Design and Analysis of User Studies for the Evaluation of a Human-Robot Interaction Scenario

- User Studies -

Project Report

Daniela Pinto

Aalborg University
Department of Electronic Systems
Fredrik Bajers Vej 7B
DK-9000 Aalborg

The typesetting system used for the creation of this report is LaTeX, specifically the online software Overleaf, was used for typesetting and compiling of the raw formatted text. The figures in the report are developed with a variety of programs while some of the figures originate from other sources.

AALBORG UNIVERSITY
STUDENT REPORT

Department of Electronic Systems
Fredrik Bajers Vej 7
DK-9220 Aalborg Ø
http://es.aau.dk

Title:
Design and Analysis of User Studies for the Evaluation of a Human-Robot Interaction Scenario

Theme:
User Studies

Project Period:
Fall Semester 2021

Project Group:

Participants:
Daniela Rodriguez Pinto

Supervisor:
Dimitris Chrysostomou

Copies: 1

Numbers of pages: 62

Date of Completion:
January 6, 2022

Abstract:

Industry 4.0 is changing the manufacturing world at a fast speed. This introduces factories with new services, technologies and approaches. However, adapting to Industry 4.0 can be challenging. In order to ease the adaptation a Learning, Training, Assistance model is developed together with the implementation of Language-enabled Virtual Assistants.
In this project, a User Study is designed where participants are expected to complete a task while interacting with a VA, named Max, and an industrial robot. In this study, the participant is expected to be able to assembly a mockup-phone provided by AAU with the help of Max and the industrial robot. Following the interaction, the participants score and rate the system and interaction on a questionnaire. The system scored an average of 68.53 in System Usability Scale and averages of 2.41 and 1.74 in overall frustration for interaction with Max and the robot, respectively. Moreover, when voting for a preferred interaction 69.0% of the participants voted for Max, 31.0% voted to follow the instructions in a User Manual, and 0.0% voted to interact with an inexperienced worker. Additionally, the participants provided valued and helpful inputs for the current version of the system, as well as, inputs for future versions.

Contents

Preface

This project was performed by a engineering student on the Master of Robotics at The School of Information and Communication Technology at Aalborg University. This report is a documentation of the project performed on the fourth semester (ROB10). A full description of the span of the project can be found in the introduction. The objective of the project is the design and analysis of a User Study for a Human-Robot Interaction Scenario.

Aalborg University, January 6, 2022

Citation Guide

This report uses the Chicago method for citation purposes. The presented citations include the author's name, and, which year the material originated in. Furthermore, the report's bibliography is organised in a alphabetic order. Furthermore, there are two methods of referencing in this report. If the punctuation is followed by the reference, it spans for the whole section of text. If the punctuation succeeds the reference, the source applies only to the current sentence.

The following is a template source, which, presents information about citation standards, from the University of Pittsburgh. [University of Pittsburgh, 2018]

Abbreviations and Terminology

HRI - Human-Robot Interaction
HRC - Human-Robot Communication
VA - Virtual Assistant
PCB - Printed Circuit Board
SUS - System Usability Scale
Q&A - Question and Answering system

This report features numbers written in the Arabic numerals with decimal point display. E.g. the number one million will be written 1,000,000.00 or simply 1000000.00.

Daniela Pinto

Daniela Rodriguez Pinto
<dpinto16@student.aau.dk >

Acknowledgements

My first and most important acknowledgement is for my family. For my parents that supported me through every single day and every single doubt that I had during both my bachelors and masters. For my brother to put up with me every time I had a problem related with living far away from home to chase my dreams. For my uncles, aunts and grandmother for sending me care packages with food and more every time I missed a holiday or a birthday. For my family that I surprisingly found in Denmark, Carmen, Samuel and Santiago, and additional family Andreas and Riley. And, to finalise the family sector, a huge thank you to Matteo, my godson, for being my reason to smile when days are grey.

A big and bigger than anything acknowledgement for my childhood friends, Sofia, Patricia, Duarte, Afonso, Pedro and more for everything they ever did, not just on these 5 years of study, but also on our 9 plus years of friendship. Not forgetting my friendships from Denmark, Jan, Jacob, Rune, Guilherme, Elona, Vivian, Jennifer, and more, for being my rock, for all the adventures both in day and night time. And last, but never least, a special and lovely acknowledgement to my boyfriend Malte, for giving me extra support and happiness, for being by my side, and also, for buying me snacks when I go on a study marathon.

Deserving his own paragraph, a big thank you to my supervisor Dimitris. Not only for being my supervisor for this project but all the other projects. For understanding how hard, complex and complicated things can be and that studies and university are not everything that a student has going on in life.

Last, but not forgotten, a big thank you to Jan Dimon and Rasmus Pedersen, for supporting me with one of the hardest decisions that I had to do but turned out to be very helpful for me. Because now, I am finishing my studies in better health.

Chapter 1

Introduction

Industry 4.0 is the prevailing industrial revolution happening in the manufacturing world. It follows the idea of "Smart Manufacturing", by using cloud services in a factory, as well as, having most processes in a digital form. Thus, Industry 4.0 is currently providing factories with new processes, services and products. [Lase et al., 2014] [Frank et al., 2019]

Adapting to Industry 4.0 is a demanding task considering the growth in technological development and complexity of manufacturing supply chains. Factories are required to learn and develop new technologies and workers are require to acquire new skills. Such as social, technical and personal competences, which involve team work, process understanding and problem-solving skills, respectively. Furthermore, it is crucial that companies which decide to adapt to Industry 4.0, have to develop new strategies in order to assist workers while they are developing the mentioned skills. Moreover, adapting to the current industrial revolution, results in the integration and implementation of robots in the factory. Thus, it is also important to analyse the Human-robot Interaction (HRI). [Goodrich and Schultz, 2007] [Li et al., 2022] [Hamada, 2019]

The *Learning, Training, Assistance - Formats, Issues, Tools* (LTA-FIT) model was proposed with the focus on guiding factories and workers through a new adaptation. As the name states, it consists of 3 levels (Learning, Training and Assistance), where during these, the worker gathers the required knowledge. Furthermore, a previous developed Language-enabled Virtual Assistant (VA), follows the LTA-model while interacting with a user. [Rehe et al., 2020] [Li et al., 2022]

For this project, a User Studies Design is designed and analysed in order to evaluate the interaction between the VA and the robot with the human, in a manufacturing scenario whilst following the LTA-model.

Chapter 2

Background Analysis

This chapter analyses the current industrial revolution in section 2.1 (Industry 4.0), and a proposed model called Learning, Training, Assistance - Formats, Issues, Tools in order to help the employees adapt to the changes Industry 4.0 brings to factories. Section 2.2 (Human-robot Interaction) analyses the different types of HRI together with the implementation of HRI context in factories. In section 2.3 (Current Approaches), an analysis is performed regarding the current implementations of VAs in a workplace. Lastly, section 2.4 (Natural Language-enabled Virtual Assistance - Max) describes the system that is used for this project.

2.1 Industry 4.0

Throughout the years, industry has been developing and growing, which led into technological leaps which, nowadays, are referred to as industrial revolutions. The first industrial revolution has, as a central element, the field of mechanization; the second, welcomed the use of electrical energy; the third introduced the world of digitalisation; and, lastly, the fourth, so called Industry 4.0, establishes the use and adaptation to smart manufacturing. (see figure 2.1) [Lase et al., 2014]

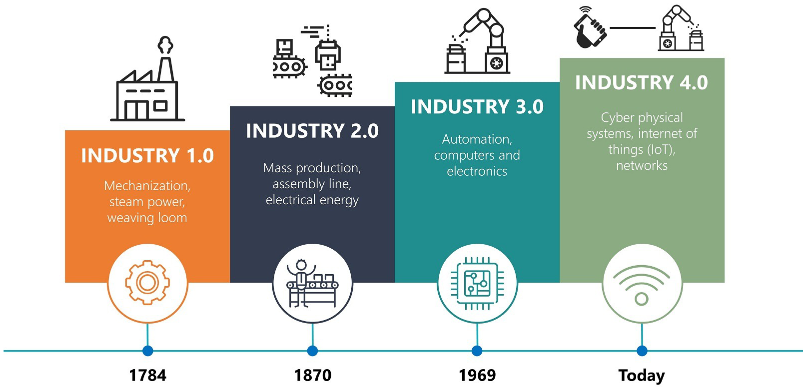


Figure 2.1: Timeline of the four industrial revolutions. [Presentation Point, 2018]

Industry 4.0 was firstly introduced over a decade ago in 2011 in Germany by a federation consisting of multiple private companies and universities as a way to improve the manufacturing industry. It suggested a completely new approach to production and manufacturing. One of the main concepts of Industry 4.0 is the implementation of smart manufacturing. Smart manufacturing enables the usage of cloud services in a factory, as well as, the technology to gather real time data and analyse it. Furthermore, it provides flexible and customised lines of production which leads to an increase of productivity and quality. Thus, adapting to Industry 4.0 requires a socio-technical evolution on the workers. [Lase et al., 2014] [Frank et al., 2019]

2.1.1 Learning, Training, Assistance - Formats, Issues, Tools

As mentioned above, Industry 4.0 is the current industrial revolution, thus, more and more companies are adapting to this new concept. Adapting to Industry 4.0 can be considered a challenging task. With an introduction to new services, products and processes, an addition of complexity to manufacturing supply chains is observed. Thus, workers are expected to learn and adapt to a new ideology, technology, services, products and processes, which requires the presence of an expert. [Hamada, 2019] [Rehe et al., 2020]

In order to address the challenges previously stated, [Rehe et al., 2020] proposed the Learning, Training, Assistance - Formats, Issues, Tools (LTA-FIT) model. This model offers a solution to the adaptation to Industry 4.0 by identifying and specifying multiple and different requirements that qualify workers and give the opportunity for companies to move towards a digital transformation.

LTA-FIT consists of three different levels:

- **Level 1 - Learning:** During this level, the goal is to teach the workers and company the required knowledge. Here, general and broad issues are discussed and cover. Moreover, *Learning* aims to further the existing knowledge; [Rehe et al., 2020]
- **Level 2 - Training:** This level focuses on building upon and solidifying the gained knowledge from *Learning*. Here, new skills can be improved though practical workshops; [Rehe et al., 2020]
- **Level 3 - Assistance:** The last level consists mainly of support throughout the practical work by addressing specific requirements. [Rehe et al., 2020]

Furthermore, this model was implemented and tested throughout 80 LTA-FIT workshops and, at least, 750 participants. It covered multiple topics such as, Human-Robot Collaboration, Augmented and Virtual Reality, IT Security, among others. All-embracing, the results of the experience were positive. [Rehe et al., 2020]

2.2 Human-robot Interaction

Human-robot interaction (HRI) has been a deeply and widely researched field of study in robotics. It is mainly focused on designing, understanding and evaluation of robotics systems for use with/by humans. By definition, *interaction* is an occasion where there is communication or reaction between two or more parties [Cambridge Dictionary, 2021]. Thus, a HRI can be described as the communication between human and robot. [Goodrich and Schultz, 2007]

Depending on the proximity between human and robots, the interaction can be divided into two general categories: *Remote interaction* and *Proximate interaction*. Remote interaction occurs when the two parties are not located in the same area (e.g. robots in a workcell, in another room, etc.). On the other hand, Proximate interaction occurs when the two parties share location (for example, collaborative robots, service robots, etc.). Dividing HRI into these categories helps distinguish between different applications depending on social interaction, mobility or physical manipulation/collaboration. [Goodrich and Schultz, 2007]

Back in 2002, United Nations (U.N.) performed a robotics survey where robotics was grouped into three major categories: *Industrial Robotics*, *Personal Service Robotics*, and *Professional Service Robotics* (see figure 2.2). [Bartneck and Forlizzi, 2004]

Figure 2.2: Example of a robot for each robotics category. **Left:** *Industrial robot* - Two industrial robotic arms performing soldering [Nichols, 2020]; **Middle:** *Personal Service robot* - Vacuum and moping robot [Brains, 2020]; **Right:** *Professional Service robot* - Mine robot assisting in mining [Baggaley, 2017].

Industrial robots were the first to be commercialised and represent the category with the most widespread distribution, it is computer-controlled, it manipulates the surrounding physical environment and it is set in an industrial environment; Personal Service robots are developed with the goal to assist people in a domestic environment, e.g., vacuum cleaners, wheelchairs, robot assistants for physically impaired people, etc.. This category consists of robots that interact with humans that do not need any training or skills to operate the robot; Lastly, Professional Service robots are the most recent category of robotics. Although, similar to industrial robots, professional service robots navigate and manipulate the surroundings, these are used to assist humans in their professional goals, besides industrial related. This can be, for example, robots that navigate through mines, robots that help with the clean-up of nuclear waste, robots that assist in hospitals and surgeries, etc.. [Thrun, 2011]

2.2.1 HRI in Industry

As previously mentioned, industrial robots were the first category of robotics to be commercialised, with the earliest registered in 1954, with a robot called "Unimate" developed by George Devol. Since then, robots have been providing multiple advantages for the industry and have become a well-established manufacturing sector. These are used mainly for repetitive tasks, such as, assembly, sorting, stacking, etc.. With this, robots are programmed to perform these tasks continuously and accurately in environments that can be harmful or difficult for humans. Furthermore, once the robot is needed for a different task it can be re-programmed. Thus, it is safe to say that with robots it is possible to achieve an increase in precision, quality of production, flexibility and production. [Li et al., 2020] [Heyer, 2010]

As above-mentioned, one of the multiple advantages of industrial robots, is the flexibility and the ability to re-program it depending on the task. However, this demands that the worker operating the robot needs a specialized knowledge on the subject.

2.3 Current Approaches

When sharing a workspace, humans tend to communicate over a shared task either through verbal or non-verbal cues. The same is expected when sharing a workspace with a robot. Thus, for a HRI to take place in a workspace, it is needed to determine the appropriate Human-Robot Communication (HRC), which, can also be through verbal or non-verbal cues. [Hjorth and Chrysostomou, 2022]

For a non-verbal HRC, the interaction consists of visual methods to read body language, such as, skeleton tracking or gesture recognition, while, for a verbal HRC, an implementation of speech recognition and voice control is made. In addition, a non-verbal approach can be considered as a complementary addition to a solely verbal approach. However, while a non-verbal HRC is more researched in the area of social robots, a verbal HRC is mainly used in manufacturing applications. [Hjorth and Chrysostomou, 2022]

"Work-based learning" is a term that has been becoming more and more popular over the years. It is defined as any type of learning that occurs in the workplace, and it has been proven to be more effective than a traditional approach, such as, classroom learning. Additionally, factories that have been implementing work-based learning, have been proven to be more effective. [Lester and Costley, 2010] [Li et al., 2022]

2.3.1 Implementing Virtual Assistants in a Workplace

Voice-enabled interaction has gained popularity throughout the years, as such interaction generate emotions associated with surprise and curiosity in humans. Thus, VAs have become widely used in multiple environments. From a domestic environment, e.g. asking for traffic information or turning on the house lights, to a work environment, e.g. asking to open and start a video conference, commercialised VAs[1] have been developed and focus on providing a natural and interactive interaction. [Burbach et al., 2019] [Li et al., 2022] [Cóndor-Herrera et al., 2020]

Bot-X

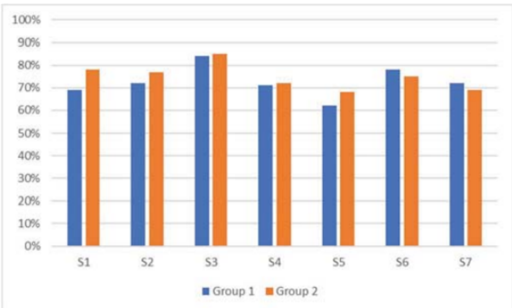
Together, the department of Materials and Production of AAU and the school of Informatics of Leicester University have proposed Bot-X, an AI-based VA that handles a variety of complex services in the manufacturing industry. Bot-X is powered by three different layers: library layer, which includes all the core libraries that provide the needed functions; dialogue layer, where the key information extraction is implemented together with a knowledge repository that holds the labeled sample dataset; and application layer, where several application interfaces are designed. Furthermore, Bot-X also consists of extra functions, such as, a calendar checker and a human emotion detection. [Li and Yang, 2021]

[1] e.g. Alexa (`https://www.amazon.com/b?node=21576558011`); Siri (`https://developers.google.com/assistant`)

In addition, Bot-X is tested in three different case studies. For the first use case, by having a natural conversation with the user, the VA extracts the sales information needed, e.g. costumer name and product name and quantity. This information is typically given to the system manually, however, the user can communicate with Bot-X instead. The second use case consists of Bot-X checking the available inventory in order to verify if the user's order is able to be produced. This scenario is usually performed by having a warehouse worker check the factory's inventory, however, Bot-X can provide a high reliability in performing such task. The last use case is where the production control is tested. Lastly, for future work, the authors have interviews planned in order to collect some feedback from industrial partners. [Li and Yang, 2021]

Chatbot

[Casillo et al., 2020] propose an implementation of a chatbot in Industry 4.0 facilities to help employees by training them during learning phases [Casillo et al., 2020]. The chatbot has integrated a live support, a customised and contextualised training, and an efficient training. Furthermore, the authors performed user studies which "covers seven specific sections of a company's training process". With a total of 30 newly hired employees as participants which are divided into 2 different groups (of 15 participants each): Group 1 performed the user studies while interacting with the chatbot, and Group 2 performed the experiment with the traditional training approach. Figure 2.3 presents a graph of the outcome of the study. As it is possible to see, in 2 sections, Group 1 had a better results, where in the remaining sections the results between the groups are comparable. [Casillo et al., 2020]



Section	Percentage		
	Negative	Neutral	Positive
A	8,33%	8,33%	83,33%
B	6,67%	0,00%	93,33%
C	10,00%	5,00%	85,00%
D	18,33%	8,33%	73,33%
E	11,67%	0,00%	88,33%

Figure 2.3: Right: Summary of the results obtained; **Left:** Table of the score of the questionnaire in percentage. (A - Usability, B - Recommendation, C - Presentation, D - Dialogue, and E - Future Development). [Casillo et al., 2020]

After the user studies, the participants were asked to answer a questionnaire. The questionnaire is divided into five sections: Usability, Recommendation, Presentation, Dialogue and Future Development. Each section consists of two assertions were the participants can rate the experiment using a five-point Likert Scale (I totally disagree | I disagree | Undecided | I agree | I totally agree). Figure 2.3 presents a table with the results of the questionnaire. [Casillo et al., 2020]

From the above mentioned, the authors concluded that the section that needed more focus on future work of the chatbot was the dialogue. However, overall the authors were presented with good results regarding their proposed solution. [Casillo et al., 2020]

2.4 Natural Language-enabled Virtual Assistance - Max

Industry 4.0 is changing the manufacturing world by providing companies with new technologies, services, products and more. Every year more and more factories adapt to Industry 4.0. Consequently, factories require the presence of an expert in order to provide an introduction of the new technologies to the workers. However, at multiple situations, the expert might not be available nor present, which can result in a delay or abandoning of the factory adapting to Industry 4.0. [Hamada, 2019] [Li et al., 2022]

In order to address this concern, [Li et al., 2022] proposed an integration of a Natural Language-enabled Virtual Assistant, named Max, which follows the LTA model for manufacturing tasks. [Li et al., 2022]

2.4.1 System Description

As it can be seen in figure 2.4, the proposed system consists of a web application, involving two actors (*Max Client* and *Max Server*), and a robotic platform. The sequence of numbers in figure 2.4 represent the workflow: 1) spoken commands; 2) and 13) display user's and Max's dialogue messages together with System Status on the web interface; 3) sends the transcripts; 4) grounds user's commands to the robot control command; 5) matches the knowledge graph[2] together with the robot services; 6) generates responses; 7) selects conversation strategies and responses templates; 8) sends responses; 9) generates verbal responses; 10) invokes API to control robots; 11) sends commands to the robotic platform; and 12) reports the robot's status. [Li et al., 2022]

[2] "(...) the knowledge graph acts as a knowledge repository used by the VA to retrieve the answers for the operators' queries." [Li et al., 2022]

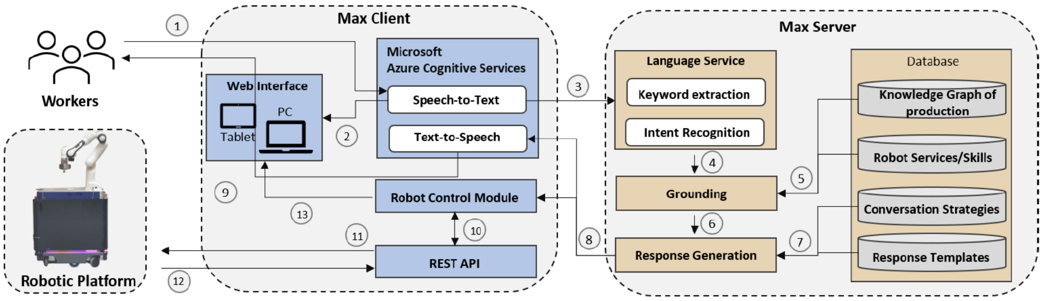


Figure 2.4: An overview of the proposed system architecture. [Li et al., 2022]

Max Client

The Max Client can be separated into three different sub-modules. The *Web interface* (see figure 2.5) is divided into different information panels, the Dialog, System Status and Robot Service Panels which display, the user's command and Max's response, real time system information and a list of the robot's services, respectively. The *Cognitive Voice Service*, which consists of two Microsoft cognitive services, speech-to-text (STT) and text-to-speech (TTS). And lastly the *Robot Control Module*, which is used to control the manipulator. [Li et al., 2022]

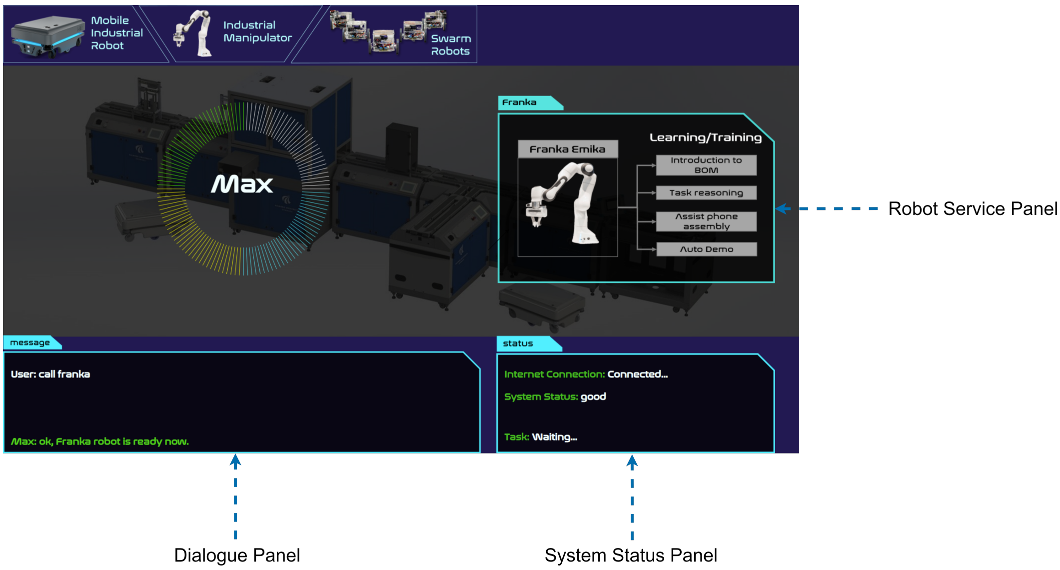


Figure 2.5: Web application Max. Based on [Li et al., 2022].

Max Server

Max Server relies on three functionalities: 1) *Language Service* in order to acknowledge the user's intent. Furthermore, the Max's language service supports two methods, a rule-based keyword extraction and the BERT [Devlin and an Kenton Lee abd Kristina Toutanova, 2019] model. 2) *Natural Language Grounding* with the purpose of grounding each user command and determine the relationship with a robot operation action. After extracting the tuple in language service, natural language grounding matches the keyword on the command with the correct answer regarding terminology or the correct robot action. 3) *Response Generator*, which supports Lexical Semantic Strategy and General Diversion Strategy. The implementation of these strategies open the possibility for Max to provide diverse responses to the user, e.g., by having Max follow a *"Don't repeat yourself rule"* from lexical semantic strategy. Moreover, this can provide the user with a more humanized verbal HRC. [Li et al., 2022]

Robotic Platform

Max is originally implemented into the latest version of the Little Helper (LH), which contains a Mobile Industrial Robot (MiR) and a Franka Emika collaborative robotic manipulator. However, for the execution of the project, only the Franka Emika is used as the robotic platform. [Li et al., 2022]

Chapter 3

Hypothesis

Gathering the information previously stated in chapter 2 (Background Analysis), the user studies focuses on three different aspects: collecting suggestions from users for improvement of the system and HRI; analysing the presence of a robot in the scenario; identification of major usability issues encountered. Thus, the following Hypothesis is created:

> *Adopting a Natural Language-enabled Virtual Assistant for LTA approach, provides an easier implementation of Industry 4.0 in factories.*

Chapter 4

Scenario

Industry 4.0 is currently being implemented by multiple factories at a rapid speed. One of the challenges this brings is the need for an expert who may not be available or present. [Hamada, 2019] Thus, for this project, a scenario where it simulates the named expert being unavailable created. Moreover, this scenario follows the LTA-model.

The chosen scenario consists of assembling and naming the components of mockup-phones, seen in figure 4.1 (bottom and top cover, fuses and PCB), provided by the Smart Production Laboratory at AAU[1]. In order to complete the task, the user has the opportunity to favour between interacting with the system and reading the User Manual[2] for a limited time.

Figure 4.1: Components of the mock-up phone. [Li et al., 2022]

[1]www.smartproduction.aau.dk

[2]The User Manual is attached to appendix A.

Furthermore, consequently, the hypothesis previously stated got divided into three sub-modules:

- Users with a background in IT/Engineering prefer the usage of the system for the experiment;
- Users tend to be more comfortable interacting with the system by the end of the experiment;
- Users that do not interact somewhat often or often with a VA are inclined to choose the User Manual.

In addition, as previously mentioned in section 2.4 (Natural Language-enabled Virtual Assistance - Max), Max uses keywords extraction in order to ground the user's commands. The keywords for each component are defined by either the shape and/or colour, mainly. These are listed below:

- **Bottom Cover:** Bottom Cover | Housing/house | Black;
- **PCB:** PCB | Green;
- **Fuses:** Fuse(s) | Small Thing(s) | Cylinder(s) | Tube(s);[3]
- **Top Cover:** Top Cover | Blue.

This chapter analyses the chosen scenario together with a general idea of interaction between the user and the VA through the different LTA-model levels.

4.1 Learning

As mentioned in Section 2.1.1 (Learning, Training, Assistance - Formats, Issues, Tools), the Learning scenario of the LTA-model provides the user with the needed primary knowledge for the task. In this scenario, this consists of knowledge on the facility, system and phone assembly process. For the learning scenario, this is done mainly through a Question Answering system (Q&A) and, primarily, by interacting with Max.

For this project, Max server works solely through a keyword extraction system, without the need for conversation strategy. Thus, the system has a knowledge graph integrated which allows for the system to answer to the user's commands. This knowledge graph can provide information such as the components of the phone, relationship between components and attributes of the phone. Figure 4.2 provides a graphical view of the content in the knowledge graph with the relationship needs (between the PCB and fuses), on (between PCB and both covers) and pair (between covers). [Li et al., 2022]

[3] Cylindfer(s) and tube(s) are both added to the system after the pilot test (Appendix B).

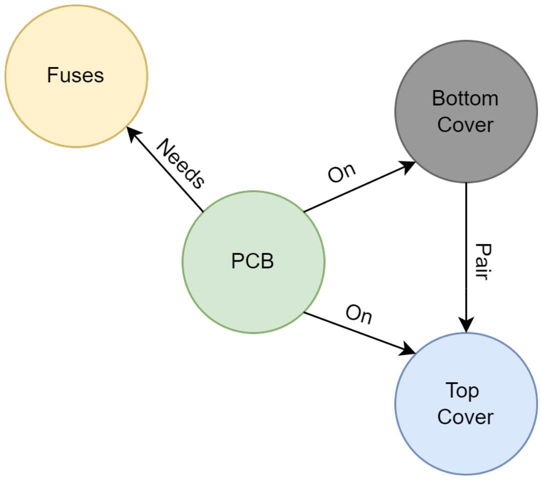


Figure 4.2: Graphical representation of the knowledge graph. Based on [Li et al., 2022].

Moreover, Max should be able to understand and answer the user's commands, gather the keyword(s) and answer the user with the needed information. Examples can be seen in figure 4.3, where the user gives two different commands to Max, and by the keyword(s) extraction of each, Max is able to retrieve the user's command.

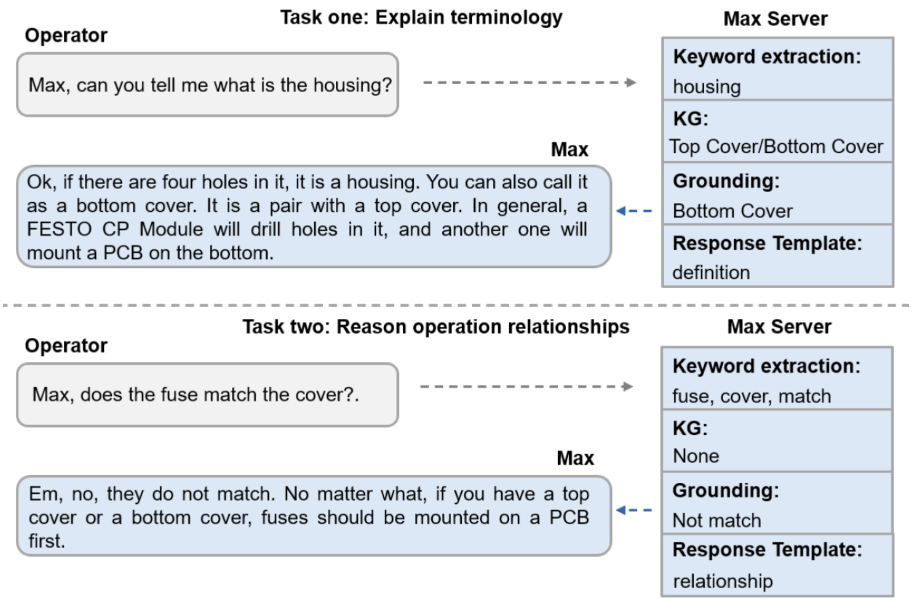


Figure 4.3: Two examples of an interaction between the user (operator) and Max: terminology of the term *housing* and relationship between components *fuse* and *cover*. [Li et al., 2022]

4.2 Training

During the training level, the user is expected to solidify and apply the knowledge gained in Learning. In order to achieve this, the system provides live demonstration of the phone assembly followed by a follow-up Q&A or hands-on learning approach. This leads into the user interacting with Max and the robot simultaneously. Therefore, the VA should be able to trigger a robot action based on the user's commands.

An example of a task can be seen in figure 4.4, where the user asks Max for a demonstration on phone assembly. Similar to the learning level, the system extracts the keyword(s) and Max provides an answer for the user and, in addition, it sends a command to the Robot Control Module, resulting in an action from the robot. [Li et al., 2022]

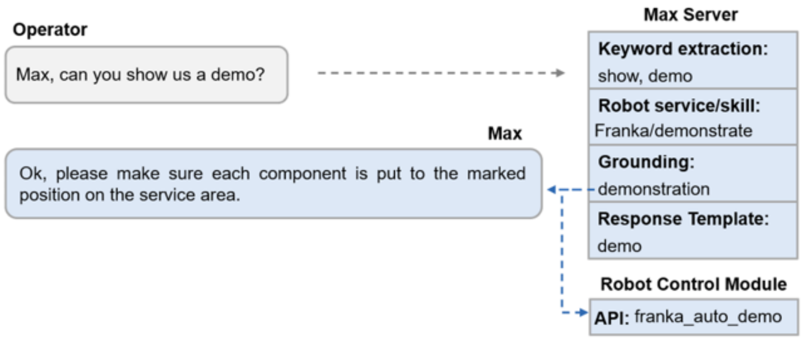


Figure 4.4: Example of an interaction between the user (operator) and Max: demonstration of the phone assembly. Based on [Li et al., 2022].

Moreover, the user can ask the system to bring specific phone components. This follows the same procedure as the demonstration of phone assembly. However, when the system brings a component, it waits for an input from the user. Figure 4.5 shows a sequence of photos of an interaction with the system. First, the user gives a command ("*Bring* me the *top cover*") to Max, once the command is grounded, the robot starts the action, in this case grasping; afterwards, the robot brings the component to the user, for the robot to release the component, the user must push down the gripper once. Afterwards, the robot goes back to the default position.

Figure 4.5: Sequence of interaction with the system. From left to right: User gives the voice command to Max; the robot starts the grasping of the top cover; user pushes the gripper; the robot releases the component.

4.3 Assistance

The last level of the model consists of assisting the user. During the assistance level, the system can provide a guided assembly where it is possible to deliver the components to the user in the correct order.

Figure 4.6 provides an example of the procedure during this level. By asking Max for a guided or assisted phone assembly, the system hands each one of the components, in the correct order, to the user, and guides through certain details (e.g. orientation of the PCB). [Li et al., 2022]

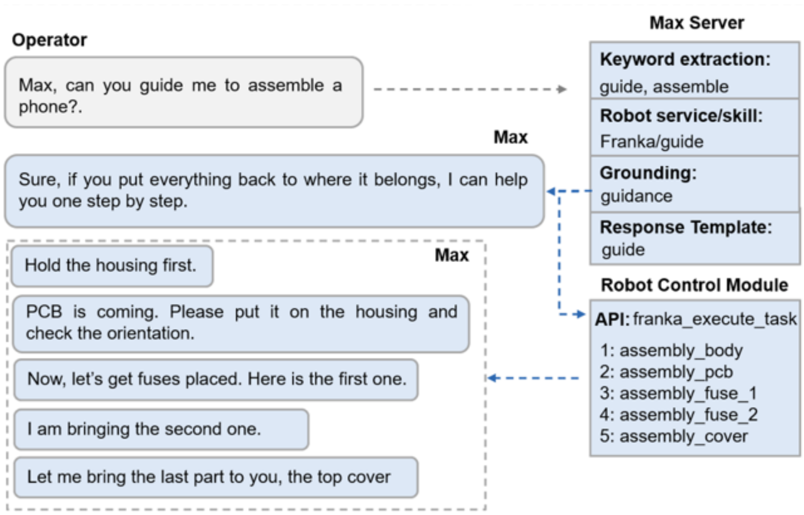


Figure 4.6: Example of an interaction between the user (operator) and Max: Guided phone assembly. Based on [Li et al., 2022].

Chapter 5

User Studies Design

With a chosen hypothesis and scenario, it is possible to design the studies in order to answer the hypothesis. This chapter analyses different possibilities of designs and concludes on a design for the studies. Section 5.1 (Study Context) explores and analyses the three most used study contexts, moreover, it also analyses the differences between within- and between-participants. In section 5.2 (Questionnaire) an analysis and description of the different sections of the questionnaire is presented.

5.1 Study Context

One hypothesis can be tested in a number of different contexts. The most popular are: laboratory, field and internet testing. Each have their own advantages and disadvantages. [Hoffman and Zhao, 2020]

As the name suggests, internet testing is conducted online. This study context has become more and more popular over the years with the growth of the internet, and with Amazon using it for the Mechanical Turk and Prolific [Chandler et al., 2013]. For this study, the participants can do everything online without the need to go to a physical place. In the context of HRI, this study context can include showing images or videos of either a human interacting with a robot, or simply, images and videos of how the robot is supposed to react. An advantage this context provides is that gathering more test subjects is simplified and faster with a lower financial cost. Moreover, by testing online, it is possible to open a broader range of test subjects with different age and, study and work backgrounds. However, a disadvantage can be that by having the test subjects watching a video of the robot, it lowers the external validity[1] of the study. Furthermore, online participants might not be as engaged in the interaction as these would be face-to-face. Lastly, by having an online testing, it can limit the researcher to analyse the reaction and interaction done with the robot. Thus, internet testing is not the chosen for this project. [Hoffman and Zhao, 2020]

[1] External Validity defines how generalised the outcome of a study can be, e.g. places, populations, situations, etc. [Hoffman and Zhao, 2020].

[Chandler et al., 2013]

Laboratory and field studies are both performed in person. One of the main differences between these two is that laboratory studies are performed in a controlled environment, while field studies are implemented in a daily situation. Choosing field studies brings an understanding of how the users that would interact with the system on a daily basis, react to the system. Thus, field studies contributes for external and ecological validity[2]. However, conducting a field study is considered to be harder and more difficult as it adds complexity to confounding variables[3]. Nonetheless, laboratory studies allow a more controlled experiment, by allowing more control over the variables. [Hoffman and Zhao, 2020]

Taking the above-mentioned into account, the experiment for this project would ideally follow the laboratory testing. Thus, it is possible to have a more controlled result and have more controlled over external factors, e.g. noise. However, a private room without any external noise and factors was not available. Thus, the experiment implements a combination of both laboratory and field study. Lastly, it is recommended that a "stage script"[4] is written for the experiment. This is done to assure that most experiments are as similar as possible to each other from the researcher part [Hoffman and Zhao, 2020].

5.1.1 Within- and Between-Participants

When conducting a study involving human-participants, it is possible to choose either within- or between-participants design. Within-participants is where each participant goes through all the conditions of the experiment and at the end the difference between participants is analysed. On the other hand, between-participants is when each participant is assigned to a group and each group goes through a stage of the experiment, thus, at the end it is possible to analyse the differences between the groups. [Hoffman and Zhao, 2020]

An advantage of choosing a within-participants design is the fact that fewer participants are needed. Moreover, choosing within-participants, the individual differences between the participants do not affect the experiment. On the other hand, this design can leave the participants with multiple order effects. Some examples are: [Hoffman and Zhao, 2020]

- *Habituation:* the participants might get bored or uninterested during the experiment;
- *Learning:* the participants might perform better at a task when repeated;
- *Familiarity:* the participants get familiarised with the task being performed.

[2]Ecological validity refers to whether the results of a study can be generalised to situations outside of a laboratory. [Hoffman and Zhao, 2020].

[3]Confounding variables are defined as external factors that can affect the outcome of the study and are not part of the theory and/or hypothesis. [Hoffman and Zhao, 2020].

[4]The stage script can be found in Appendix C.

Nonetheless, it is also possible to include a mixed-participants design, which includes features of both within- and between-participant designs. In mixed-participants design, the researcher has the opportunity to analyse not only the differences between different groups of participants, but also the changes in each individual participant. [American Psychological Association, 2021]

Taking into consideration the above-mentioned, within-participants is usually the go-to design choice as it is possible to analyse the differences of each participant in an individual matter, however, this design has its flaws. When these cannot be avoided or taken into consideration, the between-participants design is a better choice. Lastly, it is also possible to go for a combination of both by choosing the mixed-participants design. For this project, the mixed-participants design is chosen.

5.2 Questionnaire

In order to measure the HRI of the experiment, a questionnaire is used, which the participant has to fill out after testing [Rueben et al., 2020]. The questionnaire is divided into six sections: 1) Participant's information; 2) Personal Preferences; 3) Evaluation of the system's usability; 4) Evaluation of the perceived comfort while working on the task; 5) Evaluation of engaging interaction; 6) Personalised affirmations.

5.2.1 Sections 1 and 2 - Participant's Information and Personal Preferences

The first two sections of the questionnaire are designated for the participants to introduce themselves. The first section consists of the following questions:

- *Gender:* female | male | other;
- *Age:* 20-25 | 26-30 | 31-35 | over 36;
- *Background:* IT | Business | Entertainment | Science | Other; [5]
- *Previous interaction with an industrial robot:* Never | At least once | I have programmed and/or work with industrial robots;
- *Interaction with a VA:* Never | Rarely | Somewhat often | Often.

Thus, it is possible to distribute the participants into different groups.

The second section asks for the participant's preference between either reading the user manual or interacting with the system, for completion of the task, and a justification for doing so. Moreover, in this section, the participant also answers the same questions in the scenario of repeating the experiment. Hereby, the researcher gathers a better understanding of the participant's reason of choice.

[5] More options are listened in the questionnaire, however, only the selected ones are listed here.

5.2.2 Section 3 - Evaluation of the System's Usability

For this section of the questionnaire, the SUS questionnaire is used. The System Usability Scale (SUS) is a ten-item questionnaire used to measure usability. Participants have a five-point Likert scale (from strongly disagree to strongly agree) where they can rate the experiment. Later, these are used to measure the system's satisfactions, effectiveness and efficiency. [Brooke, 1995] [Weiss et al., 2016]

For this project, the SUS questionnaire is divided into three different categories: Max (table 5.1); Robot (table 5.2); Max and Robot/system (table 5.3).

Table 5.1: Section 3 of the questionnaire designated to Max. The latter subscale of NASA TLX is distributed through items 5, 6 and 7.

1. I think that I would like to use Max frequently.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
2. I found Max's screen interface unnecessarily complex.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
3. I thought that Max couldn't understand what I was saying.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
4. I felt very confident using Max.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
5. I was stressed during my interaction with Max.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
6. I felt insecure during my interaction with Max.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
7. I was annoyed during my interaction with Max.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree

Table 5.2: Section 3 of the questionnaire designated to the robot. The latter subscale of NASA TLX is distributed through items 3, 4 and 5.

1. I think that I would need the support of an expert to be able to use the robot.				
◯ Strongly disagree	◯ Disagree	◯ Neither agree or disagree	◯ Agree	◯ Strongly Agree
2. I need to learn a lot of things before I could get going with the robot.				
◯ Strongly disagree	◯ Disagree	◯ Neither agree or disagree	◯ Agree	◯ Strongly Agree
3. I was stressed during my interaction with the robot.				
◯ Strongly disagree	◯ Disagree	◯ Neither agree or disagree	◯ Agree	◯ Strongly Agree
4. I felt insecure during my interaction with the robot.				
◯ Strongly disagree	◯ Disagree	◯ Neither agree or disagree	◯ Agree	◯ Strongly Agree
5. I was annoyed during my interaction with the robot.				
◯ Strongly disagree	◯ Disagree	◯ Neither agree or disagree	◯ Agree	◯ Strongly Agree

Table 5.3: Section 3 of the questionnaire designated to the system (Max and the robot).

1. I thought that the system was easy to use.				
◯ Strongly disagree	◯ Disagree	◯ Neither agree or disagree	◯ Agree	◯ Strongly Agree
2. I found the various functions in the system were well integrated.				
◯ Strongly disagree	◯ Disagree	◯ Neither agree or disagree	◯ Agree	◯ Strongly Agree
3. I would imagine most people would learn to use the system rather quickly.				
◯ Strongly disagree	◯ Disagree	◯ Neither agree or disagree	◯ Agree	◯ Strongly Agree
4. I found the system very complicated to use.				
◯ Strongly disagree	◯ Disagree	◯ Neither agree or disagree	◯ Agree	◯ Strongly Agree

5.2.3 Section 4 - Evaluation of the Perceived Comfort While Working on the Task

In this section of the questionnaire, the NASA TLX questionnaire is used. The NASA TLX questionnaire measures the "workload" that the participant feels during the performance of the task. Originally, NASA TLX consists of six subscales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration) which the participant can rate. [National Aeronautics and Space Administration, 2020] [Hart, 2006]

In addition, the NASA-TLX questionnaire, traditionally, uses a 20-point or 100-point scale for the participants to rate the experiment, however, this could provide an unreliable score as there are more options. As an example, [Babakus and Mangold, 1992] state that "(...) the five-point format would reduce the frustration level of the respondent patients, and would thereby increase the response rate and the quality of the responses." and [Chyung et al., 2017] state "Keep an optimal balance in the scale design in order to lower the respondents' cognitive effort while maximizing the communication of information within the scale. Use a 5-point scale, as it is the best option from an information-processing perspective.". Thus, taking the above-mentioned into account, a shorter Likert scale (1 to 5) is implemented as it provides the researcher with a more reliable score from the participants. [Babakus and Mangold, 1992] [Chyung et al., 2017]

Moreover, for this project, the latter subscale is divided into affirmations and is addressed in Section 3 - Evaluation of the System's Usability (tables 5.1 and 5.2). This is done due to the original question having multiple adjectives that can be considered synonyms.

5.2.4 Section 5 - Evaluation of Engaging Interaction

The questions present in this section are based on the "Human Evaluation Questions" in [Sun et al., 2021]. In the article, the authors propose this sector to be divided into four different categories: *Engaging*, *Interesting*, *Humanlike* and *Knowledgeable*; where each category consists of two questions. For this paper, three out the four categories are chosen and each consists of one question. Furthermore, the questions are presented in table 5.4.

Table 5.4: Section 5 of the questionnaire. Engaging (Item 1); Interesting (Item 2); Knowledgeable (Item 3).

1. Who would you prefer to interact with?			
○ Max	○ An inexperienced worker	○ Follow the instructions written in the manual	
2. Does the system seem interesting to you?			
○ Totally	○ A bit	○ Not much	○ Not at all
3. Does Max sound well informed and confident for the task at hand?			
○ Totally	○ A bit	○ Not much	○ Not at all

5.2.5 Section 6 - Personalised Affirmations

The last section consists of a set of ten personalised affirmations with a five-point Likert Scale (from strongly disagree to strongly agree). This section covers the following aspects:

- *Comfortability Scale* describes the participant's sense of comfortability towards the robot both at the start and end of the experiment;
- *Perceived Competence* describes the participant's belief that the participant is able to carry on the task together with the system [Weiss et al., 2016];
- *Performance Expectancy* defines the degree to which the participants believes that by using the system helps with job performance [Weiss et al., 2016];
- *Effort Expectancy* defines the degree of ease that can be associated with the use of the system [Weiss et al., 2016].

Furthermore, the affirmations for each aspect are presented in table 5.5.

Table 5.5: Section 6 of the questionnaire. Comfortability scale (Items 1, 5 and 10); Perceived competence (Items 2 and 6); Performance expectancy (Item 3 and 7); Effort expectancy (Items 4, 8 and 9).

1. At the start, I was comfortable approaching the robot.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
2. I can operate the system good enough to carry out work.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
3. I believe that the use of the system would increase factories' effectiveness.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
4. It was easy for me to get the needed information from Max.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
5. By the end, I was comfortable approaching the robot.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
6. It was easy for me to operate the robot for this task.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
7. I can fulfil the task more efficiently with the help of the system.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
8. I was confident while interacting with Max.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
9. I felt capable of handling the system to perform the task.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree
10. I felt comfortable while interacting with Max.				
○ Strongly disagree	○ Disagree	○ Neither agree or disagree	○ Agree	○ Strongly Agree

Chapter 6

Experiment

The user studies took a total of two weeks and took place at an office at AAU. As mentioned in chapter 4 (Scenario), the task for the experiment consists of the participant being able to independently assemble the mock-up phone and name its components. In order to achieve this, the participant can either choose to interact with the system or read the user manual for 2 minutes[1]. Figure 6.1 shows a sketch of the setup of the experiment. An external microphone[2] is added, due to the microphone from the computer not being able to capture the voice commands.

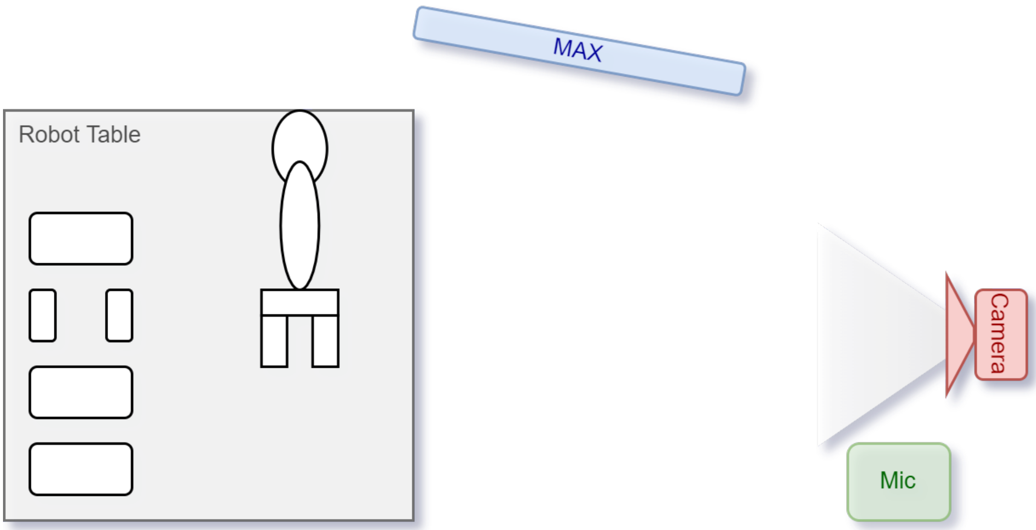


Figure 6.1: Sketch of the upgraded setup. The camera and microphone are placed at a distance of 2.2 and 1.6 meters from the edge of the robot table, respectively.

[1]This is the same time it takes the robot to do a demonstration of the phone assembly.

[2]Initially, the interaction with Max was set to be done through a headset, however, due to COVID-19, implementing an external microphone is preferred as it was found to be more hygienic for the participants.

Moreover, figure 6.2 presents the placement of the phone components on the robot table.

Figure 6.2: Placement of the phone components on the robot table from a distant (left) and closer (right) look.

The experiment follows the structure of the "stage script", which can be seen in Appendix C and the procedure is described in the following:

Pre-studies: At the start of the experiment, the participant is introduced to the overall goal and purpose of the user studies. Thus, a small introduction to Industry 4.0, the LTA-model and the problem the system is attempting to solve is described, followed by a brief explanation of the task at hand. Afterwards, the participant is introduced to the methods they can choose in order to complete the task: reading a manual for a limited time, or interacting with the robot for an unlimited time. Furthermore, each participant is noted that they are being recorded by video and audio during the whole experiment and, additionally, the researcher is taking notes in the background.

Conducting the user studies: The experiment can go different paths depending on the choice of the participant.

- *Manual:* In case the participant chooses the manual first, they have 2 minutes to read it and afterwards complete the task. In addition, independently of how the task is performed after reading the manual, the participant is asked to interact with the system.

- *System:* In case the participant chooses the system, they will not receive the manual and will only interact with the system. A small introduction on how to interact with Max and the robot is provided. For an interaction with Max, the participant is provided with a "cheat sheet" (see Appendix D) where they can check the keywords (in black) and examples (in grey) of commands they can use. The ellipsis on the "cheat sheet" represent a component. Thus, the participant has to provide Max with an adjective they think it is adequate to describe the component. Lastly, an explanation to how to interact with the robot is done.

Post-studies: The questionnaires (mentioned in section 5.2 (Questionnaire)) are filled in by the participants. Moreover, the participants are reminded that the answers to the questionnaire are anonymous.

Participants

In total, 29 participants took part in the study. Figure 6.3 presents the participants' demographics in regards to age and gender, where, between 20 and 25 years old, is the most prominent age group, 72.4% of the participants are male and 75.9% have a background in IT/Engineering. Furthermore, out of the 29 participants, 6 have never interacted with an industrial robot before, 2 have at least once and the remaining 21 have programmed and/or work with industrial robots. In addition, 9 of the participants never interact with a VA, 13 rarely and, 4 and 3 interact with a VA somewhat often and often, respectively.

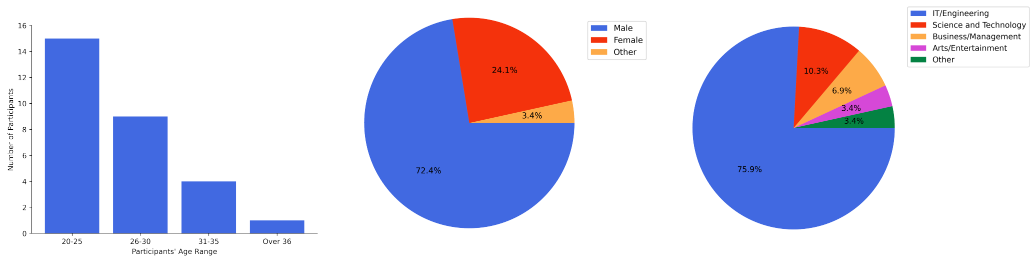


Figure 6.3: Demographics of the participants. **Left:** Age range; **Middle:** Gender; **Right:** Study/-Work Background.

Moreover, figure 6.4 presents the participants' previous interactions with industrial robots and VAs. On previous interaction with an industrial robot, 72.4% of the participants have programmed and/or work with an industrial robot, while 20.7% and 6.9% have never or have at least once interacted with an industrial robot, respectively. In addition, 44.8% of participants rarely interact with a VA and 31.0% never interact with a VA, the remaining 24.1% of the participants interact with a VA somewhat often and often.

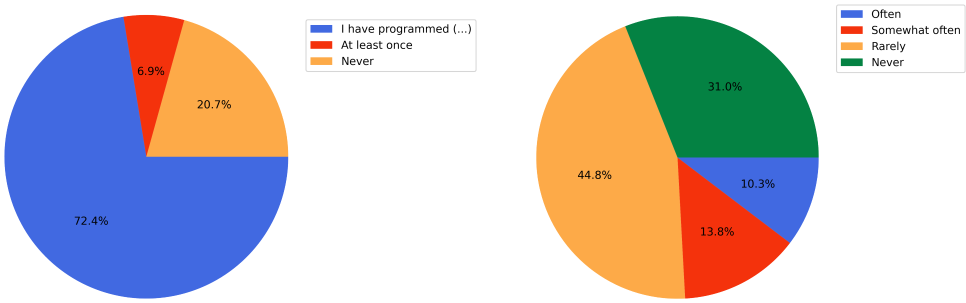


Figure 6.4: Previous interactions of participants. **Left:** Previous interaction with an industrial robot; **Left:** How often do participants interact with VAs.

Furthermore, figure 6.5 presents the participants' preferences between the User Manual and the system. As it possible to see, initially 20 out of 29 of the participants chose the system first while the remaining chose the User Manual. Additionally, out of 20 participants that chose the system first, 16 said they would choose the system, while 4 would choose the User Manual if given the opportunity to repeat the experiment. Moreover, out of the 9 participants that chose the User Manual first, 6 said they would choose the User Manual again, while 3 would change to the system.

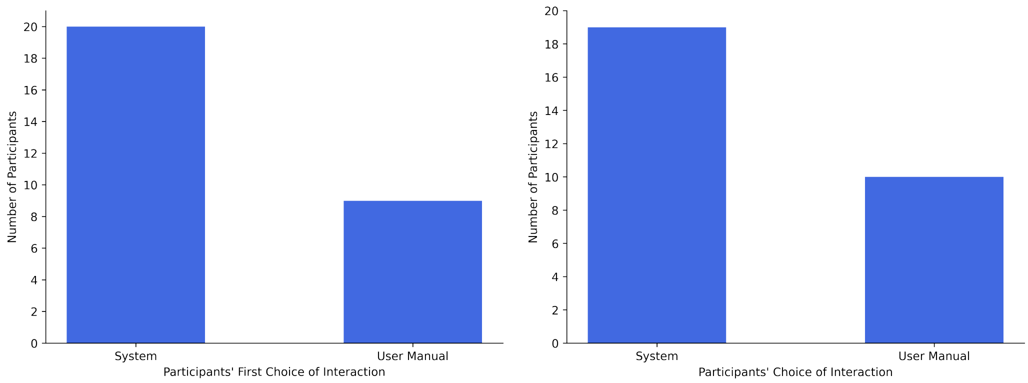


Figure 6.5: Section 2 answers. **Left:** Participants' choice when performing the User Studies; **Right:** Participants' choice if given the opportunity to perform the User Studies again.

In addition, 17 out of the 22 participants with a background in IT/Engineering chose an interaction with the system. Out of the remaining 7 participants, 3 chose an interaction with the system. Lastly, 5 out of the 7 participants that interact with a VA somewhat often and often, have chosen to interact with the system.

6.1 Results

As mentioned in Section 5.2 (Questionnaire), the questionnaire is divided into six sections. Each section, from 3 to 6, is evaluated independently and, once all the sections have been evaluated, the performance of the system can be concluded.

6.1.1 Section 3 - Evaluation of the System's Usability

The SUS is a ten-item questionnaire where each item has a five-point Likert scale for the participant to rate the experiment. In order to calculate the SUS score, it is needed to calculate the score of each item. Thus, the scale for each item gets converted to points as it follows: [Brooke, 1995]

- **Strongly Disagree:** 1 point;
- **Disagree:** 2 points;
- **Neither Agree or Disagree:** 3 points;
- **Agree:** 4 points;
- **Strongly Agree:** 5 points.

In addition, equations 6.1, 6.2 and 6.3 present how to achieve the score of the odd and even items, and the total final score, respectively, where n represents the sum of the points in odd items and m represents the sum of the points in even items. Furthermore, the total score is presented in the range of 0-100, where the higher the score the better. [Brooke, 1995]

$$score_{odd} = n - 5 \tag{6.1}$$

$$score_{even} = 25 - m \tag{6.2}$$

$$TotalScore = (score_{odd} + score_{even}) \times 2.5 \tag{6.3}$$

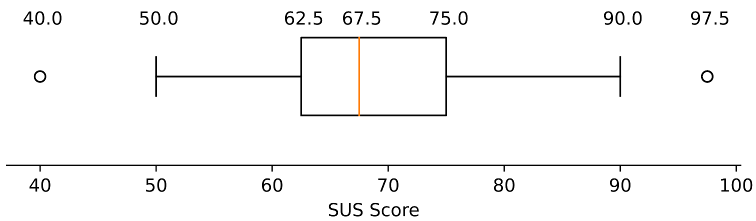


Figure 6.6: Box plot of the SUS scores.

Figure 6.6 shows the box plot of the SUS scores. The scores 40 and 97.5 are considered outliers as only one and two participants scored these, respectively. Furthermore, the minimum and maximum SUS scores are 50 and 90, respectively, and the mean is at 68.53.

6.1.2 Section 4 - Evaluation of the Perceived Comfort While Working on the Task

The NASA-TLX questionnaire consists of six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration; and the participant can rate each through a five-point Likert scale.

As mentioned in section 5.2 (Questionnaire), the last subscale of the NASA-TLX is divided into multiple affirmations in Section 3 of the questionnaire. To count the score of these, the mean of the affirmations for Max and the robot are calculated and used as the score value for the subscale.

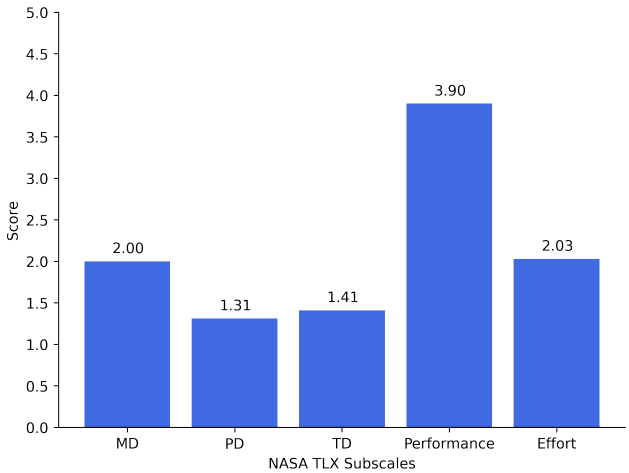


Figure 6.7: NASA TLX score. MD - Mental Demand; PD - Physical Demand; TD - Temporal Demand.

As it is possible to see in figure 6.7, the average for Mental Demand, Physical Demand, Temporal Demand and Performance are lower than or equal to 2 out of 5. Moreover, Effort scored the highest with an average of approximately 3.90 out of 5.

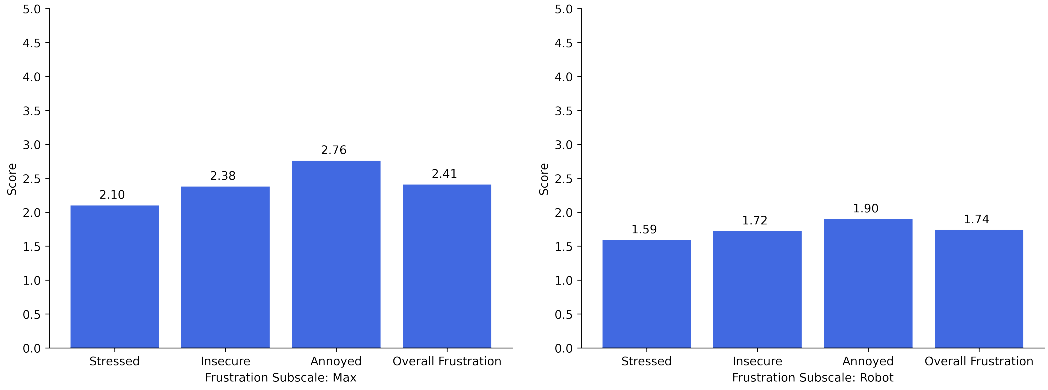


Figure 6.8: Latter subscale of NASA TLX divided into three categories: Stressed, Insecure and Annoyed. Overall Frustration presents the mean of the three categories. **Left:** Frustration subscale score regarding Max; **Right:** Frustration subscale score regarding the robot.

Furthermore, figure 6.8 presents the latter subscale of NASA TLX. Overall, Max scored higher than the robot with a difference of approximately 0.68. Furthermore, Max scored higher than 2 in all categories, while the robot scored less than 2.

6.1.3 Section 5 - Evaluation of Engaging Interaction

This section consist of three questions. To get the evaluation for this section, an analysis of the participants' answers is made.

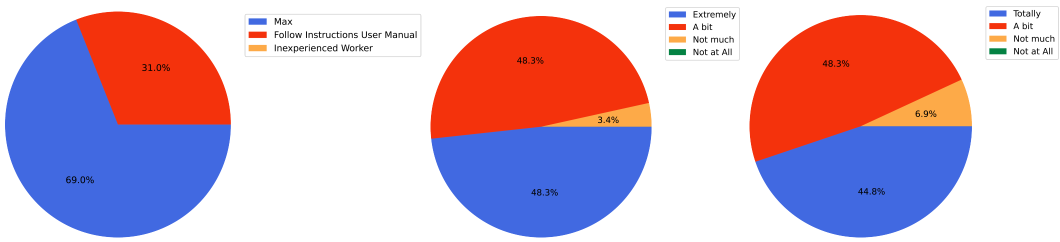


Figure 6.9: Section 5 answers. **Left:** "Who would you prefer to interact with?"; **Middle:** "Does the system seem interesting to use?"; **Right:** "Does Max sounds well informed and confident for the task at hand?".

As seen in figure 6.9, for the first question of the section 69.0% of the participants said that they would prefer to interact with Max instead of following the instruction written on the User Manual, and no participant voted for a preferred interaction with an inexperienced worker. Moreover, 48.3% for the participants rated the system as *Extremely* interesting, while the same amount of participants rated the system as *A bit* interesting. Lastly, 44.8% and 48.3% of the participants thought that Max sounded "*Totally*" and "*A bit*" well informed and confident.

6.1.4 Section 6 - Personalised Affirmations

The last section of the questionnaire consists of ten personalised affirmations with a five-point Likert Scale, as in Section 3. Thus, the scale to points conversion used previously is also used for this section. Furthermore, this section covers four aspects: Comfortability Scale (Items 1, 5 and 10), Perceived Competence (Items 2 and 6), Performance Expectancy (Items 3 and 7) and Effort Expectancy (Items 4, 8 and 9). For each, the mean of the item's score is calculated and is designated as the score for the attributed aspect. With this, the scores of each aspect vary from 1 to 5. Figure 6.10 presents the score of the last section of the questionnaire.

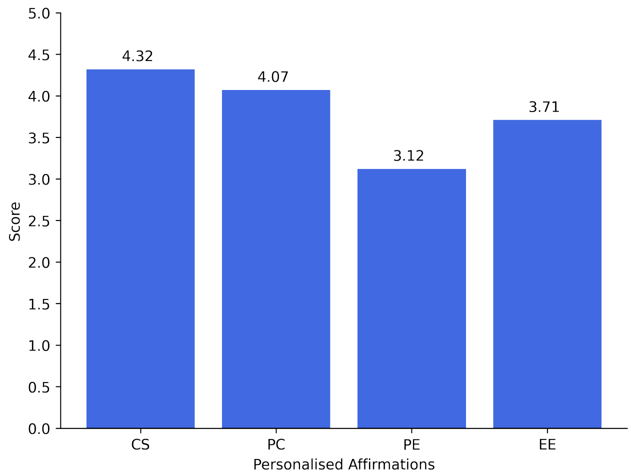


Figure 6.10: Scored of the Personalised Affirmations. CS - Comfortability Scale; PC - Perceived Competence; PE - Performance Expectancy; EE - Effort Expectancy.

In addition, for the Comfortability Scale aspect the comparison between items 1 and 5 are:

- *Rated 5 points in both items:* 12 participants;
- *Rated 4 points in both items:* 6 participants;
- *Rated 3 points in both items:* 1 participant;
- *Rated 2 in Item 1 / rated 4 in Item 5:* 1 participant;
- *Rated 3 in Item 1 / rated 4 in Item 5:* 2 participants;
- *Rated 4 in Item 1 / rated 5 in Item 5:* 4 participants;
- *Rated 5 in Item 1 / rated 4 in Item 5:* 3 participants.

Chapter 7

Discussion

This chapter presents events that occurred during the execution of this project. Furthermore, it presents an analysis and discussion regarding the results obtained from the User studies presented in chapter 6 (Experiment).

7.1 User Studies Design

During the designing of the User Studies, some complications were met. These are mainly in regards to the questionnaire and are described in the following subsections.

7.1.1 Section 3 - Evaluation of the System's Usability

Ideally, each participant would fill up a complete SUS questionnaire for each part of the system: Max, the robot and the system. However, following this path would be very bothersome and tiresome for the participants, taking into account that the participants still needed to fill up all the other sections of the questionnaire. Thus, it was favourable to divide it into three parts and evaluate it overall.

7.1.2 Section 4 - Evaluation of the Perceived Comfort While Working on the Task

Originally, the NASA TLX questionnaire includes a weighting system, where each participant should define a weight to give each subscale. When analysing the score of the questionnaire, the weight is multiplied by the participants' rates of the subscale. [Hart, 2006] For this project, adding the weight system is abstained from as the system could deviate the participants attention. In addition, the effort and frustration subscales are analysed more closely in order to get more input for future versions of the system, and Physical Demand is the subscale that is least analysed.

7.2 Results

It is important to note that the experiment took place during the first two weeks of December, with this, multiple students were at the university, together with the staff, and all these parties were using the same network. This caused the AAU network to be overloaded. [Chrysostomou et al., 2021] This caused Max to sometimes stay silence for minutes while it was attempting to capture what the participant was saying and processing the long sentences to extract the needed keywords. Thus, a lot of participants has negative comments towards the system.

Furthermore, the participants for this experiment have a broad geographic distribution, resulting in a lot of accents when speaking with Max. This caused Max to have a harder time understanding some of the participants', with a more pronounced accent, commands. Some examples of the most common mistakes are that Max would understand:

- "ring" instead of "bring";
- "collar" instead of "cover";
- "neat" instead of "need";
- "bottle" instead of "bottom".

Moreover, it is important to notice that Max's microphone would reset every 2 seconds if it did not understand or hear any command [Chrysostomou et al., 2021]. During some of the experiments, this would happen while the participant was giving a command, meaning that Max would only detect the second half or latter part of the command. With this, sometimes Max would preform the wrong action for the said command. E.g. the participant would say "Can you *bring* me the *black* piece?" and Max would only understand the latter keyword (*black*) and would provide the definition for the bottom cover.

Lastly, as mentioned in chapter 6 (Experiment), the location of the experiment was in an office/room at AAU. This room was shared with three AAU staff members. With this, Max would sometimes detect the other members' conversations, their walking and the movement of people entering and leaving the room. This caused some background noise for Max, together with Max taking longer to perform the action as it kept listening to see if new keywords would be detected.

7.2.1 Section 2 - Personal Preferences

The results from section 2 of the questionnaire presented that if given the opportunity to do the experiment again, fewer participants would choose to interact with the system. Some of the participants finalised the experiment expressing some discontent. Not towards the idea, but towards the advancement of the system.

7.2.2 Section 3 - Evaluation of the System's Usability

As previously mentioned, SUS scores range from 0 to 100 and 68 is considered to be the 50th percentile or, colloquially known as, the average score. The author in article [Klug, 2017], proposed the table presented in figure 7.1 which implements the letter grading system into the numerical score of SUS. [Klug, 2017]

Letter grade	Numerical score range
A+	84.1–100
A	80.8–84.0
A-	78.9–80.7
B+	77.2–78.8
B	74.1–77.1
B-	72.6–74.0
C+	71.1-72.5
C	65.0–71.0
C-	62.7–64.9
D	51.7–62.6
F	0–51.6

Figure 7.1: Proposed letter grading for SUS numerical score. [Klug, 2017]

This section of the questionnaire scored an average of 68.53. Following the grading system proposed by [Klug, 2017], the system for this project scores a C. In addition, as above-mentioned, the average score for the SUS is 68 and this section scored 68.53, meaning that it falls closely to the average. Thus, it is possible to conclude that the system does not "fail", however, it also barely "passes".

7.2.3 Section 4 - Evaluation of the Perceived Comfort While Working on the Task

In this section, the result of each subscale is analysed closely.

Mental Demand: Mental Demand scored a 2.00 for the experiment. Individually evaluating the results, while most participants (12) rated the experiment with 1 in this subscale, a sum of 9 participants rated 3 and 4. This rating could have originated from participants with a heavier accent as these participants needed to repeat the command multiple times in order to get the correct response from Max, which could have caused some distress. However, as the answers of the questionnaire were submitted anonymously, it is not possible to be certain of the reason.

Physical Demand: As above-mentioned, this subscale was of least interest for the user studies, in addition, it is also the subscale that scored the lowest value. Taking a look at the individual answers, two participants rated this subscale with 4, while the remaining rated with 1 and 2. This could have been a result of a bad interpretation from the participant or a misunderstanding of the question.

Temporal Demand: Just with a 0.10 difference from Physical Demand, Temporal Demand is the second lowest rating of this section. When initiating the experiment, the participant are told to interact with the system for as long as they wanted and, in addition, the participants are also reminded that they are not being evaluated during the experiment, the system is. This could have caused the participants to not feel like they were "rushed" or "hurried" and they could perform the task without time constrains.

Performance: With the highest rating in this section, the subscale Performance scored approximately 3.90. This subscale evaluates how successful the participants would rate their own performance. With only 2 participants out of 29 rating the performance with 1 and 2, the remaining rated the experiment with 3 and higher, with 9 participants rating 5. During the user studies, it was noted that 2 of the participants did not know the name of all the components by the end. It is safe to assume that the 2 participants that rated 1 and 2 are the same participant that did not meet the final goal of the task. However, as the answers are anonymous, it cannot be completely confirmed and/or reassured.

Effort: Out of the 29 participants, a conjunction of three, rated this subscale 4 and 5. Two of these participants could have been the same ones that on the previous subscale scored the lowest. Furthermore, it was observed that some participants had a harder time with Max, which could have also led into scoring higher on this subsection as it could have required harder work to accomplish the level of performance.

Frustration: As previously mentioned, this subscale is divided into a total of six affirmations distributed between the interaction with Max and the robot. Overall, Max scored higher than the robot. When analysing each affirmation individually, it is possible to notice that the feeling of "annoyed" is the highest for both interactions.

7.2.4 Section 5 - Evaluation of Engaging Interaction

This section evaluates three aspects of the VA. Overall, this section provided promising results from the participants.

Engaging: Even though, 19 out of the 29 participants would choose to interact with the system if repeating the experiment, 69.0% of the participants (20 participants) said that they would have preferred an interaction with Max compared to following instructions from the User Manual or interact with an inexperienced worker. This is a promising result as it shows high engagement and initiative from the participants towards Max.

Interesting: With only 3.4% of the participants rating the system as "*Not much*" interesting, the remaining 96.6% of the participants found the system "*Extremely*" and "*A bit*" interesting. This results shows that most participants found the system interesting, which can increase their curiosity and interaction.

Knowledgeable: Even though, this aspect was the one with the least positive result from the participants in this section, a total of 93.1% of the participants thought that Max sounds "*Totally*" and "*A bit*" well informed and confident with the task at hand. It is possible to assume that the results from this aspect came from the fact the Max does not provide follow-up questions, e.g. some participants when asking the terminology of a component ("What is the name of the black case?"), the following question would be to pass the object previously asked for ("Can you pass me **that**?"). Moreover, when Max would not understand the participants' commands or would provide a different answer from what was originally asked (e.g. the user would ask "bring me the PCB" and Max would answer with the terminology of PCB), the participants also showed some discontent.

7.2.5 Section 6 - Personalised Affirmations

This section is divided into four personalised affirmations. Overall, this section provided promising results from the participants.

Comfortability Scale: With the highest score, comfortability scale scored an average of approximately 4.32 out of 5. Looking at the items individually, Item 5 scored highest with approximately 4.52, followed by Item 1 with a score of 4.34 and lastly Item 10 with 4.10. By observing that Item 5 scored higher than Item 1, is it possible to conclude that the users got more comfortable interacting with the robot as the experiment progressed. In addition, while Item 1 had one participant that rated with a 2, Item 5 did not have any participant rating lower than a 3.

Perceived Competence: Following comfortability scale, perceived competence scored the second highest with an average of approximately 4.07. Looking at the Items individually, these scored similar results with averages of 4.03 and 4.10 for Item 2 and Item 6, respectively.

Performance Expectancy: Even though performance expectancy scored the lowest in this section with an average of approximately 3.12, it is still a promising result. By analysing the Items individually, Item 3 scored approximately 3.24 while Item 7 scored 3.00. Thus, the information gathered from the participants' rating is that they do not strongly believe that using the system helps, however, they do not disagree with the affirmations.

Effort Expectancy: With an overall score of approximately 3.71, individually looking, Item 9 scored the highest with 4.17, followed by Item 8 with 3.80 and lastly Item 4 with 3.17. Thus, the score of Item 9 presents a feeling of capability from the participant and Item 8 presents a feeling of confidence while interacting with Max. It can be assumed that Item 4 scored the lowest by the problems that some users had while interacting with Max that are mentioned above.

In conclusion, the results for the experiment were, overall, promising. Whereas it is possible to conclude that the participants had an easier and better time interacting with the robot and the system, rather than just Max.

Chapter 8

Conclusion

This report presents the reason why a User Studies is needed when developing a new and promising system to be implemented. With the information gathered in chapter 2 (Background Analysis), it is possible to conclude why factories are adapting to Industry 4.0 together with the challenges that this presents. Furthermore, the chapter presents the LTA-model approach in order to facilitate this adaptation. In addition, the implementation of VAs in a workplace is explored and analysed. This allows the main hypothesis for this project to be defined as the following:

> *Adopting a Natural Language-enabled Virtual Assistant for LTA approach, provides an easier implementation of Industry 4.0 in factories.*

The scenario chosen for the User Studies is described in chapter 4 (Scenario), together with sub-models for the hypothesis. These sub-modules are:

- Users with a background in IT/Engineering prefer the usage of the system for the experiment;
- Users tend to be more comfortable interacting with the system by the end of the experiment;
- Users that do not interact somewhat often or often with a VA are inclined to choose the User Manual.

Additionally, chapter 4 (Scenario) presents how this scenario can follow the LTA-model approach. Once the hypothesis and scenario for the User Studies are defined, it is possible to start the analysis and design of the studies. Chapter 5 (User Studies Design) presents an analysis on different study contexts and the advantages and disadvantages of each, together with a within- and between-participants analysis. For this project, it is concluded that a mix of laboratory and field study with a mixed participants design was the most beneficial option. Furthermore, the chapter provides an analysis of the chosen questionnaires in order for the participants to rate the system. Lastly, chapter 6 (Experiment) presents the results gathered from the participants.

Prior to answering the hypothesis, it is needed to analyse the results gathered from the hypothesis sub-modules.

Users with a background in IT/Engineering prefer the usage of the system for the experiment

As mentioned in chapter 6 (Experiment), 17 out of the 22 participants with a background in IT/Engineering chose to interact with the system. From the remaining 7 participants, less than half, 3, chose to interact with the User Manual. Thus, it is possible to see that approximately 77% of the participants with IT/Engineering background chose the system, while 42% of the remaining participants with another background chose the User Manual.

Users tend to be more comfortable interacting with the system by the end of the experiment

By analysing the results presented in section 6.1.4 (Section 6 - Personalised Affirmations), in Item 1 a total of 15 participants rated the affirmation with 5, 10 rated with 4, 3 rated with 3 and only 1 rated with 2. However, when analysing Item 5, no participant rated the affirmation with 2, only 1 participant rated the affirmation with 3, 12 rated with 4 and 16 rated with 5. This caused a higher average score in Item 5 compared with Item 1, concluding that the participants got more comfortable with the interaction by the end of the experiment.

Users that do not interact somewhat often or often with a VA are inclined to choose the User Manual

Observing the results in chapter 6 (Experiment), out of 29 participants a total of 7 participants interact somewhat often and often with a VA. Out of these 7 participants, 5 chose to interact with the system. From the remaining 22 participants, 15 chose to interact with the system. Thus, it is possible to see that approximately 68% of the participants that rarely or never interact with a VA chose the interaction with the system.

Furthermore, by analysing each section of the questionnaire it is possible to gather promising inputs from the participants. For example, even though the system would fail for some of the participants, 69.0% of the participants still voted that they would prefer to interact with Max, rather than following the instruction written on the User Manual (31%) or interact with an inexperienced worker (0.0%).

Taking a look at the hypothesis again, by gathering all the above-mentioned information together with the information presented in chapter 6 (Experiment), it is possible to conclude that the participants provided mainly positive ratings and scores towards the system. Nevertheless, a more complex task and a bigger sample size would have been ideal to answer such hypothesis. However, an overall positive rating and scores throughout the experiment, provides a promising view towards the system.

Inputs from the Participants

Before submitting the questionnaire, the participants are presented with an open question where they can provide some input towards the developers of Max. Even though a lot of information can be provided from the users when answering the questionnaire, the open questions opens possibilities for the developers to get an outside input. The answers in this question are not positive or negative, but rather informative and helpful.

Some users presented the idea of implementing of type of "training" before interacting with Max, so the complications regarding specific accents could be avoided. Furthermore, some participants did not enjoy the "listening at all time" approach and would rather prefer a "press to talk" approach, while other participants suggested Max to start the interaction with an introduction of its functionalities. In addition, multiple participants projected a satisfactory feeling and commented that the interaction with Max was pleasant and fun.

Gathered Knowledge from the User Studies

From the researcher point of view, when designing a User Study multiple aspects can be missed or not taken into account. To address this, a pilot test is performed. For this project, the pilot test consisted of one participant and a lot was learned and noted from the one participant test. However, different people can provide different inputs and can detect different missed aspects, thus, a hand-full of participants might have been more beneficial.

Furthermore, when deciding the questionnaires for the participants, the researcher is always presented with a large variety of options. It is important for the researcher to analyse each option carefully and determine if that option should be included. Moreover, a questionnaire can also partly be included if the researcher finds it as such.

When performing the experiment, it is helpful to have a "stage script". However, it is important to keep in mind that every participant is different. So even with the help of a script, some participants might differ more than expected.

Lastly, it is important for the researcher to keep in mind that any result is a good result. When performing a User Study, a negative or positive score from the participants, provides the researcher and developers with a better idea of what problems the system might face upon implementation. Thus, any input from participants is helpful and valued in order to improve the system.

Bibliography

American Psychological Association (2021). *mixed design.* https://dictionary.apa.org/mixed-design.

Babakus, E. and W. G. Mangold (1992). Adapting the servqual scale to hospital services: an empirical investigation. *Health Services Research 26*, 767–786.

Baggaley, K. (2017). *Robots are replacing humans in world's mines. Here's why.* https://www.nbcnews.com/mach/science/robots-are-replacing-humans-world-s-mines-here-s-why-ncna831631.

Bartneck, C. and J. Forlizzi (2004). Shaping human-robot interaction: Understanding the social aspects of intelligent robotic products. *Human Factors in Computing Systems*.

Brains, J. (2020). *I've tested more than 20 robot vacuums, but the Roborock S6 has the best suction I've seen.* https://www.businessinsider.com/roborock-s6-robot-vacuum-mop-review?r=US&IR=T.

Brooke, J. (1995). Sus: A quick and dirty usability scale.

Burbach, L., P. Halbach, N. Plettenber, J. Nakayama, M. Ziefle, and A. C. Valdez (2019). "hey, siri", "ok, google", "alexa". acceptance-relevant factors of virtual voice-assistants. *2019 IEEE International Professional Communication Conference*, 101–111.

Cambridge Dictionary (2021). *interaction.* https://dictionary.cambridge.org/dictionary/english/interaction.

Casillo, M., F. Colace, L. Fabbri, M. Lombardi, A. Romano, and D. Santaniello (2020). Chatbot in industry 4.0: An approach for training new employees. *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 371–376.

Chandler, J., P. Mueller, and G. Paolacci (2013). Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behaviour Research Methods 46*, 112–130.

Chrysostomou, D., C. Li, and D. Pinto (2021). *Title of the page.* Source representing the interaction between the student/researcher (Daniela Pinto), the supervisor (Dimitris Chrysostomou) and one of the developers of Max (Chen Li).

Chyung, S. Y., K. Roberts, I. Swanson, and A. Hankinson (2017). Evidence-based survey design: The use of a midpoint on the likert scale. *Performance Improvement 56*, 15–27.

Cóndor-Herrera, O., J. Jadán-Guerrero, and C. Ramos-Galarza (2020). Virtual assistants and its implementation in the teaching-learning process. *Advances in Intelligent Systems and Computing 1269*.

Devlin, J. and M.-W. C. an Kenton Lee abd Kristina Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Association for Computational Linguistics*, 4171–4186.

Frank, A. G., L. S. Dalenogare, and N. F. Ayala (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal of Production Economics 210*, 15–26.

Goodrich, M. A. and A. C. Schultz (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction 1*, 203–275.

Hamada, T. (2019). Determinants of decision-makers' attitudes toward industry 4.0 adaptation. *Social Sciences 8*, 1335–1347.

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50*, 904–908.

Heyer, C. (2010). Human-robot interaction and future industrial robotics applications. *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4749–4754.

Hjorth, S. and D. Chrysostomou (2022). Human–robot collaboration in industrial environments: A literature review on non-destructive disassembly. *Robotics and Computer-Integrated Manufacturing 73*.

Hoffman, G. and X. Zhao (2020). A primer for conducting experiments in human-robot interaction. *Association for Computing Machinery 10*, 6:1–6:31.

Klug, B. (2017). An overview of the system usability scale in library website and system usability testing. *Weave: Journal of Library User Experience 1*.

Lase, H., P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann (2014). Industry 4.0. *Business & Information Systems Engineering 6*, 239–242.

Lester, S. and C. Costley (2010). Work-based learning at higher education level: value, practice and critique. *Studies in Higher Education 35*, 561–575.

Li, C., A. K. Hansen, D. Chrysostomou, S. Boeg, and O. Madsen (2022). Bringing a natural language-enabled virtual assitant to industrial mobile robots for learning, training and assistance of manufacturing tasks. *The 2022 IEEE/SICE International Symposium on System Integration*.

Li, C. and H. J. Yang (2021). Bot-x: An ai-based virtual assistant for intelligent manufacturing. *Multiagent and Grid Systems 17*.

Li, M., A. Milojevic, and H. Handroos (2020). Robotics in manufacturing - the past and the present. *Technical, Economic and Societal Effects of Manufacturing 4.0*, 85–95.

National Aeronautics and Space Administration (2020). *NASA TLX: Task Load Index*. https://humansystems.arc.nasa.gov/groups/TLX/.

Nichols, M. R. (2020). *The Rise of Robots in Industry 4.0*. https://www.thomasnet.com/insights/the-rise-of-robots-in-industry-4-0/.

Presentation Point (2018). *Industry 4.0: Importance of Data and Signals and Triggers*. https://www.presentationpoint.com/blog/data-signals-triggers-industry-4-0/.

Rehe, G., R. Schneider, and D. D. Thimm (2020). Transformation towards human workforce 4.0 through the qualification model lta-fit. *53rd Hawaii International Conference on System Sciences*, 4601–4610.

Rueben, M., S. A. Elprama, D. Chrysostomou, and A. Jacobs (2020). *Human-Robot Interaction: Introduction to (Re)Using Questionnaires in Human-Robot Interaction Research*. Springer.

Sun, K., S. Moon, P. Crook, S. Roller, B. Silvert, B. Liu, Z. Wang, H. Liu, E. Cho, and C. Cardie (2021). *Adding Chit-Chat to Enhance Task-Oriented Dialogues*. https://arxiv.org/abs/2010.12757.

Thrun, S. (2011). Toward a framework for human-robot interaction. *Human-Computer Interaction 19*, 9 – 24.

University of Pittsburgh (2018). *Citation Styles: APA, MLA, Chicago, Turabian, IEEE: Chicago 16th/Turabian 8th*. http://pitt.libguides.com/c.php?g=12108&p=64732 [Accessed: 2018-09-07].

Weiss, A., A. Huber, J. Miniichberger, and M. Ikeda (2016). First application of robot teaching in an existing industry 4.0 environment: Does it really work? *Societies 6*.

Part I

Appendices

Appendix A

User Manual

User Manual

- Smart Production Lab and Phone Assembly -

Aalborg University
Department of Electronic Systems
Fredrik Bajers Vej 7B
DK-9000 Aalborg

Contents

Chapter 1

Introduction

This user manual is intended to be used as a guidance for phone assembly and provide the user the basics of the Smart Production laboratory and phone assembly. In this manual, you can find a broad description and introduction to:

- Smart Production Lab and its services:
- Phone components;
- Relationship between the phone components;
- Correct process of phone assembly.

Furthermore, it is expected that the user can answer questions regarding the Smart Production Lab and phone components, as well as, be able to assemble the phone independently and correctly after reading this manual.

Chapter 2

Smart Production Laboratory

The AAU Smart Production Lab is a state-of-the-art reconfigurable demonstration of technologies comprised of multiple Festo modules, also known as Cyber-Physical modules (CP Factory) (see figure 2.1). The CP factory is Festo's interpretation of Industry 4.0 by having high flexibility, and being reconfigurable to changes on the production line. The system is made up of small individual modules. Each module can operate independently from one another, and can be connected to other modules, resulting in one big production line. This makes it possible to change the setup of the factory by simply splitting apart the modules and combining them in new ways with minimal setup time.

Figure 2.1: Festo cyber-physical Laboratory.

The main task of the Smart Production Lab is the assembly of a phone, where the colour of the cases and the number of fuses can be customised. However, for this project, the assembly of the phone is performed by an industrial robotic manipulator (see figure 2.2) in order to teach the user how to assemble the phone correctly.

Figure 2.2: Industrial robotic manipulator used for demonstration of assembly of the phone. The phone components are placed within reach from the robot.

For more information regarding the Smart Production Lab, you can check the official webpage: www.smartproduction.aau.dk.

Chapter 3

Phone

Phone Components

The phone consists of a total of 5 components (see figure 3.1).

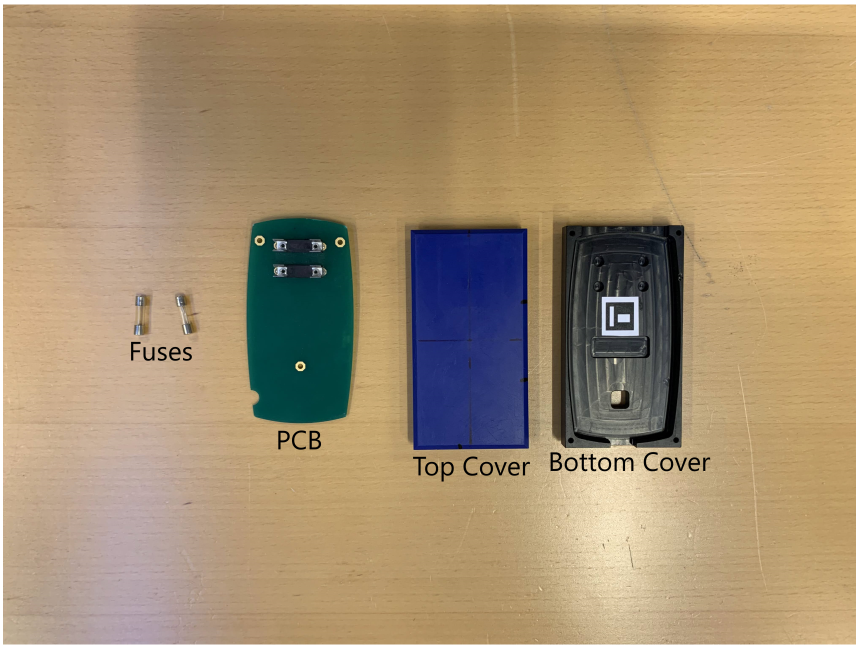
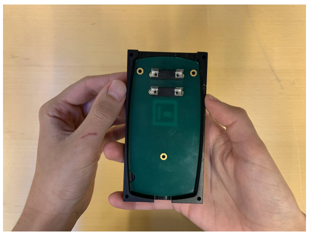
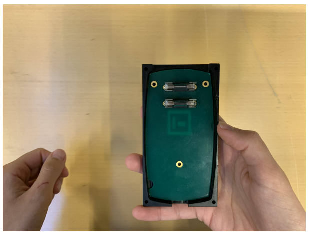


Figure 3.1: Components of the phone. From left to right: 2 Fuses, 1 PCB, 1 Top Cover, 1 Bottom Cover.

Phone Assembly

The following tables present the assembly process of the Phone. Please note that the objects that are mentioned inside parenthesis () are the objects from the previous action that are still needed.

Action	Components	Notes	Image
Identifying Bottom Cover	Bottom Cover	In this example, the bottom cover is black and can be identified as the case with 4 holes	
Identifying the PCB and PCB Orientation	(Bottom Cover) PCB	The PCB is green and has 4 soldering points in the back that match with the 4 holes in the bottom cover	
Demonstration of the PCB placed	(Bottom Cover) (PCB)	When the PCB has been placed correctly on the black cover, it should look like in the picture on the right.	
Identification of both fuses	(Bottom Cover) (PCB) 2 Fuses	The image demonstrates the identification of one fuse, as both look the same	

Action	Components	Notes	Image
Verification of fuses placement	(Bottom Cover) (PCB) (Fuses)	In order to assert the fuse on the PCB, it is needed to slightly push the fuse down into position	
Identifying Top Cover and its orientation	(Bottom Cover) (PCB) (Fuses) Top Cover	In this example, the top cover is blue and can be identified as the case without 4 holes. It has the correct orientation when its deepest holes can match the fuses that are present on the PCB	
Finalising the assembly	(Bottom Cover) (PCB) (Fuses) (Top Cover)	A final press is needed once the phone is fully assembled, to tightly fit the two parts together.	

Appendix B

Pilot Test

When designing a user studies, it is very likely that certain aspects are not taken into account, or are missed by the researcher. Therefore to avoid these, a pilot test is performed. Pilot test is a test performed before recruiting participants and performing the user studies. It is designed as trial run of the experiment. From the pilot test, the researcher can gather valued information from the participant which can lead to a change of setup, "stage script" or even adding or removing certain aspects from the experiment procedure. [Hoffman and Zhao, 2020] Moreover, for this project, the pilot testing consists of one participant.

B.1 Setup

The setup for the user study was, initially, placed at the robotics lab at AAU. It is an open space with multiple machinery around and constant people and robots movement. Figure B.1 shows a moment of the pilot study. The table placed next to the robot has the computer displaying Max web interface and the white-board, that can be seen on the right side of the figure, has a list of the keywords that can be used with Max, together with some examples to guide the participants.

Figure B.1: Photo taken during the pilot testing.

B.2 Outcomes

During the pilot, one of the first aspects that was noted, was that the participant would repeat the activation sentenced for Max every time they wanted to give a new command. Furthermore, it was also noted that the participant did not know how to approach the robot in order to make it release a phone component. Moreover, the researcher also took note of the multiple adjectives that the participant would use to describe each of the phone components that were not initially added to the code.

Taking into consideration the pilot testing, the following aspects are added and changed for the experiment:

- Instead of using a white-board with the commands, it is preferable to use a paper so the participant does not need to concentrate in multiple boards;
- Instead of having Max web interface just on a computer, it is preferable to have it in a larger screen so it would be possible to get all the written information more easily;
- The setup needs to be moved because a lot of external and background noise happen at the lab;
- A more detailed description on how to interact with Max is added to the script (e.g. no need to repeat the activation sentence before every command);
- A more detailed description on how to push the gripper of the robot to release a component;
- The keywords "tube" and "cylinder" are added in order to describe the fuse;
- The keyword "steps" is added in order to get the answer for "How many steps are in the assembly?" (same answer for "How many processes are in phone assembly?").

The participant also filled in the questionnaire, but did not have any comments for it.

Overall the pilot test presented helpful and valued inputs that were taken into account when changing the experiment design.

Appendix C

Stage Script

Pre-introduction

- This is an experiment that will be used for a user study for the master thesis
- You will be recorded both visually and in audio
    - Your name will be anonymous and not mentioned in the report
    - You can tell me now if you want or not pictures of your experiment in the report
    - Only I will have access to the recordings and they will not be shared. They are used in case I did not notice anything during the experiment and want to analyse it after
- Before the experiment, there will be a small introduction
- After there will be an online survey that I would ask you to fill in
    - It is fast to answer and it can take max. 15 minutes for some people.
- I will be present during the experiment taking notes and outside the camera field of view

Introduction

- Do you know or have an idea of what Industry 4.0 is?
    - New industrial revolution taking place now
    - Making everything digital and in the same cloud
- When hiring someone new, it is needed to give an introduction to the worker of the product together with the facility
    - For this, an expert is needed
    - A lot of times the expert is busy or not in the office so it can make things more complicated
- The point of this experiment is to introduce a Language Enabled Virtual Assistant for Learning, Teaching and Assist approach that factories can implement when a new employee comes
    - This Virtual Assistant cannot provide you as much information as the expert but is expected to cover at least the basics of a task
- However, instead of interacting with the virtual assistant and robot, you can perform the task following this manual

Task

- The task consists of Assembling a dummy phone and being able to know the names of each component
- You can either opt for the manual or the virtual assistant for this task
    - Manual: You have 2 minutes to perform the task and I will be timing
        - I will not take part in it, so it is only the subject performing the task only with the manual
    - VA: You can interact with it as much as you want
- By the end, I will ask you to assemble the phone independently, without the help of the manual/robot, while naming each of the components

Manual

- Even though you chose the manual, I would like to ask you to interact with the VA/robot

VA/Robot

- The system consists of a Virtual Assistant called Max and a robot manipulator called Franka
  - The robot is collaborative so you will be interacting with it
  - I will be holding the emergency button just in case, but this robot has been tested multiple times and it is safe
- The Virtual Assistant interface can be seen on the screen
  - You will be talking with the Max
  - To talk with Max, there is this paper with a list of actions/commands that can be used
  - In black are the keywords for the action and in grey is an example of a sentence
    - You can use the keywords and interact with the robot however you want
    - I will not interfere in the interaction unless needed
- Pay attention to how the components of the phone are placed on the table because the orientation of each matter
  - If you forget, tell me
  - It has scratches and marks to indicate where the screw meets the component
- Remember to speak clearly and “loud” to the microphone as there are people always walking around and other noises
- Give time for the subject to look at the keywords/paper and table
- Whenever you want to start: say the words under activate
- If asked
  - No, I cannot tell you the name of any component, Max will do that
  - You don’t need to follow the order of the paper; you can ask and interact with Max however you please
- Notes of side-question and questions task-related

After the experiment

- As I mentioned before, there is a survey online that I would like you to fill
  - You can either do it here on my computer or your phone
  - Or I can share the link to your email, and you do it later. However, I would ask you to fill it up still today as the experiment is fresh on your head and that way provides more true and reliable answers
- There is cake and soda and beer, you can take one
- Thank you for being part of my experiment

Appendix D

"Cheat Sheet"

Activate

- Call Franka

Service

- Service
- Introduction
- What kind of <u>service</u> can you provide?

Terminology

- Lab
- Steps
- How many <u>steps</u> on the assembly?
- What is …?
- Can I assemble … with …?

Grasp

- Bring
- Need
- Pass
- I need …
- Can you bring me …?

Assistance

- Help
- Assist
- Guide

+ assemble

- Can you <u>help</u> me <u>assemble</u> the phone?

Demo

- Demonstration
- Show

+ assemble

- Can you <u>show</u> me how to <u>assemble</u> the phone?