Lessons Learned from Designing and Evaluating CLAICA: A Continuously Learning AI Cognitive Assistant

Samuel Kernan Freire
Delft University of Technology
Delft, The Netherlands
s.kernanfreire@tudelft.nl

Evangelos Niforatos
Delft University of Technology
Delft, Netherlands
e.niforatos@tudelft.nl

Chaofan Wang
Delft University of Technology
Delft, The Netherlands
c.wang-16@tudelft.nl

Santiago Ruiz-Arenas
Universidad EAFIT
Medellin, Colombia
sruizare@eafit.edu.co

Mina Foosherian
BIBA - Bremer Institut für Produktion und Logistik GmbH
Bremen, Germany
fos@biba.uni-bremen.de

Stefan Wellsandt
BIBA - Bremer Institut für Produktion und Logistik GmbH
Bremen, Germany
wel@biba.uni-bremen.de

Alessandro Bozzon
Delft University of Technology
Delft, Netherlands
a.bozzon@tudelft.nl

ABSTRACT

Learning to operate a complex system, such as an agile production line, can be a daunting task. The high variability in products and frequent reconfigurations make it difficult to keep documentation up-to-date and share new knowledge amongst factory workers. We introduce CLAICA, a Continuously Learning AI Cognitive Assistant that supports workers in the aforementioned scenario. CLAICA learns from (experienced) workers, formalizes new knowledge, stores it in a knowledge base, along with contextual information, and shares it when relevant. We conducted a user study with 83 participants who performed eight knowledge exchange tasks with CLAICA, completed a survey, and provided qualitative feedback. Our results provide a deeper understanding of how prior training, context expertise, and interaction modality affect the user experience of cognitive assistants. We draw on our results to elicit design and evaluation guidelines for cognitive assistants that support knowledge exchange in fast-paced and demanding environments, such as an agile production line.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Interactive systems and tools**; **Natural language interfaces**.

KEYWORDS

cognitive assistant, chatbots, industry 5.0, human-centered AI, knowledge sharing, knowledge-based AI

IUI '23, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0106-1/23/03.
https://doi.org/10.1145/3581641.3584042

ACM Reference Format:
Samuel Kernan Freire, Evangelos Niforatos, Chaofan Wang, Santiago Ruiz-Arenas, Mina Foosherian, Stefan Wellsandt, and Alessandro Bozzon. 2023. Lessons Learned from Designing and Evaluating CLAICA: A Continuously Learning AI Cognitive Assistant. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3581641.3584042

1 INTRODUCTION

Operating a complex machine, such as a production line, can be a daunting task for an inexperienced worker, requiring extensive job training. Currently, training and knowledge sharing rely heavily on human interaction, where novices are often paired with experienced workers and receive a lot of one-on-one guidance for an extended period. However, this approach can be time-consuming and expensive, and expert workers may not have the capacity to share all relevant knowledge. Moreover, the departure of experienced workers or their retirement can result in a loss of valuable (tacit) knowledge, making it challenging to train new operators. Additionally, instruction materials is often inaccessible or outdated, inhibiting its use. As a result, training new operators is highly resource-intensive and requires a considerable amount of time [45].

Recently, intelligent Products, Systems and Services (iPSS) are transforming how technicians share knowledge in manufacturing environments [12]. However, older systems are based primarily on predefined knowledge bases, which cost a significant amount of resources to develop and maintain [14]. Previous studies attempted to automatically uncover knowledge using natural language processing (NLP) in existing maintenance reports, but many data quality problems were discovered [16]. Others came to the conclusion that technicians frequently provide informal descriptions of issues, which results in discrepancies and errors in the data; certain maintenance data, such as the actual root cause of a problem, are not always collected; and once the data is collected, it is often not used for a subsequent diagnosis [44]. Clearly, the poor quality of reports inhibits (AI-facilitated) knowledge sharing between workers, requiring a more effective and user-friendly method of recording and sharing knowledge between factory workers [26].

A cognitive assistant is a smart system designed to "augment human intellect" by endowing one with cognitive capacities beyond what is humanly possible. Cognitive assistants have been found to reduce cognitive load when making decisions and taking actions [1, 8, 31]. By continuously learning from workers' experience, cognitive assistants are able to adapt to changing physical environments, dynamic social context, and user needs [1]. Natural language mechanism and context awareness could also enable cognitive assistants to efficiently acquire high-quality knowledge shared by experienced workers and pass it on to novices [26].

We developed CLAICA (Continuously Learning Artificial Intelligence Cognitive Assistant) to support workers by providing on-the-job training and knowledge sharing. The CLAICA prototype can interact through a conversational user interface in a web browser across mobile and desktop devices. Its primary function is to learn best practices from (expert) workers and share this with others. In addition, it can recommend existing task instructions and training. The knowledge it acquires is stored in a continuously growing knowledge graph. It has several key advantages over a human mentor for novices; namely, (1) it is always available, (2) it will never forget information, and (3) it has direct and real-time access to production-line data. These advantages are critical in a modern agile production line, where a small number of human workers operate a complex connected system and many different products are produced, each of which requires specific setups. CLAICA's context awareness is powered by direct access to live production-line data (e.g., machine status, sensor data, error codes) that enable it to streamline interactions with users (e.g., it can detect what product is being produced without having to ask the user) and to help match the current situation to knowledge stored in its knowledge base. Here, we describe CLAICA's system architecture and share results from a user study with 83 participants. The study participants performed eight knowledge exchange tasks with CLAICA related to representative manufacturing activities, such as requesting cognitive assistance to solve a specific problem and sharing new machine settings. Our contributions can be summarized as follows.

(1) We showcase how a cognitive assistant can exchange knowledge with workers about operating an agile production line.
(2) We provide a deeper understanding on how context expertise, modality, and training affect the task performance, perceived workload, usability, and user experience of cognitive assistants.
(3) We elicit requirements and design guidelines for future cognitive assistants beyond the manufacturing domain.

2 BACKGROUND

2.1 Human-centered manufacturing

A key paradigm for modern production is the broad adoption of digital tools to increase, for example, productivity and sustainability. Thoben et al. concretized this paradigm with research challenges and application examples for Industry 4.0 and smart manufacturing (SM) [48]. The latter are two initiatives to systematically detail and implement this paradigm in manufacturing. While Industry 4.0 originated in Germany, SM emerged in the United States. Countries such as Japan, Korea, and China created similar initiatives. These initiatives are largely technology-focused and do not adequately account for human needs, involvement, and collaboration with AI systems [36]. The next paradigm addresses this shortcoming through human-centric manufacturing, called Industry 5.0 [36, 51]. Müller reports the findings of an expert group regarding technologies contributing to more human-centric manufacturing [35]. They include, among others, individualized human-machine interaction technologies accounting for the strengths of humans and machines. Also, using AI to assist humans in understanding causalities in complex and dynamic systems is a critical technology. Romero et al. introduced similar ideas stressing the importance of socially sustainable manufacturing [39]. They proposed the vision of Operator 4.0, which focuses on trusting and interaction-based relationships between humans and machines. Eight types of operators illustrate how technology could enhance human capabilities and skills.

Meanwhile, agile manufacturing enables product customization and on-demand production to respond quickly to customer needs and market changes. Moreover, because of the low cost of internet of things (IoT) devices, IoT devices have gained popularity in manufacturing settings to optimize business workflows and processes, improve safety and improve research and development. Nevertheless, such devices also generate unprecedented volumes of sensor data, necessitating operators' constant attention and dramatically increasing the incurred cognitive load. Interestingly, working in an agile connected production line is a real challenge for inexperienced workers. Even for experienced operators, operating an agile production line is a knowledge-intensive task that requires enormous cognitive resources [45]. For example, a single worker may need to operate and fix numerous machines along a production line simultaneously, while also (re)optimizing the setup for more than 100 different products. Therefore, systems that could reduce cognitive demands and allow quick feedback are needed to effectively support workers on a personal basis.

2.2 Cognitive Assistants

Unlike systems designed to replace humans in specific tasks (e.g., industrial robots), cognitive assistants strive to complement human abilities to accomplish complex tasks, such as aiding life-long education and machine operation [1, 2, 31]. In addition, such assistants often outperform human capacities for communication and memory in a variety of ways, such as simultaneously providing dependable and repeatable communication between numerous users [1, 2]. To achieve the aims mentioned above, cognitive assistants should support efficient human-machine interfacing via natural language processing, interpretation of gestures, perception, vision, and sounds, augmented reality to provide additional layers of information, and others [1, 2]

Of these, the most widely used interaction method for cognitive assistants is conversational agent-based natural language communication, which involves natural language understanding, generation, and dialog for implementation [1]. Conversational agents engage with people using natural language, which could perform labor-intensive jobs at low cost in a variety of industries, such as customer service, healthcare, education, e-banking, and personal assistants [34]. For example, recent work has proven conversational agents to be an effective educational tool for communicating breast cancer risk and suggested medical guidelines to women, leading to a significant increase in breast cancer genetics knowledge [52]. Furthermore, advances in context awareness make it possible to improve a conversational agent's usefulness. For example, acquiring more accurate knowledge about city locations by asking questions when users are there [10] or inferring context from user utterances to provide more relevant tourist recommendations [13]. A virtual assistant called Amber was proposed by Kimani et al. using a sensing framework that could record users' faces, speech, and app usage in order to aid users with job prioritization, provide reminders, and inhibit social media diversions [28].

Conversational agent-based cognitive assistants have also shown promise in supporting one's cognitive processing. Le and Wartschinski [31] developed LIZA, a cognitive assistant aiming at enhancing users' reasoning and decision-making abilities. By holding a conversation with test subjects to help them solve common heuristic and bias problems, LIZA helped test subjects to improve their reasoning skills and achieve significantly higher learning gain [31]. Numerous application scenarios have also been proposed in which cognitive assistants have positive cognitive effects. From education and training, such as by employing lifelong learning to retrain adult workers to meet shifting technological demands, to elderly care by facilitating interaction with those suffering by cognitive decline [1].

2.3 Cognitive Assistants in Manufacturing

Industrial applications for cognitive assistants (CA)—similar to Alexa, Google Assistant, or Siri—are an emerging research topic. These prototypes emerged in different research communities with different names (e.g., intelligent (virtual/personal) assistants, digital assistants, software robots, or simply chatbots). Cognitive assistants in manufacturing can bear significant benefits [50]. These include, for instance, central access to heterogeneous information systems, the delegation of tasks, and gaze-free and hands-free interactions during work. Furthermore, cognitive assistants can be used for training workers on-the-job [27] and adjusting machine parameters in a simulation [32].

Most of the literature regarding CAs in manufacturing focuses on knowledge and information delivery, for example, context-aware assistance (e.g., [19]), recommendations for predictive maintenance (e.g., [49]), decision support based on business analytics of shop-floor data (e.g., [3, 37]). For instance, Rodriguez et al. [38] present a mixed reality assistance system to support real-time assembly operations. They evaluated the operation context through a recognition system that determines the completion of each assembly step to derive the next instruction. Belkadi et al. [7] proposes a context-aware knowledge-based system aiming to support manufacturing operators. They integrate knowledge management, context management, and simulation management modules in order to support the decision making of the workers in real-time. As for Rodriguez et al. [38], context management gets contextual information in order to understand the current user's situation and implements simulation techniques to anticipate the effect of the worker decisions. Büttner et al. [11] implemented a hand-tracking algorithm in order to identify wrong picking actions and errors in the assembly process. Tao et al. [47] implements wearable device sensing and environmental sensing to capture worker activity in the workplace in order to guide them in the execution of their tasks, and Josifovska et al. [25] integrates a context manager module which includes a Digital twin of a human that simulates specific human abilities and preferences in order to enable assistance system adaptation. Longo et al. [33] demonstrated a digital assistant integrated into an augmented reality application to train machine operators. Their prototype provides information about safety measures, potential hazards, machine status and operations, and quality control procedures. Besides, it instructs users on lubrication, greasing, cleaning, checking and restoring hydraulic pressure or fluids and lube for maintenance.

Researchers have used several approaches to evaluate the effect of instruction delivery on workers' mental workload and its effectiveness. Funk et al. [19] evaluated the workload effect associated with the delivery of instructions on assembly work by monitoring biosignals (such as heart rate, galvanic skin response, electroencephalography, and electromyography) and indicators such as task completion time and error rates. Likewise, Kosch et al. [29] evaluated the cognitive workload produced by in situ projections during the execution of manual assembly tasks. They implemented electroencephalography (EEG) to monitor cognitive workload and compared the results with those obtained by traditional paper-based instructions. In line with the above-mentioned approaches, Funk et al. [20] proposed a standardized experiment design for evaluating the effect of interactive instructions in heterogeneous assembly tasks.

Existing literature has also explored acquiring knowledge from workers but to a significantly lesser extent. This could be partially attributed to the less immediate benefits; however, we argue that the ability to continuously learn is necessary for the long-term success of CAs on the shop floor. Fenoglio et al. [17] propose a system for capturing explicit and tacit knowledge (through best practices) from experienced workers in industrial domains. They implement a role-playing game where a virtual agent interacts with human experts and knowledge engineers in order to extract and represent knowledge in an iterative way. Despite this system providing means to capture tacit knowledge, it requires the intervention of a human agent, as they argue that "is not possible to fully capture tacit knowledge (usually nonverbal and unexpressed) with a purely algorithmic approach". Likewise, Soliman and Vanharanta [46] suggests a model for knowledge creation and retention through artificial intelligence. However, there is no practical application of the model reported in literature, and Hoerner et al. [22] propose a digital assistance system to support operator troubleshooting processes on the shop floor. For this purpose, a method to capture and structure expert's tacit knowledge was developed. However, this method is not executed by the digital assistant. It is performed by human experts who are in charge of extracting and representing the obtain knowledge and to deliver it as input to the system.

However, and to the best of our knowledge, no existing solution captures (tacit) knowledge from experienced operators with the purpose to structure it, store it, and re-share it with novices in real time and on the shop floor. CLAICA eases the learning curve for novice operators by serving as a dialectic mediator among novices and current or past experienced operators. Even for experienced personnel, CLAICA provides multiple opportunities to rehearse and test acquired knowledge and skills, while constantly discovering and formalizing new knowledge or knowledge that has been overlooked (e.g., tacit knowledge).

3 CLAICA

The primary goal of CLAICA is to enable knowledge exchange between shop floor workers. It continuously learns by acquiring knowledge from workers through dialectic interactions, allowing it to efficiently share up-to-date knowledge. In addition, it can recommend existing work instructions and perform information retrieval tasks. Furthermore, workers can provide feedback on the knowledge they receive to improve recommendations over time. What sets CLAICA apart from the state-of-the-art is its ability to efficiently acquire knowledge from workers on the shop floor without human involvement and store it in a knowledge graph along with contextual information. However, knowledge managers may still be needed to perform quality control by approving, reviewing, and removing elements of knowledge. Ultimately, CLAICA aims to reduce the burden on knowledge managers and improve knowledge sharing.

3.1 Co-designing CLAICA

CLAICA was developed in close collaboration with an industrial company, a detergent producer. Their ambitions for the assistant are to result in faster training of new operators and higher overall equipment effectiveness (OEE) [15], that is, the percentage of manufacturing time that is truly productive. During the early design phase of CLAICA, we conducted semi-structured interviews and focus groups with operators and management at two detergent factories. We explored three main topics, namely what opportunities are there to support machine operators with a cognitive assistant, what wishes do factory employees have regarding the (interactive) capabilities of the assistant, and what challenges the assistant might face from a user acceptance perspective. We interviewed six machine operators, two maintenance technicians, two shift leaders, four engineers, and one factory director. We pinpointed the following opportunities: (1) identify the best practices of operators, elicit this knowledge, and share it with others, (2) make existing work instructions more easily accessible, (3) help operators identify the root cause of problems, (4) provide operators with access to machine data anywhere on the production line, and (5) create high-quality issue reports. The operators were primarily interested in receiving ubiquitous access to machine data, while their supervisors were also concerned with facilitating knowledge sharing among operators, providing access to instructions, and creating better issue reports. The shift leaders suggested that operators, even experts, could benefit from suggestions; however, the operators themselves disagreed. We noted that the operators were very proud of their skills and proud that they did not need instruction material. However, they thought that novice operators could benefit from easy access to up-to-date documentation. Existing documentation resides in paper and digital form, but is poorly structured and frequently outdated. As such, novice operators are trained almost exclusively by experienced operators on-the-job. Due to the complexities of operating an agile production line, this process is lengthy and, therefore, costly. Furthermore, it is risky from the company's perspective, as a lot of valuable (tacit) knowledge will be lost when experienced operators leave.

We collected 100 issue descriptions on one of the production lines over three days. We asked the operators to verbally describe the location, symptoms, and cause of each issue as it occurred. The operators used a mono headset during data collection (see Fig. 1).

Figure 1: An operator describing a problem with the production line

We analyzed the resulting reports to identify opportunities and challenges for CLAICA. We observed that operators used different terms and acronyms when describing the same machines (e.g., depa, palletizer, depalletizer). In addition, the production line breaks down frequently (up to 30 times in an eight-hour shift) and the operators are under intensive time pressure. As a result, many problems go undocumented. Once we built a working prototype, we presented it to the operators (see Fig. 2) to elicit further feedback.

Figure 2: An operator using CLAICA at the production line

The insights from the interviews, the focus groups, and the collection of issue descriptions resulted in the following design requirements for CLAICA.

- It must provide accurate cognitive support to novices
- It must interact efficiently and reliably

- It must support user feedback
- It must continuously learn from its (expert) users
- There must be incentives for (expert) users to share their knowledge (this responsibility is shared with the company)
- It must reduce the reliance on expert operators for training and supporting novices
- It must be transparent about what data it collects and how it provides recommendations
- It must be able to handle divergent phrasing and misunderstandings gracefully

3.2 System

3.2.1 Capabilities. CLAICA employs a conversational user interface (CUI) as its primary means of interacting with users, a knowledge graph for storing its knowledge, and a cognitive engine for processing and integrating all the information streams (see Fig. 3). CLAICA can collect information about the context autonomously (e.g., machine states). As such, the context awareness provided by the live production data enables CLAICA to streamline interaction, minimize its duration, and input burden (e.g., by auto-filling some of the information for the user).

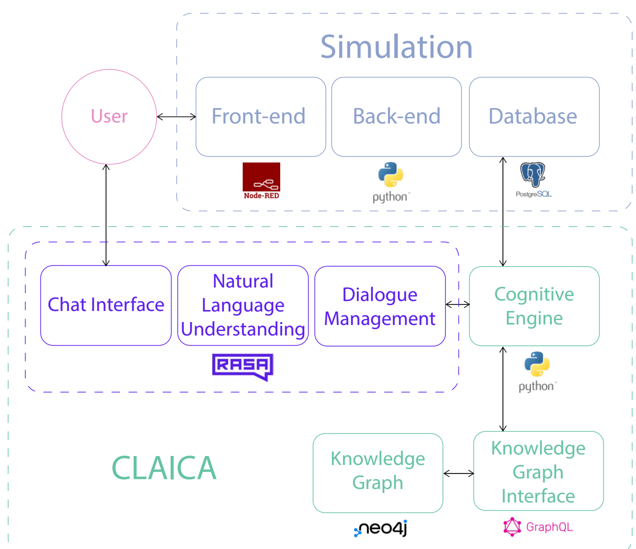

Figure 3: CLAICA's architecture (below) and the simulation (above)

In its current state, CLAICA can continuously learn from operators in the following scenarios.

(1) *A worker shares how they diagnosed and solved an issue.* Upon classifying the worker's intent to provide an issue report, CLAICA will start requesting information from the user. Each issue report is split into three sections, namely symptoms, cause, and solution. CLAICA asks the worker to provide a description of each of these in text. While processing the worker's response, CLAICA attempts to extract several named entities such as any machine components (e.g., nozzle 23), machine component states (e.g., overheated), product components (e.g., label) and product component states (e.g., crooked), error codes, and worker actions. If CLAICA is unable to extract at least one machine component associated with the symptom of the issue, it asks the worker to specify one manually. Furthermore, it uses its context awareness to suggest additional entries as buttons in the CUI, such as machine error codes. Finally, the workers' descriptions, extracted entities, and context information are stored in the knowledge graph.

(2) *A worker shares machine settings for a product.* In this scenario, CLAICA uses its machine integration to collect some of the information automatically, for example, which product is being produced and with what settings. However, it still asks the worker for confirmation and the option to specify something manually. Additionally, CLAICA asks the user for comments on the settings, for example, if they are safe or risky. In turn, this information is stored in the knowledge graph.

(3) *A worker rates existing machine settings.* Upon receiving recommended machine settings from CLAICA, workers are encouraged to rate them from 1 to 5. CLAICA takes these ratings into account when recommending the settings to other works. In addition, a rating of 1 will automatically triger the settings to be retracted and flagged for review by a knowledge manager.

In addition, CLAICA can provide knowledge in the following scenarios.

(1) *A worker asks for help.* When a worker asks for help, CLAICA asks them to describe the situation. It attempts to extract several named entities, such as machine component(s) and their states. In turn, it uses these entities to search its knowledge graph for the most relevant existing solutions and/or documents and recommends these to the worker.

(2) *A worker asks for machine settings for a product.* When the user requests machine settings, CLAICA checks which product is currently being produced and asks for user confirmation. Then, it searches its knowledge base for the highest rated settings for that product and presents it (see Fig. 4).

(3) *A worker asks how to perform a specific task.* CLAICA's knowledge base also includes existing documentation on how to perform several tasks. The worker can ask for them explicitly, for example, "how do I perform a prerun?" (see Fig. 5).

3.2.2 Assistant Framework. We use an open source conversational AI assistant framework, Rasa[1], to build CLAICA (see Fig. 3. We use Rasa due to its flexibility and performance; it can easily be customized for specific use cases and connected to additional components, such as knowledge bases. This includes adding custom Python scripts (Cognitive Engine), adjusting the NLP pipeline, and using custom entity extractors, such as Duckling[2] (e.g., to extract pressure values from user utterances). The natural language understanding (NLU), dialogue management uses several base features that have universally applicable intents and conversation patterns, for example, greeting and restarting the assistant. Additionally, we

[1]https://rasa.com/—last accessed February 17, 2023
[2]https://duckling.wit.ai/—last accessed February 17, 2023

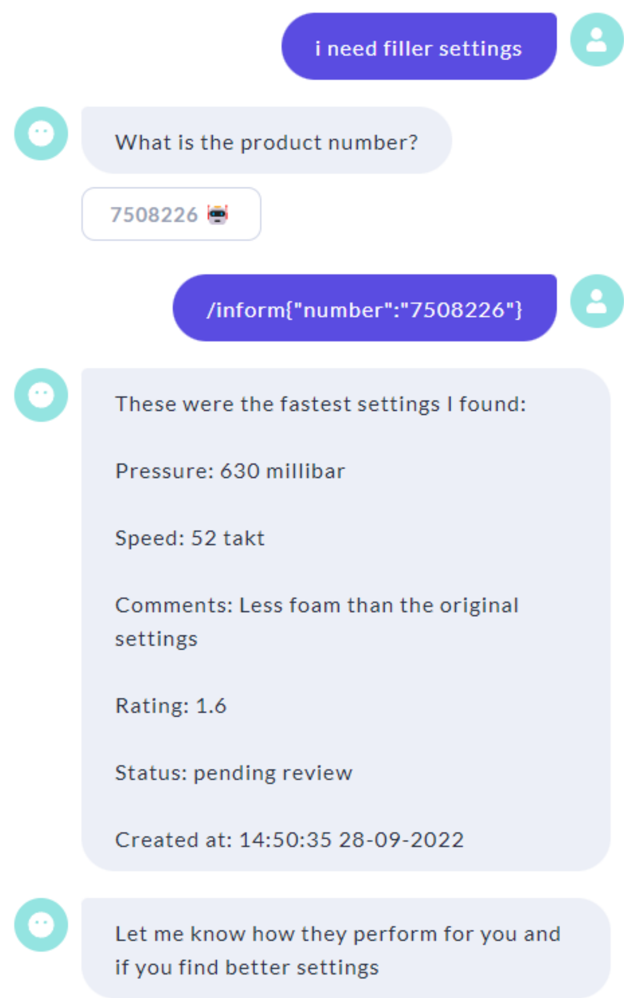

Figure 4: A user requests filling machine settings for a product

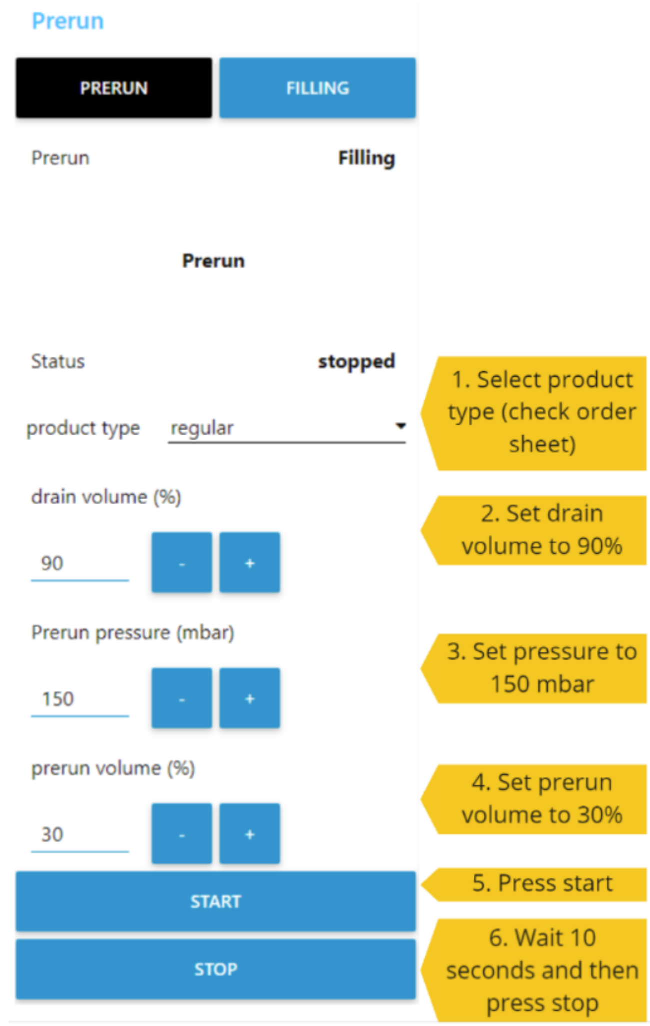

Figure 5: Task instructions for a prerun

have added features that contain specific training data, domain descriptions, and custom actions for CLAICA. The Cognitive Engine, also known as the custom actions server, is the place where all incoming streams of information, such as machine data, user-provided information, and knowledge base, are handled, analyzed, and responded to. It also contains the logic to ensure all the relevant information is collected from user's during knowledge acquisition (e.g., during an issue, CLAICA needs to collect the machine component ID, symptom, cause and applied solution). We run a Rasa X server that supports a browser-based chat interface (see Fig. 4). It also provides a browser-based interface to review user conversations and apply improvements.

3.2.3 Continuously Growing Knowledge Base. The knowledge base is a Neo4j[3] knowledge graph, also known as a semantic web. It can be queried by the assistant using GraphQL[4] resolvers. A knowledge graph is a type of data representation that uses nodes and edges to map relationships and information between entities. It helps machines, like CLAICA, to "understand" the meaning and context of data, providing a more structured and interconnected view of information. For CLAICA, we define a description of an issue symptom as an "event" node that will have a "caused by" relationship to another event that describes the cause. In turn, the symptoms event will have a "solved by" relationship to a solution node. Each of the event nodes can have relationships with other entities, such as machine components and their states. This gives CLAICA a robust way to find relevant information (e.g., a solution), to a situation described by the worker.

3.2.4 Simulating the Production Line. Testing an intervention, such as a cognitive assistant, in a detergent factory and in situ is challenging due to the potential for dangerous and costly operating errors. Problems can arise from incorrect recommendations and the cognitive load or distractions imposed by interactions with the assistant. Furthermore, since the target audience for the assistant is

[3]https://neo4j.com/—last accessed February 17, 2023
[4]https://graphql.org/—last accessed February 17, 2023

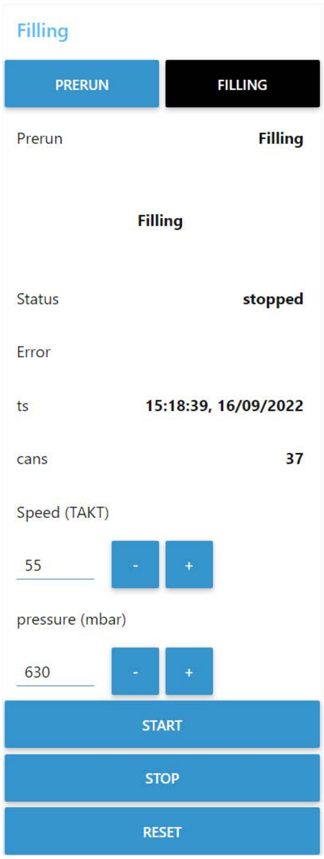

Figure 6: The simulated user interface for the filling machine

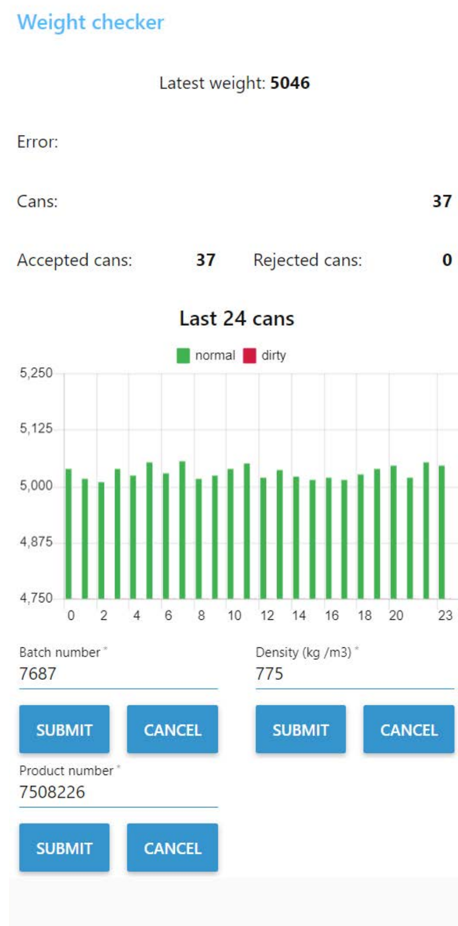

Figure 7: The simulated user interface for the weight checker machine

new employees, the pool of available test participants is small and unreliable. Therefore, we created a detergent factory simulation to evaluate the assistant in the lab before introducing it in the wild. The simulation features several graphical user interface (GUI)s that connect to a model in the back-end. These include GUIs where users can control the most important machines on the production line, namely the detergent container filling machine (Figs. 6) and the detergent container weight checker machine (Fig. 7).

The front-end was built using Node-RED[5], which connects to a PostgreSQL[6] database and machine models written in Python[7]. It supports several simple tasks related to the preparation of machines for production, as well as more complex optimization and problem solving tasks. The GUIs are browser-based and designed to be accessed on a tablet or laptop. To simulate a real factory scenario as closely as possible, we examined the set-up found in a detergent factory and modeled our lab set-up on it. For example, for the GUIs, we matched the information displayed, the user controls, their physical height, and the distance between them. The production line simulator was validated through several focus group sessions with factory employees ranging from shop floor workers to process improvement engineers.

[5]https://nodered.org/—last accessed February 17, 2023
[6]https://www.postgresql.org/—last accessed February 17, 2023
[7]https://www.python.org/—last accessed February 17, 2023

4 STUDY AND METHODOLOGY

We conducted a between-subjects user study with 83 participants to evaluate user performance, usability, user experience and perceived workload when interacting with CLAICA. The user study took approximately one hour and consisted of the following three parts: a demonstration of the (simulated) production line, a demonstration of the assistant, and a user test. The user test involved performing eight knowledge exchange tasks with CLAICA. These included information retrieval (e.g., "Find instructions on how to perform a prerun"), knowledge sharing, and requesting cognitive support. Note that the participants only interacted with the assistant; they did not have to perform any actions with the (simulated) production line. Participants then completed a survey, including a NASA Task Load Index questionnaire (NASA-TLX) [21], a User Experience Questionnaire (UEQ) [43], and a System Usabilty Scale (SUS) questionnaire [6]. At the end of the study, we posed several open-ended questions to initiate a discussion and obtain qualitative feedback from our participants.

4.1 Research Questions

In this user study, we sought to answer the following Research Questions (RQs):

RQ1 *How does prior training affect user performance, perceived workload, user experience and perceived usability when using CLAICA?*
When introducing a new tool or system like CLAICA to end users, its introduction will probably include some training on what it can do and how to use it. Managers would like to set expectations and manage the change process with care. To ensure adoption, managers should highlight the perceived usefulness of a system, especially the improvements that it may bring to efficiency [9]. Furthermore, training could help users become more effective with the new system. We plan to provide training to end users when CLAICA is deployed in situ; therefore, we would like to evaluate the effect of providing training or not ("trained" vs. "untrained" participants). As we expect users will need time to learn how to interact with CLAICA, we think that the trained group will be able to interact more efficiently with CLAICA and, therefore, have a better use experience.

RQ2 *How does context expertise affect user performance, perceived workload, user experience and perceived usability when using CLAICA?*
Due to the challenges of conducting studies with real end-users, many researchers opt to recruit participants locally (e.g., students or colleagues). The same is true for companies developing new products - access to end users can be difficult for many reasons, for example, confidentiality [42]. Therefore, it is valuable to understand how context expertise affects the user experience of AI assistants such as CLAICA ("worker" vs. "layman" participants). We do not expect context expertise will help users interact with CLAICA more efficiently; however, we think that it may affect the subjective user experience as they understand the implications of CLAICA in the context better.

RQ3 *How does the modality of interaction with CLAICA affect user performance, perceived workload, user experience and perceived usability?*
From the formative study interviews with end users, we learn that they currently use a fixed computer to access instructions and create issue reports. However, the production line can be long, so having to walk back to the computer takes time and may be a barrier to its use. Furthermore, they may not remember all the details of an instruction when they return to the relevant machine on the production line. Regardless, the laptop's larger screen and keyboard might help users retrieve and provide information faster and we use it as a baseline to compare against ("smartphone" vs. "laptop"). Previous research has shown that typing on computers is faster, but that smartphones can benefit more from suggested words [40]. Therefore, we expect that the laptop group will be able to perform tasks faster and have a better user experience.

4.2 Participants

We recruited 83 participants (44 male, 34 female, three "other" and four "preferred not to say") with age ranges from 18 to 64. Most of the participants ($N = 50$) were 17–29 years old, followed by the 30–39 age bracket ($N = 16$). All participants were able to communicate clearly in written and spoken English. The participants were recruited from the following four groups: factory employees ($N = 12$), external human computer interaction (HCI) researchers ($N = 16$), local HCI researchers ($N = 18$) and master's students ($N = 37$).

4.3 Procedure

Before commencing a user trial, participants were asked to read and sign an informed consent form that had been approved by an ethical board. Participants were randomly but equally split between the "smartphone" group and the "laptop" group. Next, participants were shown a short video introduction to the context of manufacturing and AI assistants. We then presented a three-minute video to introduce the (simulated) production line. The first part of the video introduced the detergent production line, a few of the worker's primary tasks and the machines we included in the simulation (filling machine and weight checker). Then, participants were shown how we simulated the filling machine (see Fig. 6) and the weight checker (see Fig. 7). Following the first video, the participants were given three minutes to write down comments on the realism of the simulation and how it could be used and improved. Participants were asked to write their comments as digital post-it notes. Then, the participants from the "trained" group were shown a five-minute training video demonstrating the capabilities of the assistant and how to use it. Again, they were given three minutes to write their comments. The "untrained" group, who performed the study at another time and place, was asked to write down comments on a related topic. Therefore, we ensure a comparable total experiment time and workload. The participants in the "smartphone" group were asked to interact with CLAICA using a web browser on their smartphone, while the participants in the "laptop" group used a web browser on their laptop. Then 10 minutes were given to complete eight assigned knowledge exchange tasks with CLAICA. Participants were asked to interact naturally in English. Immediately after completion, they were instructed to complete several surveys, namely NASA task load index (NASA-TLX), system usability scale (SUS), user experience questionnaire (UEQ), demographics (age, gender, occupation, years of experience in occupation), and personal factors related to chatbot experience. After completing the survey, we moderated a 10-minute discussion on their experience using the assistant and opportunities for improvements.

4.4 Measures

We measured the following quantitative objective variables: Task completion time, Task completion rate (yes/no), Conversation turns between user and CLAICA, User utterance length (words), Conversation breakdowns (when CLAICA cannot classify the user's intent), and User typing errors. The task completion time was automatically calculated by the Qualtrics[8] survey using the time difference between the first and last click on the task instruction page. The chat

[8]https://www.qualtrics.com/

logs were extracted from an SQL Tracker Store [9] using SQL queries and a Python script. We then semi-automatically extracted participant IDs to match the chats to the survey responses. Quantitative objective variables related to the interactions were automatically calculated using a Python script. We used the Enchant [10] library to identify user typos and checked the assistant's failure response "/restart" to count conversation breakdowns. For the sake of brevity, we omit these metrics in this paper.

We measured the following quantitative subjective variables: workload using the NASA-TLX, usability using the SUS, and user experience using the UEQ. We collected the following demographics and personal factors: Age, Gender, Occupation, Years of experience in occupation. In addition, we collected the following self-reported factors related to chatbots and technology: prior experience with chatbots and tech-savviness. We asked the following questions with likert scales as a response: "*I am familiar with chatbot technologies*," "*I use chatbots frequently*," "*I consider myself an advanced technology user*," and "*I am eager to try new technologies*," and "*I am good at solving technical problems*." The first four of these questions have been used in prior work on chatbot breakdown strategies [4, 18].

4.5 Qualitative Analysis

The last part of the user study consisted of asking the participants several open-ended questions to gather detailed insights from users about their experience executing tasks with the assistant. The questions were as follows: "How did you experience the assistant?", "What worked and what didn't?" and "How could it be improved?". After completing the tasks and filling our the survey, the participants were given 5 minutes to write digital post-it notes on a Miro board [11]. A total of 148 text entries were collected. We performed a content analysis to investigate the user responses. Lazar et al. [30] define content analysis as "*the process of developing a representative description of text or other unstructured input*". We used two subjective coders, who read the data multiple times and then followed a mixture of emergent coding and a priori coding to create the key categories. The first and second level are based on a priori coding to clean and sub-categorize the data according to this study's goals. Comments were not processed further if they were not directly relevant to answer the research questions.

5 RESULTS

To decide on our statistical methods, we first performed all the necessary pre-tests, such as Shapiro-Wilk tests of normality and Levene's tests of homogeneity of variance. We omit the pre-tests for brevity. Depending on the statistical test at hand, we report averages and standard deviations (parametric), or median values (non-parametric). We started our analyses by computing correlations between personal factors (e.g., demographics and prior experience with chatbots), as they can confound the relationship between the independent and dependent variables. We found the following two significant correlations: age is positively correlated with SUS score ($r(81) = .238, p < .05$) and self-reported technical problem solving skills are positively correlated with completion rate ($r(83) = .308, p < .05$). These correlations indicate that (a) the older our participants were, the higher their usability ratings about CLAICA, and (b) the higher the self-reported problem-solving skills, the more probable it is for our participants to successfully complete a knowledge exchange task with CLAICA. We then investigated if there were any significant differences between the median age and technical problem solving skills of the groups. A Mann-Whitney U test showed that there is a significant difference in the median age category between the layman group (Mdn = 1) and worker group ($Mdn = 4\ /mid\ U = 80.500, p < .001$), between the untrained group (Mdn = 2) and the trained group ($Mdn = 1 \mid U = 311.500, p < .05$), but no significant difference between the median age category of the smartphone group (Mdn = 1) and laptop group ($Mdn = 1 \mid U = 775.500, p = .746$). Regarding self-reported technical problem-solving skills, a series of Mann-Whitney U tests did not reveal any significant differences between the laymen (Mdn = 4) and the workers ($Mdn = 4$)($U = 379, p = .519$), or between the untrained (Mdn = 4) and the trained group ($Mdn = 4$)($U = 391.500, p = .336$), or between the smartphone (Mdn = 4) and the laptop group ($Mdn = 4$)($U = 791.500, p = .596$).

5.1 Effects of Prior Training (RQ1)

At first, we investigated whether receiving instructions (training) on what a cognitive assistant can do would have any effect on user performance, expressed as overall task completion times and task completion rates. However, a non-parametric Mann-Whitney U test displayed no significant difference in the median overall task completion times between untrained (Mdn = 423.463) and trained participants ($Mdn = 370.165 \mid U = 302, p = .099$). Similarly, a non-parametric Mann-Whitney U test displayed no significant difference in the median overall task completion rate between untrained (Mdn = 6.5) and trained participants ($Mdn = 7 \mid U = 389.500, p = .332$). These findings indicate that **prior training on how to use CLAICA did not affect significantly task performance when using CLAICA (RQ1).**

Next, we investigated if prior training on CLAICA impacts perceived workload (NASA-TLX) and usability (SUS). An independent-samples t-test revealed a significant difference in the average NASA-TLX scores between the untrained ($M = 35.651, SD = 18.261$) and the trained group ($M = 46.058, SD = 13.080 \mid t(69) = -2.590, p < .05$)(see Fig. 8).

Similarly, an independent-samples t-test revealed also a significant difference in the average SUS scores between the untrained ($M = 59.559, SD = 19.946$) and the trained group ($M = 45.046, SD = 17.324 \mid t(69) = (2.905, p < .01$, see Fig. 9). Contrary to our expectations, **the trained group reported significantly higher workload and significantly lower usability than the untrained group did** after completing knowledge exchange tasks with CLAICA (RQ1).

Last, we explored whether previous training influences user experience (UX) as reported by our participants using the six dimensions of the UEQ. A series of independent-samples t-tests revealed significant differences between the two groups reflected in the average scores of **attractiveness** (untrained: $M = .392, SD = .854$ vs. trained: $M = -.133, SD = .848 \mid t(69) = (2.905, p < .05$), **perspicuity** (untrained: $M = .853, SD = 1.284$ vs. trained:

[9]https://rasa.com/docs/rasa/tracker-stores/
[10]https://abiword.github.io/enchant/
[11]https://miro.com/whiteboard/

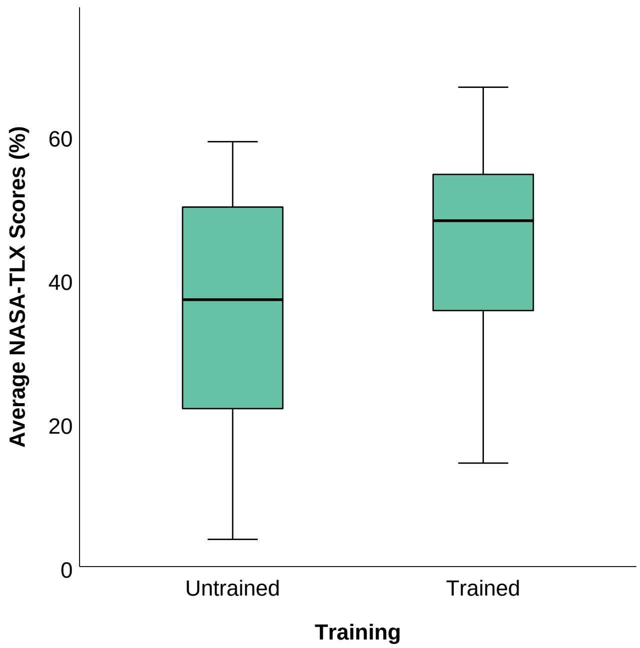

Figure 8: NASA-TLX score versus training

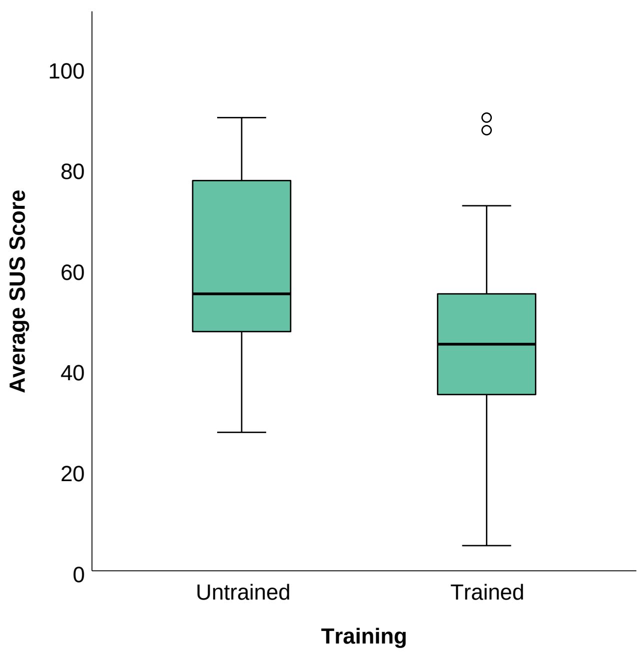

Figure 9: SUS score versus training

$M = -.133, SD = .848 \mid t(69) = (2.905, p < .01)$, **efficiency** (untrained: $M = .7647, SD = 1.191$ vs. trained: $M = .046, SD = .874 \mid t(69) = (2.905, p < .01)$, and **dependability** (untrained: $M = .471, SD = .824$ vs. trained: $M = .130, SD = .867 \mid t(69) = (1.430, p = .157)$, but not in the average scores of **stimulation** (untrained: $M = .088, SD = .824$ vs. trained: $M = -.301, SD = .822 \mid t(69) = (1.702, p = .093)$ and **novelty** (untrained: $M = .1765, SD = .822$ vs. trained: $M = -.181, SD = .972 \mid t(69) = (1.395, p = .167)$. For an overview, see Fig. 10. These findings indicate that **prior training in using CLAICA impacted negatively UX notions such as perceived attractiveness, perspicuity, efficiency, and dependability, but did not affect perceived stimulation and novelty (RQ1)**.

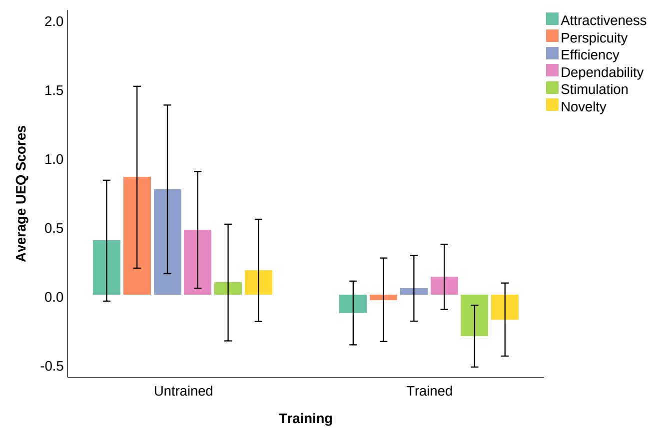

Figure 10: UEQ scores versus training

5.2 Effects of Context Expertise (RQ2)

Next, we investigate if and how context expertise—being a **worker** vs. being a **layman**—affects task performance, perceived workload, usability, and UX when using a cognitive assistant such as CLAICA. Two non-parametric Mann-Whitney U test revealed significant differences in the median overall **task completion times** (worker: $Mdn = 502.985$ vs. layman: $Mdn = 380.775 \mid U = 261, p < .05$) and in the median overall **task completion rates** (worker: $Mdn = 7.5$ vs. layman: $Mdn = 7 \mid U = 333, p = .211$) between the two context-expertise groups (see Fig. 11). Interestingly, **laymen performed better than workers in knowledge exchange tasks when using CLAICA (RQ2)**. A follow-up non-parametric Mann-Whitney U test did not reveal significant differences in the median NASA-TLX scores between worker ($Mdn = 37.88$) and layman ($Mdn = 45.45$) participants ($U = 334.5, p = .236$).

However, an independent-samples t-test revealed a significant difference in the average SUS scores between worker ($M = 74.167, SD = 14.706$) and layman ($M = 48.521, SD = 18.896$) participants ($t(81) = -4.470, p < .001$)(see Fig. 12). The mean SUS score from the worker and layman groups is equivalent to a "good" and "poor"/"ok" grade, respectively [5] or the 70th and 10th percentile, respectively [41]. Note that these equivalent scores predate widespread user testing of chatbots and may be unreliable [24]. These findings suggest that **laymen did not experience CLAICA as more cognitive demanding than workers in completing knowledge exchange tasks. However, workers rated CLAICA significantly higher than laymen in terms of usability. (RQ2).**

Finally, we investigated whether and how context expertise has an impact on user experience (UX) as reported by our participants

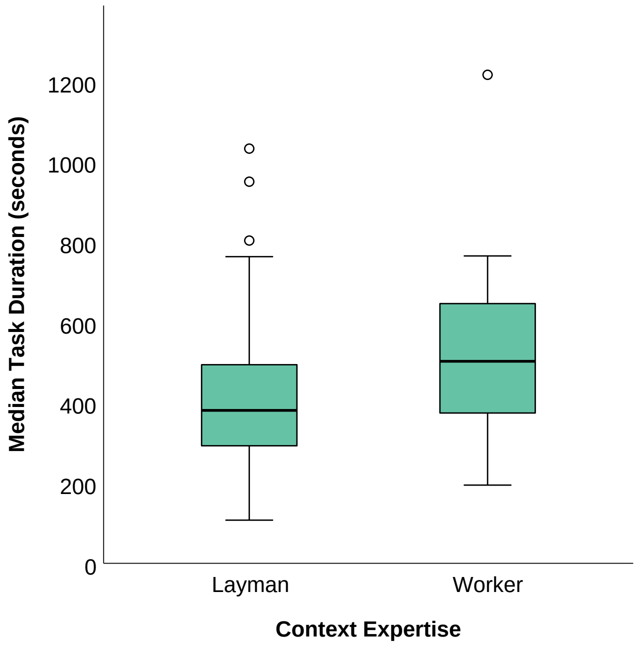

Figure 11: Task duration versus context expertise

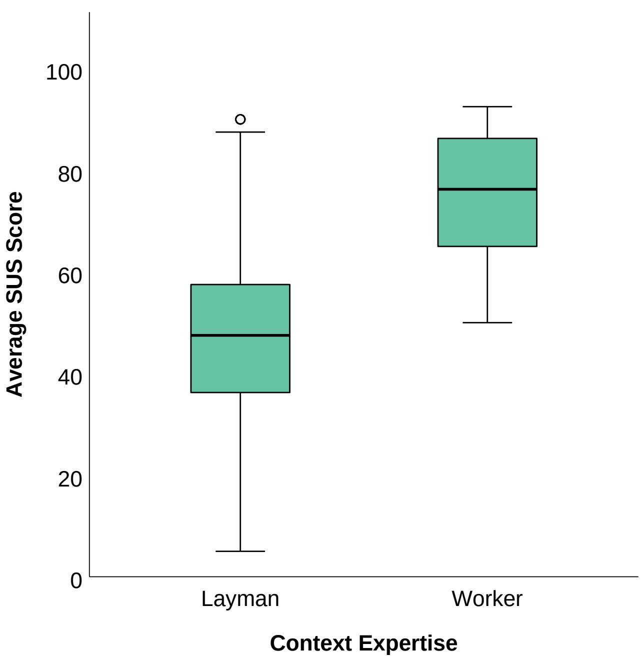

Figure 12: SUS score versus context expertise

using the six dimensions of the UEQ. A series of independent-samples t-tests revealed significant differences between the two groups reflected in the average scores of **attractiveness** (worker: $M = 1.305, SD = .887$ vs. layman: $M = -.007, SD = .872 \mid t(81) = -4.808, p < .001$), **perspicuity** (worker: $M = 1.458, SD = .851$ vs. layman: $M = .176, SD = 1.207 \mid t(81) = -3.525, p < .001$), **efficiency** (worker: $M = 1.416, SD = .807$ vs. layman: $M = .218, SD = .999 \mid t(81) = -3.937, p < .001$), **dependability** (worker: $M = 1.208, SD = .744$ vs. layman: $M = .211, SD = .863 \mid t(81) = -3.765, p < .001$), **stimulation** (worker: $M = 1.333, SD = .807$ vs. layman: $M = -.207, SD = .833 \mid t(81) = -5.949, p < .001$), and **novelty** (worker: $M = .854, SD = 1.245$ vs. layman: $M = -.095, SD = .926 \mid t(81) = -3.117, p < .05$). For an overview, see Fig. 13. **These findings indicate that context expertise plays a substantial role in UX when using a cognitive assistant such as CLAICA (RQ2).**

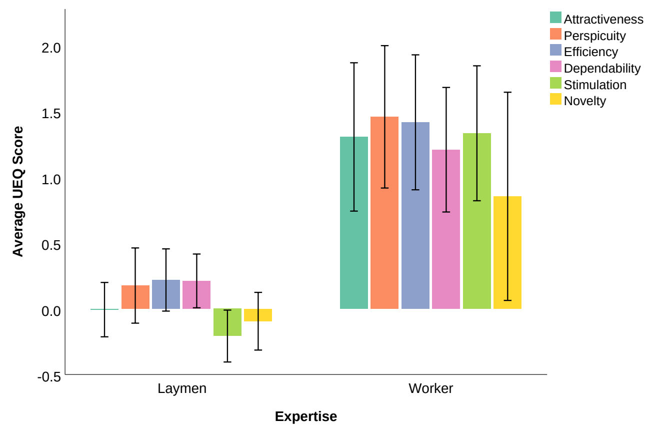

Figure 13: UEQ aspects versus context expertise

5.3 Effects of Interaction Modality (RQ3)

Then, we investigated if and how the interaction modality participants used (smartphone vs. laptop) had any impact on user performance, perceived workload, user experience and perceived usability. The aim here is to compare between a popular interaction modality (smartphone) and an established one (laptop) in how they influence information exchange with cognitive assistants such as CLAICA. From the outset, two non-parametric Mann-Whitney U tests displayed no significant differences in the median overall **task completion times** (smartphone: $Mdn = 385.981$ vs. laptop: $Mdn = 379.522$ /mid $U = 765, p = .827$) and in the median overall **task completion rates** (smartphone: $Mdn = 7$ vs. laptop: $Mdn = 7 \mid U = 736.500, p = .296$). These findings indicate the **interaction modality (smarthphone vs. laptop) had no effect on task performance when executing knowledge exchange tasks with a cognitive assistant such as CLAICA (RQ3).** In the same guise, two follow-up independent samples t-tests did not unveil any significant differences in the average NASA-TLX (smartphone: $M = 41.646, SD = 17.420$ vs. laptop: $M = 43.859, SD = 12.663 \mid t(61.410) = -.643, p = .523$) and in the average SUS (smartphone: $M = 53.333, SD = 25.663$ vs. laptop: $M = 51.383, SD = 15.462 \mid t(53.994) = .403, p = .688$) scores between the two interaction modality groups. These findings suggest that **interaction modality (smartphone vs. laptop) does not bear a substantial impact on perceived workload and reported usability when using a**

cognitive assistant such as CLAICA to perform knowledge exchange tasks (RQ3).

Last but not least, we inquired into if and how interaction modality has influences UX as reported by our participants using the six dimensions of the UEQ. However, a series of independent-samples t-tests and Mann-Whitney U tests displayed no significant differences between the two groups as reflected in the average and median scores of **attractiveness** (smartphone: $M = .153, SD = 1.121$ vs. laptop: $M = .206, SD = .880 \mid t(81) = -.241, p = .810$), **perspicuity** (smartphone: $M = .361, SD = 1.495$ vs. laptop: $M = .362, SD = 1.029 \mid t(59.110) = -.002, p = .998$), **efficiency** (smartphone: $M = .542, SD = 1.157$ vs. laptop: $M = .2766, SD = .973 \mid t(81) = 1.133, p = .260$), **dependability** (smartphone: $M = .333, SD = 1.023$ vs. laptop: $M = .372, SD = .832 \mid t(81) = -.192, p = .849$), **stimulation** (smartphone: $Mdn = 0.250$ vs. laptop: $Mdn = 0.0 \mid U = 821, p = .818$), and **novelty** (smartphone: $M = -.208, SD = .992$ vs. laptop: $M = .234, SD = 1.021 \mid t(81) = -1.981, p = .051$). These findings show that **interaction modality (smartphone vs. laptop) has no effect on UX when using a cognitive assistant such as CLAICA to perform knowledge exchange tasks (RQ3).**

5.4 Qualitative Insights

We conducted a content analysis to create a structured representation of the comments received from our participants. At the first level, two coding categories were created: Conversational AI perspective and user interface (UI) perspective. At the second level, for the conversational AI perspective, text inputs were categorized according to the main technical components of the assistant which was built using the Rasa framework; such as NLU component, dialog management component, and external data bases. Two other categories include the assistant's functionalities and UX directly bound to the assistant. Comments related to the UI perspective were categorized into second-level aspects such as interaction modes, input modes, accessibility, general UX bound with the UI interactions, and message visualization. The comments were then further refined through emergent coding, as shown in Fig. 14 and 15.

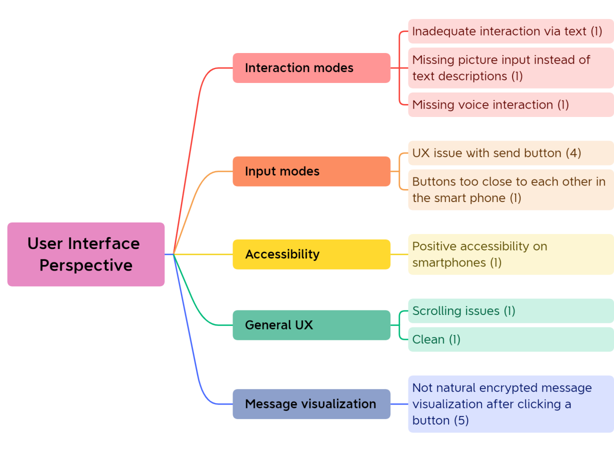

Figure 14: Results of the content analysis performed on participants' comments about the user interface (UI) perspective

We identified 34/148 comments on the limited abilities of the NLU in understanding divergent phrasing resulting in conversational breakdown (NLU fallback). Four participants ($N = 4$) found the applied repair strategy and restart option to recover from the breakdown ineffective. Five participants ($N = 5$) wished they could talk to the assistant by using keywords instead of complete sentences. Participants expressed interest in the prospect of using CLAICA; One worker excitably asked "*when will it be able available in the app store*"(W13) and "*the ease of use and quick access to information are great*" (W13). Seven participants ($N = 7$) voiced concerns regarding safeguarding the quality of acquired knowledge and how to select the most appropriate recommendations as the knowledge base grows. Eleven participants ($N = 11$) stated that CLAICA might not be realistic or complex enough to be adopted in the manufacturing use case. One participant ($N = 1$) said "*seemed to work well for the prescribed tasks on the whole. Hard to relate this to real 'life on the floor', where presumably numerous kinds of errors (not all related to the machine) might occur.*" (L103) Another participant stated that "*worker should fix the issue or contact a supervisor if they cannot fix it, before they talk with chatbot*" (L13).

6 DISCUSSION

One of CLAICA's major strengths is its ability to continuously learn from (expert) users; however, there are several hurdles that CLAICA must overcome to be effective. Many of these were raised by the participants; namely, (1) that its knowledge base was currently too limited, (2) how to ensure (long-term) accuracy, and (3) how to prevent information overload. As CLAICA can learn, its knowledge base does not need to be (quasi)complete before introduction; however, it would probably promote user acceptance if it could provide some level of assistance straight away. Therefore, we populated it with information about 100 previous problems that were solved by workers and a few product settings. We considered this sufficient for this user study; however, we aim to collect more before deploying it on a production line. We have already introduced a rating and approval mechanism that allows continuous feedback and flagging of inaccurate information. We may also consider prompting expert users to review information that CLAICA suspects is outdated (e.g., by its age or low rating). To tackle information overload, we rank the information in the knowledge base by several factors (e.g., rating and age) and only display the top three to the users. However, we plan to use similarity algorithms on the knowledge graph to make this ranking more intelligent, for example, by suggesting knowledge about similar problems. Based on our experience designing and evaluating CLAICA, we created a set of design guidelines for future cognitive assistants.

6.1 Deprioritize training when users are technologically adept

The main advantage of conversational agents lies in that user interaction is greatly facilitated by the use of natural language. That said, we theorized that to perform knowledge exchange tasks with a cognitive assistant such as CLAICA, prior training should be necessary to fully realize its potential. However, our results showed that the trained group found the workload (NASA-TLX) to be higher, usability (SUS) lower, and attractiveness, perspicuity, and efficiency

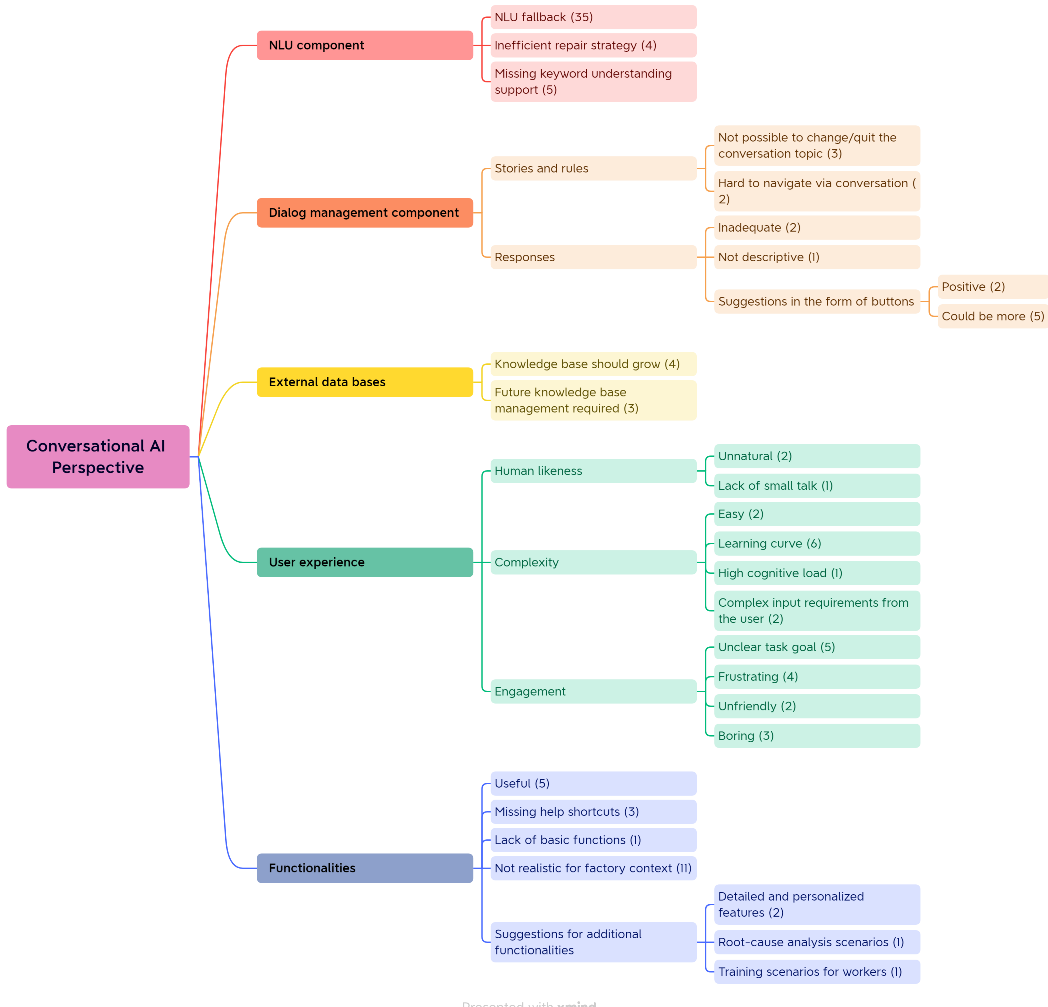

Figure 15: Results of content analysis performed on participants' comments about the conversational AI perspective

UEQ also lower. All other measures were found to be not significantly different between the trained and untrained groups. This is remarkable, as providing training for the assistant did not result in higher overall task completion rates or faster task completion time. Perhaps users that received training tried to remember what they had seen in the training session. In turn, this may have resulted in experiencing a higher workload, lower perceived usability and worse UX than the untrained participants did. Alternatively, the training may have set their expectations too high and they were disappointed when CLAICA failed to understand their utterances, resulting in inefficient and unclear exchanges. Future work is necessary to deep-dive and reliably pinpoint the cause of this effect.

6.2 Test with diverse user groups to account for vested interest

Social desirability is a type of response bias and describes the tendency of people to present themselves in a generally favorable manner [23]. Typically, one would assume that the social desirability bias would manifest equally in the worker and layman participant groups, thus canceling itself out. Workers rated CLAICA's usability (SUS) and user experience (UEQ) significantly higher than laymen. This result could be explained by the following factors: First, workers may (unconsciously) compare CLAICA with systems in their workplace that have poor usability and UX; second, workers are positively biased toward CLAICA as it was designed to help them and their colleagues (vested interest); and third, older workers might be more likely to be impressed by new technology. In fact, we found a significant correlation between age and SUS scores and there was a significant difference between the median age of the workers and laymen group. This may (partially) explain the difference in SUS scores between workers and laymen. Although there were no significant differences in perceived workload (NASA-TLX) scores or overall task completion rates, workers took significantly longer to complete the tasks. Although the workers group had experience with factory production lines, this clearly did not help them interact more efficiently with CLAICA. In fact, it appears to be detrimental. Maybe the workers took longer because they were on average older and had worse digital competencies. These findings have implications for studies that use laymen to test systems for a specific group of end-users. Namely, that context expertise did not help the user interact more efficiently with the assistant, help them complete more tasks, or reduce workload. However, context expertise positively affected subjective usability and UX.

6.3 Intent shortcuts override interaction modality types

The interaction modality (smartphone or laptop) appeared to not significantly affect any of the measures. This is unexpected, as the participants who used a laptop to interact with CLAICA could probably type faster [40], had a larger screen at their disposal, and in theory should have completed the knowledge exchange tasks faster. Perhaps the speed-typing advantage on laptops was counteracted by the availability of auto correct on the participants' smartphones and, perhaps more importantly, by CLAICA's use of buttons/shortcuts. Alternatively, it is also possible the time required for the typing did not constitute a significant part of the duration of the task. Considering that there appears to be no benefit to using a laptop over a smart, we suggest interacting with cognitive assistants like CLAICA on a smartphone, as many workers already carry one around in their pocket, and it can be used at whatever location is relevant to the required cognitive assistance or knowledge sharing.

6.4 Salvage conversational breakdowns and support single-word input

We grouped the comments of our participant into the following two categories: Conversational AI and UI. The effectiveness of CLAICA's NLU was the most prominent subtopic, for example, how well CLAICA could understand user intents and handle fallback. This is not surprising considering the challenges AI has in understanding natural language and the limited training examples (20-50 examples per intent) we used to train CLAICA. Regarding the divergent phrasing of user utterances, additional training data (e.g., from this user study) can be used to improve its accuracy. As it is important to test frequently and early with users, we cannot avoid NLU breakdowns; however, having a robust fall-back mechanism (e.g., offering the top three intent predictions) can soften the impact on UX.

Five participants, including factory workers who wanted to have short interactions, suggested providing a better NLU for keywords (that is, utter a single word to indicate intent). However, training an NLU model to classify intents purely based on a keyword may reduce its accuracy if the same keyword could be associated with several intents. Nevertheless, using keywords for frequently used features may work well, as we recognize that this is natural for humans and can save time.

6.5 Use suggested responses to streamline interaction

In contexts where users frequently use the same assistant (e.g., at a workplace or home), it is more reasonable to expect users to learn specific commands to invoke a feature. Indeed, several participants also mentioned that there was a learning curve in interacting with CLAICA as fully natural language did not always work. Nevertheless, many participants were positive about the interactive efficiency they could achieve once they learnt how to make themselves understood by CLAICA, for example, through the use of its shortcut buttons or specific phrasing. These buttons enabled faster responses and gave users hints as to what response was expected from them. Considering their success, we will continue to implement them and explore how to make the suggestions more intelligent and informative.

6.6 Ethical Considerations

The use of a system such as CLAICA raises several ethical concerns. First, we do not know what the long-term effects of using CLAICA may be. CLAICA can help share knowledge among workers; however, they can become overly dependent on its recommendations. Perhaps CLAICA could ask open questions to workers to invoke critical thinking. Second, the real-time data that CLAICA uses can be used by management to infringe on workers' privacy rights. Therefore, it is important that managers (or other workers) cannot use CLAICA to track people's performance or interactions without their knowledge. As such, guidelines and policies to protect and respect user privacy and rights, such as data management plans, are fundamental to Industry 5.0. Finally, it is not clear who owns the knowledge that workers share with CLAICA. One could claim that the knowledge acquired from the workers and formalized by CLAICA, becomes company property. In turn, by sharing their knowledge, workers effectively reduce their (perceived) value. Therefore, it is important that CLAICA or their employer employs reciprocal strategies. For example, the factory collaborating with the development of CLAICA has introduced a monetary reward system for workers who shared the most high-quality knowledge over a month period.

6.7 Limitations

This study focused on the task performance, usability, and user experience through several interactions with CLAICA; however, participants did not have to act on the exchanged information or share their own knowledge. Furthermore, the interactions were performed in the lab ("in vitro") as opposed to an actual production line. Therefore, the results reflect a part of CLAICA's UX. Assessing the UX of CLAICA "in vivo" will be the focus of future studies. To keep the duration of the user study to a reasonable duration (about an hour), we informed test participants that they could begin completing the survey after spending ten minutes trying to complete the assigned tasks. Although the vast majority of participants completed the tasks in less than ten minutes, several participants took longer, including multiple factory workers (see Fig. 11). Considering that this affected all groups and that the sample size for the factory was relatively small ($N = 12$), we decided not to remove these data points. We found a correlation between age and SUS. This may partially explain the difference in SUS for the context expertise dimension; however, we believe the difference in age between the groups to be indicative of the reality in factories. That is, many workers are >50 years old and have poorer digital competencies than many of the younger laymen we recruited for the study. Finally, we conducted the study in six separate user trials with two different moderators and en masse. First, conducting the study in a group setting may affect the results, as participants may feel pressure to finish faster than their peers or learn from each other. We tried to minimize this effect by asking participants not to converse with each other, anonymizing the results, and collecting the task performance measures automatically so that participants did not have to indicate when they were done. Second, differences between how the sessions we conducted, such as, the wording or demeanor of the moderators, could also affect the results. We believe that we minimized the effects sufficiently by following a script for the instructions and using the same slides, task instructions, and video material for the user trials.

7 FUTURE WORK

After we improve CLAICA on the basis of the findings of this work, we will conduct a laboratory study to evaluate CLAICA compared to the current training situation on the production line (using manuals and human-human training). Then, we will deploy CLAICA in factories and conduct in situ studies. Possible future research directions include investigating additional modalities (e.g., voice), long-term effects on workers' task performance, incurred cognitive workload, knowledge retention, workers' well-being, and explainable AI recommendations. We aim to continue to develop CLAICA in adherence to the (more than) human-centered principles of user experience, knowledge management, operator safety, and AI ethics.

8 CONCLUSION

We presented CLAICA, a continuously learning AI cognitive assistant that provides support to agile manufacturing workers by exchanging knowledge and providing quick access to instructional material. CLAICA is the product of a co-design process with factory workers and the focus of a user study with 83 participants. Our findings contribute to a deeper understanding of how prior training, context expertise, and interaction modality affect the user experience of cognitive assistants. Drawing on our findings, we elicit design and evaluation guidelines for cognitive assistants that support knowledge exchange in cognitively-demanding tasks and in challenging environments (e.g., an agile production line).

ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon 2020 research and innovation program via the project COALA "COgnitive Assisted agile manufacturing for a LAbor force supported by trustworthy Artificial Intelligence" (Grant agreement 957296).

REFERENCES

[1] 2016. *Intelligent Cognitive Assistants: Workshop Summary and Recommendations.* https://www.nsf.gov/crssprgm/nano/reports/2016-1003_ICA_Workshop_Final_Report_2016.pdf

[2] 2018. *Intelligent Cognitive Assistants: Workshop Summary and Research Needs.* https://www.nsf.gov/crssprgm/nano/reports/ICA2_Workshop_Report_2018.pdf

[3] Brunno Abner, Ricardo J. Rabelo, Saulo P. Zambiasi, and David Romero. 2020. Production Management as-a-Service: A Softbot Approach. In *Advances in Production Management Systems. Towards Smart and Digital Manufacturing*, Bojan Lalic, Vidosav Majstorovic, Ugljesa Marjanovic, Gregor von Cieminski, and David Romero (Eds.). IFIP Advances in Information and Communication Technology, Vol. 592. Springer International Publishing, Cham, 19–30. https://doi.org/10.1007/978-3-030-57997-5{_}3

[4] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300484

[5] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Studies* 4, 3 (may 2009), 114–123.

[6] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594. https://doi.org/10.1080/10447310802205776

[7] Farouk Belkadi, Mohamed Anis Dhuieb, José Vicente Aguado, Florent Laroche, Alain Bernard, and Francisco Chinesta. 2020. Intelligent assistant system as a context-aware decision-making support for the workers of the future. *Computers & Industrial Engineering* 139 (Jan. 2020), 105732. https://doi.org/10.1016/j.cie.2019.02.046

[8] Florian Brachten, Felix Brünker, Nicholas RJ Frick, Björn Ross, and Stefan Stieglitz. 2020. On the ability of virtual agents to decrease cognitive load: an experimental study. *Information Systems and e-Business Management* 18, 2 (2020), 187–207. https://doi.org/10.1007/s10257-020-00471-7

[9] Florian Brachten, Tobias Kissmer, and Stefan Stieglitz. 2021. The acceptance of chatbots in an enterprise context – A survey study. *International Journal of Information Management* 60 (2021), 102375. https://doi.org/10.1016/j.ijinfomgt.2021.102375

[10] Luka Bradeško, Michael Witbrock, Janez Starc, Zala Herga, Marko Grobelnik, and Dunja Mladenić. 2017. Curious Cat–Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition. *ACM Trans. Inf. Syst.* 35, 4, Article 33 (aug 2017), 46 pages. https://doi.org/10.1145/3086686

[11] Sebastian Büttner, Oliver Sand, and Carsten Röcker. 2017. Exploring Design Opportunities for Intelligent Worker Assistance: A New Approach Using Projetion-Based AR and a Novel Hand-Tracking Algorithm. In *Ambient Intelligence (Lecture Notes in Computer Science)*, Andreas Braun, Reiner Wichert, and Antonio Maña (Eds.). Springer International Publishing, Cham, 33–45. https://doi.org/10.1007/978-3-319-56997-0_3

[12] Mario Casillo, Francesco Colace, Loretta Fabbri, Marco Lombardi, Alessandra Romano, and Domenico Santaniello. 2020. Chatbot in Industry 4.0: An Approach for Training New Employees. In *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. 371–376. https://doi.org/10.1109/TALE48869.2020.9368339

[13] Fabio Clarizia, Francesco Colace, Massimo De Santo, Marco Lombardi, Francesco Pascale, and Domenico Santaniello. 2019. A Context-Aware Chatbot for Tourist Destinations. In *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. 348–354. https://doi.org/10.1109/SITIS.2019.00063

[14] Jet Cullen and Alan Bryman. 1988. The knowledge acquisition bottleneck: time for reassessment? *Expert Systems* 5, 3 (1988), 216–225. https://doi.org/10.1111/j.1468-0394.1988.tb00065.x

[15] Bulent Dal, Phil Tugwell, and Richard Greatbanks. 2000. Overall equipment effectiveness as a measure of operational improvement–a practical analysis. *International Journal of Operations & Production Management* 20, 12 (2000), 1488–1502. https://doi.org/10.1108/01443570010355750

[16] Brett Edwards, Michael Zatorsky, and Richi Nayak. 2008. Clustering and classification of maintenance logs using text data mining. *Volume 87-Data Mining and Analytics 2008* (2008), 193–199.

[17] Enzo Fenoglio, Emre Kazim, Hugo Latapie, and Adriano Koshiyama. 2022. Tacit knowledge elicitation process for industry 4.0. *Discover Artificial Intelligence* 2, 1 (March 2022), 6. https://doi.org/10.1007/s44163-022-00020-w

[18] Mina Foosherian, Samuel Kernan Freire, Evangelos Niforatos, Karl A. Hribernik, and Klaus-Dieter Thoben. 2022. Break, Repair, Learn, Break Less: Investigating User Preferences for Assignment of Divergent Phrasing Learning Burden in Human-Agent Interaction to Minimize Conversational Breakdowns *(MUM '22)*. Association for Computing Machinery, New York, NY, USA, 151–158. https://doi.org/10.1145/3568444.3568454

[19] Markus Funk, Tilman Dingler, Jennifer Cooper, and Albrecht Schmidt. 2015. Stop Helping Me - I'm Bored! Why Assembly Assistance Needs to Be Adaptive. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (Osaka, Japan) *(UbiComp/ISWC'15 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1269–1273.

https://doi.org/10.1145/2800835.2807942

[20] Markus Funk, Thomas Kosch, Scott W. Greenwald, and Albrecht Schmidt. 2015. A benchmark for interactive augmented reality instructions for assembly tasks. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia (MUM '15)*. Association for Computing Machinery, New York, NY, USA, 253–257. https://doi.org/10.1145/2836041.2836067

[21] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[22] Lorenz Hoerner, Markus Schamberger, and Freimut Bodendorf. 2022. Using Tacit Expert Knowledge to Support Shop-floor Operators Through a Knowledge-based Assistance System. *Computer Supported Cooperative Work (CSCW)* (Sept. 2022). https://doi.org/10.1007/s10606-022-09445-4

[23] Ronald R Holden and Jennifer Passey. 2010. Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences* 49, 5 (2010), 446–450. https://doi.org/10.1016/j.paid.2010.04.015

[24] Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael Mctear. 2019. Usability Testing of a Healthcare Chatbot: Can We Use Conventional Methods to Assess Conversational User Interfaces?. In *Proceedings of the 31st European Conference on Cognitive Ergonomics* (BELFAST, United Kingdom) *(ECCE '19)*. Association for Computing Machinery, New York, NY, USA, 207–214. https://doi.org/10.1145/3335082.3335094

[25] Klementina Josifovska, Enes Yigitbas, and Gregor Engels. 2019. A Digital Twin-Based Multi-modal UI Adaptation Framework for Assistance Systems in Industry 4.0. In *Human-Computer Interaction. Design Practice in Contemporary Societies (Lecture Notes in Computer Science)*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 398–409. https://doi.org/10.1007/978-3-030-22636-7_30

[26] Samuel Kernan Freire, Evangelos Niforatos, Zoltan Rusak, Doris Aschenbrenner, and Alessandro Bozzon. 2022. A Conversational User Interface for Instructional Maintenance Reports. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 5, 6 pages. https://doi.org/10.1145/3543829.3544516

[27] Samuel Kernan Freire, Sarath Surendranadha Panicker, Santiago Ruiz-Arenas, Zoltán Rusák, and Evangelos Niforatos. 2022. A Cognitive Assistant for Operators: AI-Powered Knowledge Sharing on Complex Systems. *IEEE Pervasive Computing* (2022), 1–9. https://doi.org/10.1109/MPRV.2022.3218600

[28] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. 2019. A conversational agent in support of productivity and wellbeing at work. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*. IEEE, 1–7. https://doi.org/10.1109/ACII.2019.8925488

[29] Thomas Kosch, Markus Funk, Albrecht Schmidt, and Lewis L. Chuang. 2018. Identifying Cognitive Assistance with Mobile Electroencephalography: A Case Study with In-Situ Projections for Manual Assembly. *Proceedings of the ACM on Human-Computer Interaction* 2, EICS (June 2018), 11:1–11:20. https://doi.org/10.1145/3229093

[30] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. Chapter 11 - Analyzing qualitative data. In *Research Methods in Human Computer Interaction (Second Edition)* (second edition ed.), Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser (Eds.). Morgan Kaufmann, Boston, 299–327. https://doi.org/10.1016/B978-0-12-805390-4.00011-X

[31] Nguyen-Thinh Le and Laura Wartschinski. 2018. A cognitive assistant for improving human reasoning skills. *International Journal of Human-Computer Studies* 117 (2018), 45–54. https://doi.org/10.1016/j.ijhcs.2018.02.005

[32] Franz Georg Listl, Jan Fischer, and Michael Weyrich. 2021. Towards a Simulation-based Conversational Assistant for the Operation and Engineering of Production Plants. In *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 1–4. https://doi.org/10.1109/ETFA45728.2021.9613681

[33] Francesco Longo, Letizia Nicoletti, and Antonio Padovano. 2017. Smart operators in industry 4.0: A human-centered approach to enhance operators' capabilities and competencies within the new smart factory context. *Computers & Industrial Engineering* 113 (2017), 144–159. https://doi.org/10.1016/j.cie.2017.09.016

[34] Bei Luo, Raymond YK Lau, Chunping Li, and Yain-Whar Si. 2022. A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 1 (2022), e1434. https://doi.org/10.1002/widm.1434

[35] J Müller. 2020. Enabling Technologies for Industry 5.0—Results of a Workshop with Europe's Technology Leaders. *Directorate-General for Research and Innovation* (2020). https://doi.org/10.2777/082634

[36] Saeid Nahavandi. 2019. Industry 5.0—A human-centric solution. *Sustainability* 11, 16 (2019), 4371. https://doi.org/10.3390/su11164371

[37] Ricardo J. Rabelo, Saulo Popov Zambiasi, and David Romero. 2019. Collaborative Softbots: Enhancing Operational Excellence in Systems of Cyber-Physical Systems. In *Collaborative Networks and Digital Transformation*, Luis M. Camarinha-Matos, Hamideh Afsarmanesh, and Dario Antonelli (Eds.). IFIP Advances in Information and Communication Technology, Vol. 568. SPRINGER NATURE, 55–68. https://doi.org/10.1007/978-3-030-28464-0{_}6

[38] Leonardo Rodriguez, Fabian Quint, Dominic Gorecky, David Romero, and Héctor R. Siller. 2015. Developing a Mixed Reality Assistance System Based on Projection Mapping Technology for Manual Operations at Assembly Workstations. *Procedia Computer Science* 75 (Jan. 2015), 327–333. https://doi.org/10.1016/j.procs.2015.12.254

[39] David Romero, Johan Stahre, Thorsten Wuest, Ovidiu Noran, Peter Bernus, Åsa Fast-Berglund, and Dominic Gorecky. 2016. Towards an operator 4.0 typology: a human-centric perspective on the fourth industrial revolution technologies. In *Proceedings of the 46th International Conference on Computers and Industrial Engineering*, Mohamed Dessouky, Yasser Dessouky, and Hamed K. Eldin (Eds.). 608–618.

[40] Quentin Roy, Sébastien Berlioux, Géry Casiez, and Daniel Vogel. 2021. Typing Efficiency and Suggestion Accuracy Influence the Benefits and Adoption of Word Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 714, 13 pages. https://doi.org/10.1145/3411764.3445725

[41] Jeff Sauro. 2011. *A practical guide to the system usability scale: Background, benchmarks & best practices.* Measuring Usability LLC.

[42] Kaisa Savolainen. 2021. User-Centred Design without Involving Users: A Longitudinal Case Study in a Human-Centred-Design–Mature Company. *The Design Journal* 24, 6 (2021), 887–905. https://doi.org/10.1080/14606925.2021.1980267

[43] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Construction of a Benchmark for the User Experience Questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (06 2017), 40–44. https://doi.org/10.9781/ijimai.2017.445

[44] Thurston Sexton, Michael P. Brundage, Michael Hoffman, and K C Morris. 2017. Hybrid datafication of maintenance logs from AI-assisted human tags. In *2017 IEEE International Conference on Big Data (Big Data)*. 1769–1777. https://doi.org/10.1109/BigData.2017.8258120

[45] Amit Sheth, Hong Yung Yip, Arun Iyengar, and Paul Tepper. 2019. Cognitive services and intelligent chatbots: current perspectives and special issue introduction. *IEEE Internet Computing* 23, 2 (2019), 6–12. https://doi.org/10.1109/MIC.2018.2889231

[46] Yehya Soliman and Hannu Vanharanta. 2020. A Model for Capturing Tacit Knowledge in Enterprises. In *Advances in Human Factors, Business Management and Leadership (Advances in Intelligent Systems and Computing)*, Jussi Ilari Kantola and Salman Nazir (Eds.). Springer International Publishing, Cham, 141–148. https://doi.org/10.1007/978-3-030-20154-8_14

[47] Wenjin Tao, Ze-Hao Lai, Ming C. Leu, Zhaozheng Yin, and Ruwen Qin. 2019. A self-aware and active-guiding training & assistant system for worker-centered intelligent manufacturing. *Manufacturing Letters* 21 (Aug. 2019), 45–49. https://doi.org/10.1016/j.mfglet.2019.08.003

[48] Klaus-Dieter Thoben, Stefan Wiesner, and Thorsten Wuest. 2017. "Industrie 4.0" and Smart Manufacturing – A Review of Research Issues and Application Examples. *International Journal of Automation Technology* 11, 1 (2017), 4–16. https://doi.org/10.20965/ijat.2017.p0004

[49] Stefan Wellsandt, Mina Foosherian, Katerina Lepenioti, Mattheos Fikardos, Gregoris Mentzas, and Klaus-Dieter Thoben. 2022. Supporting Data Analytics in Manufacturing with a Digital Assistant. In *Advances in Production Management Systems. Smart Manufacturing and Logistics Systems: Turning Ideas into Action*, Duck Young Kim, Gregor von Cieminski, and David Romero (Eds.). IFIP Advances in Information and Communication Technology, Vol. 664. Springer Nature Switzerland, Cham, 511–518. https://doi.org/10.1007/978-3-031-16411-8{_}59

[50] Stefan Wellsandt, Karl Hribernik, and Klaus-Dieter Thoben. 2021. Anatomy of a Digital Assistant. In *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems*, Alexandre Dolgui, Alain Bernard, David Lemoine, Gregor von Cieminski, and David Romero (Eds.). IFIP Advances in Information and Communication Technology, Vol. 633. Springer International Publishing, Cham, 321–330. https://doi.org/10.1007/978-3-030-85910-7{_}34

[51] Xun Xu, Yuqian Lu, Birgit Vogel-Heuser, and Lihui Wang. 2021. Industry 4.0 and Industry 5.0—Inception, conception and perception. *Journal of Manufacturing Systems* 61 (2021), 530–535. https://doi.org/10.1016/j.jmsy.2021.10.006

[52] Shuo Zhou and Timothy Bickmore. 2022. A Virtual Counselor for Breast Cancer Genetic Counseling: Adaptive Pedagogy Leads to Greater Knowledge Gain *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 436, 17 pages. https://doi.org/10.1145/3491102.3517553