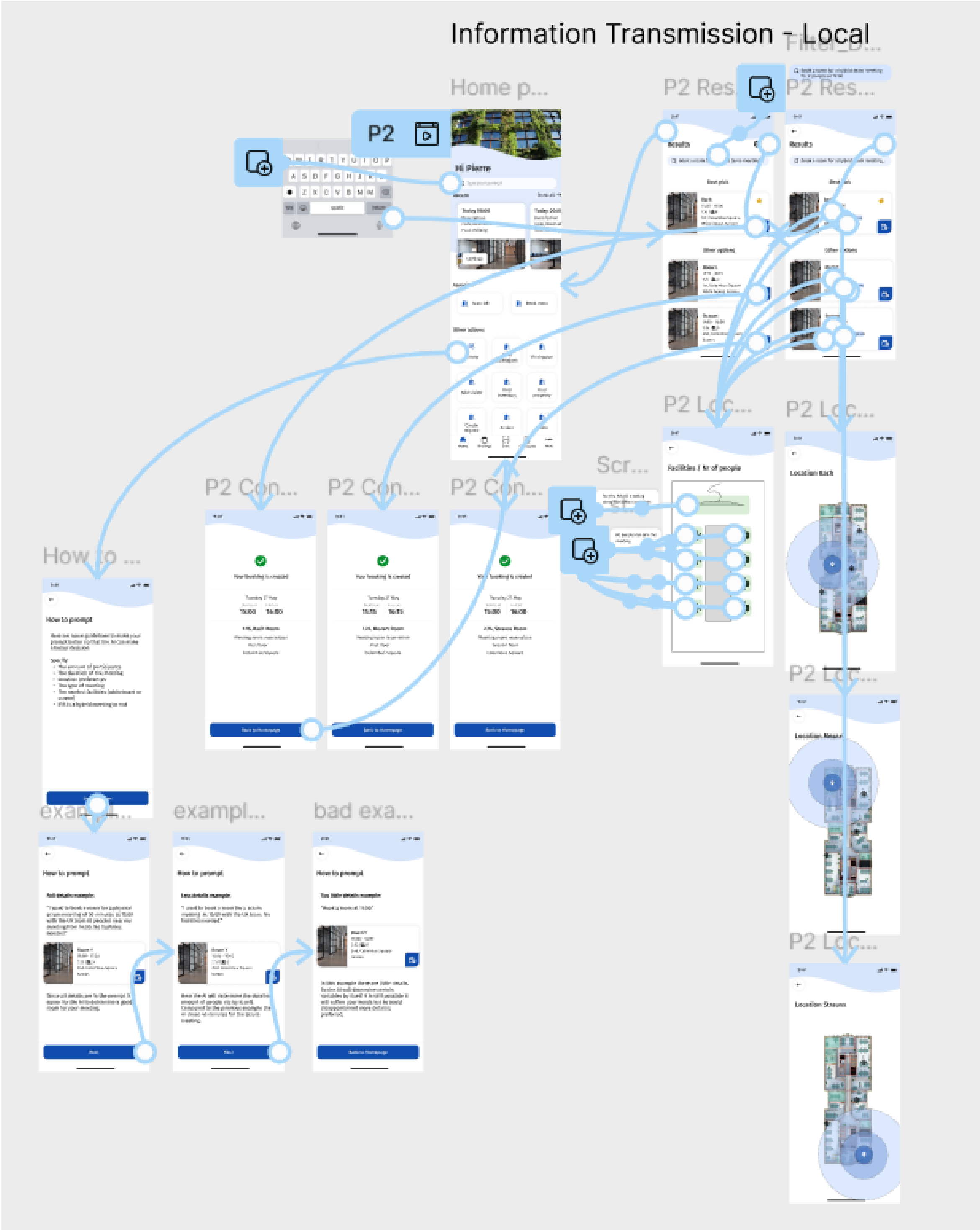


Figure D.3: Flow Transmission Local

Appendix E

Scenario Number generator Code

The number generator used to give participants a number is made in Google Apps Script. My program is based on a version of BPWEBS.com but altered to fit my goal. The script consists of 3 separate files.

E.1 Code.gs

```
    /*
# CREATED BY: BPWEBS.COM
# URL: https://www.bpwebs.com
# Altered By Ruben Koole
*/

function doGet() {
  return HtmlService.createTemplateFromFile('Index').evaluate()
  .setXFrameOptionsMode(HtmlService.XFrameOptionsMode.ALLOWALL);
}

//GET DATA FROM GOOGLE SHEET AND RETURN AS AN ARRAY
function getData() {
  var scriptProperties = PropertiesService.getScriptProperties();

  // Get the current value of i, default to 2 if not set
  var i = parseInt(scriptProperties.getProperty('rowCounter')) || 2;
  var spreadSheetId = "1LKGhuLtk0kVz5oyq48z3GQQOiujdMe4xYpQ6U9Y1bwk";
  //Connect to google sheet
  var dataRange = "Sheet1!A"+String(i); //Data range in sheet
```

```
  var range = Sheets.Spreadsheets.Values.get(spreadSheetId, dataRange);
  var values = range.values;
  scriptProperties.setProperty('rowCounter', i + 1);
  // scriptProperties.setProperty('rowCounter', 2);   // Reset list
  // var i = 2;
  Logger.log("Current rowCounter value: " + i);
  return values;

}

//INCLUDE JAVASCRIPT AND CSS FILES
//REF: https://developers.google.com/apps-script/guides/html/
best-practices#separate_html_css_and_javascript

function include(filename) {
  return HtmlService.createHtmlOutputFromFile(filename)
    .getContent();
}

//Ref: https://datatables.net/forums/discussion/comment/145428/#Comment_145428
//Ref: https://datatables.net/examples/styling/bootstrap4
```

E.2 Index.html

```
    <!DOCTYPE html>
<html>
<head>
 <base target="_top">
 <!--INCLUDE REQUIRED EXTERNAL JAVASCRIPT AND CSS LIBRARIES-->
 <script src="https://code.jquery.com/jquery-3.5.1.js"></script>
 <script src="https://cdn.datatables.net/1.10.23/js/
    jquery.dataTables.min.js"></script>
 <script src="https://cdn.datatables.net/1.10.23/js/
    dataTables.bootstrap4.min.js"></script>
 <link rel="stylesheet" type="text/css"
  href="https://cdnjs.cloudflare.com/ajax/libs/twitter-bootstrap/4.5.2/
        css/bootstrap.css">
 <link rel="stylesheet" type="text/css" href="https://cdn.datatables.net/1.10.23/
    css/dataTables.bootstrap4.min.css">
```

```
 <?!= include('JavaScript'); ?><!--INCLUDE JavaScript.html FILE-->
</head>
<body>
 <div class="container">
  <br>
  <div class="row">
   <table id="data-table" class="table table-striped table-sm
            table-hover table-bordered">
    <!-- TABLE DATA IS ADDED BY THE showData() JAVASCRIPT FUNCTION ABOVE -->
   </table>
  </div>
 </div>
</body>
</html>
```

E.3 JavaScript.html

```
    <script>
 /*
  *THIS FUNCTION CALL THE getData() FUNCTION IN THE Code.gs FILE,
  *AND PASS RETURNED DATA TO showData() FUNCTION
  */
  google.script.run.withSuccessHandler(showData).getData();

  //THIS FUNCTION GENERATE THE DATA TABLE FROM THE DATA ARRAY
  function showData(dataArray){
    $(document).ready(function(){
      $('#data-table').DataTable({
        data: dataArray,
        //CHANGE THE TABLE HEADINGS BELOW TO MATCH WITH YOUR
        SELECTED DATA RANGE
        columns: [
          {"title":"Number"}
        ]
      });
    });
  }
</script>
```

Appendix F

User test

This appendix will include the questions asked in the user test and the tasks that the participant has to perform.

F.1 Explanation Satisfaction Scale

The Explanation Satisfaction Scale consists of nine, 5-point Likert items. From ”I agree strongly” to ”I disagree strongly.”

- From the explanation, I understand how the [software, algorithm, tool] works.
- This explanation of how the [software, algorithm, tool] works is satisfying.
- This explanation of how the [software, algorithm, tool] works has sufficient detail.
- This explanation seems complete.
- This explanation shows me how accurate the [software, algorithm, tool] is.
- This explanation shows me how reliable the [software, algorithm, tool] is.
- This explanation tells me how to use the [software, algorithm, tool].
- This explanation is useful to my goals.
- This explanation helps me know when I should trust and not trust the [software, algorithm, tool].

F.2 Self-made Questions

- It was clear where to find the explanation (5-point Likert scale question)
- What feature (examples, legend, disable/enabling, highlights, etc) is most helpful, and why
- What feature are you missing?
- What feature would you consider redundant, and why?
- Do you have any suggestions to improve the features?
- Was the example page helpful?
- Which Prototype did you prefer?
- Can you elaborate why?

F.3 Tasks

There are two sets of tasks depending on which type of interaction the prototype has. Transmission or Dialogue.

Dialogue tasks:

- Find the AI Assistant.
- Book a Room (the prompt will automatically fill in when you tap on the on-screen keyboard.)
- Look for the explanation of the result.
- Look for alternative options. And their explanations.
- Search for the AI Help on the home page.

Transmission tasks:

- Find the AI Assistant.
- Book a Room (the prompt will automatically fill in when you tap on the on-screen keyboard.)
- Look for/turn on the explanation of the result.
- Explore the alternative options.
- Search for the AI Help on the home page.

Appendix G

Additional Analysis output

This appendix show some addition analysis that have been done during the evaluation phase.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.807	.798	9

Figure G.1: Reliability analysis output Dialogue Global prototype (rounded to 3 decimals)

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.931	.933	9

Figure G.2: Reliability analysis output Dialogue Local prototype (rounded to 3 decimals)

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.924	.926	9

Figure G.3: Reliability analysis output Transmission Global prototype (rounded to 3 decimals)

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.952	.952	9

Figure G.4: Reliability analysis output Transmission Local prototype (rounded to 3 decimals)