universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Design of Human-Chatbot Interaction

in Stressful Scenarios

verfasst von / submitted by

Veronika Vishnevskaia

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 013
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Joint Degree Programme MEi:CogSci Cognitive Science
Betreut von / Supervisor:	Dr. Marcin Skowron
Mitbetreut von / Co-Supervisor:	

Abstract

This thesis explores the topic of Human-Chatbot interaction from the interdisciplinary perspective of cognitive science, empirical psychology, computer science, and neuroscience. In the scope of the thesis, an interaction framework was developed based on the reviewed HAI guidelines. The framework was tested empirically in a full-factorial, between-subject experiment with 33 participants. During the experiment, physiological measures of electrodermal activity were used to obtain objective signals from sympathetic nervous systems. Additionally, a subjective questionnaire was conducted to the measure self-reported of stress and perceived usefulness, easiness, learnability, and satisfaction from the interaction. The results show statistical significance of the effect that Interaction design has on the user experience. The main contributions of the thesis include i) grounding of the Human-Chatbot interaction approaches in a comprehensive theoretical basis; ii) development of an original experimental design; iii) empirical testing of the HAI guidelines by collecting objective physiological data; iv) development of original data analysis pipeline for electrodermal data applied in a full-factorial design; v) empirical contributions to the discussion about subjective and objective methods of usability research; vi) suggestions for future research in the area of empirical studying the Human-Chatbot interaction.

Keywords: Human-AI Interaction, HAI Design guidelines, Human-Chatbot Interaction, Human-centered design, conversational AI, physiological measurements of user experience

Acknowledgements

If I set myself the goal of mentioning all the people who, in one way or another, had a hand in my successful completion of the master's program, this list would take half the work. Therefore, I will have to limit myself to only a tiny part, remaining infinitely grateful to all involved.

First of all, I would like to thank my supervisor, Dr. Marcin Skowron, for his dedication, thoughtful comments, support, and delicate guidance. He was always there for me, showing some new perspective and providing valuable advice each time we discussed the thesis.

Secondly, my deepest gratitude goes to professors, coordinators, the program committee, and everyone working on the MEi-CogSci program. Thank you for giving me the chance to make this fantastic journey that changed my life and professional development. You are doing an outstanding job building this exceptional community. I am delighted and proud to be a part of it.

I want to thank colleagues from the Competence Team for implanted devices, Center for Medical Physics and Biomedical Engineering of the Medical University of Vienna, namely Dr. Manfred Bijak and his students Clara Schmidtmann, Ahmed Ahmed, and Pia Hisberger. They showed me unlimited support in learning the topic of electrodermal activity measurement and shared their experience and best practices. Thank you, colleagues!

Next, I want to thank my team from AI Lab (Raiffeisen Bank International). My colleagues managed to upskill me in programming, cloud infrastructure, and chatbots development quickly and easily during our daily work. Without these skills, I would not bring this research to life.

Last but not least and most important, I want to thank my family, especially my parents, son, and husband. I was supported and accepted during my whole life, finding energy and determination for new adventures inside my family circle. Everything I am and everything I achieve comes from my family. Thank you for all the unconditional love you always surround me with.

Table of Contents

List of figures

List of Tables

List of Boxes

1. Theoretical Grounds

1.1. Introduction

"The future of customer conversation: More than words, more than AI" report by Accenture (Accenture, 2021) highlights that chatbots are falling behind customer expectations. Less than half of customers (48%) that used online-chat with the conversational AI assistant confirmed that they would like to use it again. At the same time, many industries such as customer support, financial services, tourism, and others are planning to expand the utilization of chatbots even more. The same report states: by 2022, 70% of the customers will interact with the chatbots, compared to 15% in 2018. If chatbots are the technology to become a common one in upcoming years, how can we improve it?

The report assumes that the root cause of the customers' disappointment is the current approach to the conversation design, which is *human-like* and *technology-centered* instead of *human-centered*. In the comprehensive overview based on more than 890 *Human-AI Interaction Design* papers, Xu et al. (2021) state that the current pain-points of conversational AI are "teething problems" similar to those of general computer systems and interfaces they had in the 1980-1990s. Back then, in the 1980s, it caused the emergency of the *Human-Computer Interaction field*; the necessity to solve problems of AI technology led to the fast growth of the *Human-AI Interaction* domain.

In 2016, Luger & Sellen published a paper where users compare the experience of dealing with the chatbot with a "really bad panic attack". The main downside reported by AI assistants' users was that they didn't understand what the chatbot could do and had to figure it out by themselves. Users complained that they "felt let down and lost", as they didn't receive proper feedback from the assistant. Meanwhile, the clear explanation of the possibilities that the system provides and timely given feedback is one of the main principles of human-centered design (see subsection "Human-centered design for more details). These findings confirm that the technological maturity of natural language processing (NLP), which chatbot developers often focus on, does not cover all the users' needs. A more systematic approach to the interaction design is required.

"Big players" with strong R&D departments, such as Microsoft and Google, have published guidelines for Human-AI interaction already in 2019 (see Amerschi et al., 2019; Google PAIR, 2019[1]). These guidelines summarize the industry experience and academic research and provide concrete recommendations on how to develop AI that will be taking into account human thinking and behavior. Yet still, they are not fully applied in the practice, as can be judged based on the analytical reports.

[1] Google PAIR. *People + AI Guidebook*. Published May 8, 2019. pair.withgoogle.com/guidebook

Human-AI interaction design proved its effectiveness in chatbot-related empirical studies, published during the latest years. Adam, Wessel, and Benlian (2021) measured the user compliance and satisfaction for banking FAQ chatbot; Tsai, Liu, and Chuan (2021) evaluated the interaction satisfaction, user engagement, and brand likeability depending on the chatbots social presence (social validity as "real" companion). Weber and Ludwig (2020) interviewed users to clarify what affects the perception of voice assistants; Skjuve et al. (2019) examined how the chatbot transparency influenced the user evaluation; De Cicco, Silva, and Alparone (2020) tested how the interaction style of the chatbot will improve the attitude toward it in millennials.

The above empirical studies are only examples of the rapidly growing Human-AI Interaction (HAI) field. Still, most of the existing works in the area of HAI are based on interviews, questionnaires, and conversation logs (Chaves & Gerosa, 2020) – primarily self-reported subjective data. Significantly fewer studies use physiological measurements as a source of more objective quantitative data (for example, see Ciechanowski et al., 2019). The present work aims to contribute to this domain, using electrodermal activity data (EDA) to measure the participants' physiological response to the interaction with chatbot models.

Another critical point is that often the users interact with the chatbots in a state of stress. Many scenarios where chatbots are widely applied – customer support, banking services, technical support, traveling, dealing with authorities – are associated with the negative arousal. Such scenarios place higher demands on the system design, as a more "stressed" focused brain is especially sensitive to the user experience (Norman, 2004). This work strives to prove that the Human-AI Interaction Design approach can mitigate the negative effects of stressful scenarios for the participants.

The research question of the current work is "**To what extent an improved Interaction Design can reduce the levels of the participants' sympathetic arousal that they experience during the interaction with the chatbot when a mild stress factor is introduced?**"

The main objectives of this master thesis are: a) based on existing studies and guidelines, to develop conversational AI models with the default and improved interaction design as V1; b) with open-source packages and cloud infrastructure, to implement web-application versions with and without the added stress factor as V2; c) to recruit participants and to test the experimental hypothesis in four groups, based on 2x2 full-factorial between-subject design; d) to measure the participants' levels of electrodermal activity to obtain objective physiological data as the dependent variable; e) to conduct the data pre-processing, statistical analysis and hypothesis testing; f) to summarize and comment on the results.

The master thesis has the following structure: Section 1, “Theoretical Grounds,” provides a comprehensive introduction into related concepts of cognitive science, Psychology, and Computer Science. Theoretical frameworks in Human-Computer Interaction (HCI), Human-AI Interaction, and Human-AI Interaction Design serve as a foundation for the “Conversational Model Development”, described in Section 2. The experimental design, procedure, and data acquisition are covered in Section 3, “Method”. Section 4, “Results,” describes the data pre-processing, analysis, and statistical significance testing. The obtained results are commented on in Section 5, “Discussion”.

1.2. Theoretical Concepts Related to Present Work

The present work is developed inside the interdisciplinary frame of cognitive science, also employing concepts from psychology and computer science. In this section, the main theories used for grounding the study are shortly summarized: postcognitivistic approaches to the consciousness, conceptual models, human-centered design, technology acceptance model, and affective interaction design. The main goal of the section is not to provide a detailed description of the theories (which would be just an excessive quotation) but rather to focus on the central ideas used.

All the concepts described are interconnected and were deduced from one another through the 1970-the 2010s. Together they create a comprehensive understanding of human cognition, emotions, and behavior in their interaction with the environment and artifacts, such as computer systems and AI agents.

1.2.1. Postcognitivism in Cognitive Science: Activity Theory & 4E-approaches

Postcognitivistic theories question the “classical” *Cognitivism* understanding of cognition, formed by Noam Chomsky, George A. Miller, and other leaders of the Cognitive revolution. For the 1950s, this view was revolutionary; cognition was seen as a biological analog of the computer, independent from the body and the environment in forming its inner representation. By the 1970s, this perspective was revised in various studies aiming to understand the relationship between cognition, body, and the external world. The *Autopoiesis* theory developed by Humberto Maturana and Francisco Varela (Maturana & Varela, 1972) proposed a new understanding of the living systems. It shows that these systems can interact with the environment to support their existence even without any cognition, for example, as each separate living cell or protozoa does. This innovative vision gave a root for the whole list of new theories, considering interaction as the main starting point for cognition development.

For the Human-Computer Interaction domain, the primary input made the *Activity theory* and so-called "*4E-Cognition approaches*": *Embodied, Embedded, Extended, and Enacted cognition.*

Activity theory emerged in the works of the Soviet school of developmental psychology throughout the 1930s-1950s, particularly in the papers of Alexei Leont'ev & Lev Vygotsky. Later found in the western cognitive science and psychology field, it heavily affected the *Information systems* theory and *Human-computer Interaction* area. Activity theory considers the development of each individual as impossible in isolation and taking place due to the interaction with the social environment and community, following the existing rules, and using available tools for problem-solving.

Bonnie Nardi and Kari Kuutti (Bannon & Kuutti, 1993; Kuutti, 1991, 1996; Nardi, 1996) applied the activity theory to Information systems and Human-Computer Interaction. Activity theory is used to understand the work activities implemented through various information systems and interfaces as tools. Cognition is considered as distributed, embodied, and realized in the external devices phenomenon, emerging from daily human activity.

The concepts used in the activity theory – activity, mediating artifacts (tools), community – became the core ideas of the *4E-approaches* that emerged two decades later, this time directly under the cognitive science domain. In studies by J. Gibson (1979), E.Hutchins (1995), M.Wilson (2002), E. Di Paolo (2014), and other authors, first notions of *embodied, embedded, enactive* and *extended* cognition were given, later united by R. Menary (2010) and M. Rowlands (2010) in *4E framework*.

Embodied cognition theory states that cognition is body-based and inherent to body reactions, chemicals, and signals. *Embedded* or *situated cognition* highlights that it heavily depends on the perception of the external world and interaction with it in specific situations and scenarios. *Enacted cognition* approach inherits the ideas of the Active theory directly. It points out that the environment is a part of our cognitive system, and studying the mind apart from its environment and dynamics makes no sense. Finally, extended cognition declares that human information processing abilities are limited; therefore, we extend our cognition into the environment and create means (artifacts) that help us to reduce our cognitive workload.

Overall, 4E-approaches introduce a perfect scene for explaining how the development of computer systems, which humanity becomes more dependent on, affects our cognition and behavior. These theories were central for the further growth of the human-computer interaction and human-centered design domains.

"Good conceptual models are the key to understandable, enjoyable products: good communication is the key to the good conceptual model."

1.2.2. Conceptual/Mental Model

In 1983 Johnson-Laird introduced the notion of the mental model as "*structural analogs of the world*." The author considered the mental models as a mapping between the real world and internal mental representations.

Shortly after publication, this theory was applied to the Human-Computer Interaction domain by Don Norman. In his "*The Design of Everyday Things"* (Norman, 1989; 2013 - the revised version), the author uses the term "mental" or "*conceptual model"* to describe a cognitive construct that the user creates during interaction with the system. For users, it is a simplified, high-level explanation of how things work. Norman states that the perceived device structure, including its main interactive components – signifiers, affordances, constraints, and mappings (see box 1), – can form the conceptual models natively.

When the system does not provide a visible structure, the user can rely only on its interaction experience and information provided (manuals, online information, marketing description). These inputs create a *system image*, which serves as a source of users' mental model. The system image, in its turn, is a derivative of the designer's mental model of the product.

To help users develop a good conceptual model, it is crucial to overcome the creator's bias. The system will not be self-explanatory because it seems so to the designer; proper communication is the key, especially when something does not work.

Don Norman's Main Principles of Interaction
Affordances - based on James J. Gibson theory (Gibson, 1966) - the possible actions offered by the item
Signifiers - indicators that communicate to the user how and where the actions can take place
Constraints - communication of existing physical, cultural, semantic, and logical limitations
Mappings - relation between the controls used and the systems controlled
Feedback - communication of the system work on the provided request and the results of it

Box 1. Main Principles of Interaction by D.Norman

How much explanation is enough for a proper mental model, and which kind of information is essential? Kulesza et al. (2013) answer this question in a study on a music-

recommending intelligence agent. Participants received musical recommendations and explanations about how and why the system made its choice during the experiment. Researchers used *soundness* (validity and credibility) and *completeness* (if all required information is provided) as two variables describing the explanation quality. Based on the values of these variables, they evaluated the resulting mental models of the users. Experimental results showed that the completeness of the explanation is more critical for comprehensive mental models than soundness. At the same time, oversimplified explanations lead to less trust and attention paid to them.

For the present work, the conceptual model is an outcome of the interaction design framework used. Therefore, the chatbot with the improved interaction framework should provide enough explanations and feedback about the system so that users will have a clear conceptual model of it. The default interaction design, on the contrary, communicates insufficiently and creates a poor conceptual model.

1.2.3. Human-Centered Design and Technology Acceptance Model

Don Norman is considered a "godfather" of User experience (UX) and Human-centered design approaches. In "*User-centered system design: New perspectives on human-computer interaction"* (1986), he calls for the creation of a "science of user-centered design," an approach where the needs of the user will define the implementation and interaction will drive the technology. Computer systems should be built based on human psychology and cognitive processes, stages of actions, and ways of interaction.

Communication is again highlighted as the main requirement for good design: the machine should always indicate what actions are possible, what is happening right now, and what will happen. Communication becomes especially crucial when things go wrong; "designing for an error" – one of the substantial parts of the overall design process. Error, as Norman emphasizes, is the point that may bring the most satisfaction from the interaction. When the machine indicates what is going wrong and the problem, the user understands the issue clearly and can act to solve it.

Norman's approach can be considered as primarily focused on user experience and interaction. However, a quarter of a century later, "human-centered design" is no longer monolith. Auernhammer (2020) highlights eight various design approaches, all of which have their implications to the AI: *human-centered system* studies the impact on social systems of such organizations; *social design* considers an interplay of the technologies with socio-economic structures and ideologies; the *participatory design* includes different social groups and their ethical perspectives; *inclusive design* tries to prevent the discrimination of specific groups of people in the design means; *interaction design* focuses on the usability of the systems and their influence on people behavior and experience;

persuasive technology approach reveals the hidden patterns of technologies that nudge users toward some intended choices; *human-centered computing* studies how to remove constraints in people capabilities by the technologies; and *need-design response* investigates how to develop the systems based on the human needs, but not to exploit them.

During the past 20 years, **interaction design** (IxD) has been studied in detail in the general Human-computer interaction field. It got a lot of attention in the IT industry as one of the product characteristics that increase its profitability. Usability and interaction experience also have a central place in the **technology acceptance theory** (Davis, 1989; Lee, Kozar & Larsen, 2003). The theory describes the stages in which an individual accepts the new technology. In the "classical" version of the idea, the technology is evaluated by two variables: Perceived Usefulness (PU) and Perceived Easiness of Use (PEOU), predicting target variables Behavioral Intention (BI) and Behavior (B). Segars and Grover (1993) introduced the third independent variable, not correlating with two others – Effectiveness [in solving the problem], but the descending studies did not support this proposal.

The present study focuses mainly on interaction design as the factor of chatbot technology acceptance. In this case, following the interdisciplinary approach of Cognitive Science, the perceived usability evaluation is completed by objective data from the peripheral nervous system.

1.2.4. Affective Interaction Design

Affect is a subjective emotional experience; contrary to emotions, which are often considered a reflected and conscious process, affect is regarded as a "feeling that you might experience without knowing why" (Norman, 2004). Affect theory was first developed by Silvan Tomkins (Tomkins, 1962), who distinguished positive, negative, and neutral affects (see box 2). Affects are often evaluated on a two-dimensional scale based on Russel's "Circumplex model of affect" (Russel, 1980). Two-dimensional space contains *arousal* measurement on one of the axes and *valence* on another. Four resulting quadrants correspond to happy (high arousal, positive valence), sad (low arousal, negative valence), angry (high arousal, negative valence), and calm (low arousal, positive valence) emotional states (see fig.1)

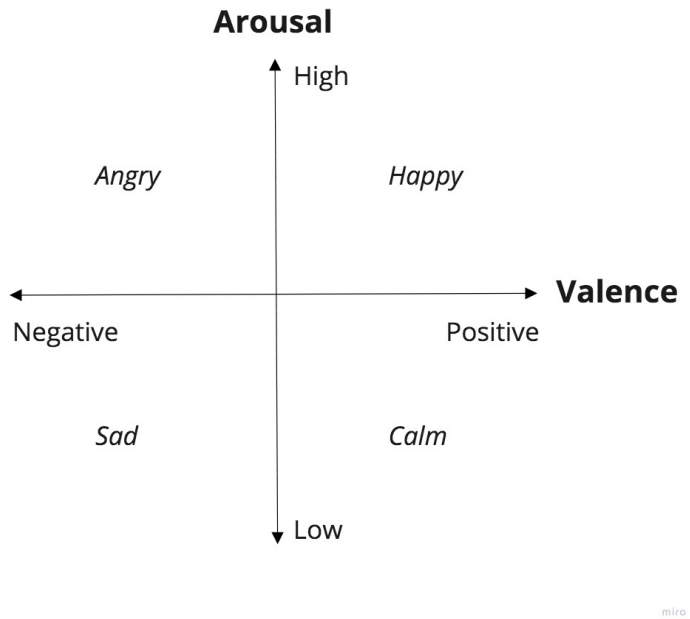

Fig.1 Russel's Two-dimensional Emotional Space

Affective design is a design approach in which the user's emotional experience is considered and planned during the product development. As a result, the product can create positive and eliminate negative emotions of the user during the interaction.

Don Norman described the emotional design in his book "*Emotional Design: Why we love (or hate) everyday things"* (Norman, 2004). Based on the architecture of the human nervous system, he introduced three levels of the design: *visceral design* - the lowest, intuitive level, primarily associated with the product appearance; *behavioral design* - middle level, responsible for the pleasure and effectiveness of use; *reflective design* - the highest level related to the self-image of the user, his satisfaction and memories. When the designer accounts on these levels during product development, he can create a product that the users will highly evaluate and demand.

Another important aspect is how our nervous system processes the information while in different emotional states. Norman argues that during the *negative affect* neurotransmitters, focus the brain processing on solving the emerging problem (and staying alive). On the other hand, when in a positive affect, the brain processing is broadened - we are relaxed, curious, and ready to learn how to use new opportunities.

If the product is developed for stressful situations, it is essential to count on the emotional state. For example, suppose we know that the user will most probably experience a negative affect. In that case, we should adjust to the focused brain, provide

only the required information and timely feedback, and make the system more transparent and unambiguous.

Silvan Tomkins's Nine Affects

Positive

Enjoyment/Joy - reaction to success/impulse to share

Interest/Excitement - reaction to new situation/impulse to attend

Neutral

Surprise - reaction to sudden change/reset impulses.

Negative

Anger - reaction to threat/impulse to attack

Disgust - reaction to bad taste/impulse to discard

Dissmell - reaction to bad smell/impulse to avoid

Distress - reaction to loss/impulse to mourn

Fear - reaction to danger/impulse to run or hide

Shame - reaction to failure/impulse to review behavior

Box 2. Nine Affects by S.Tomkin

The conceptual framework used in the present work thus can be summarized as follows. Person's interaction with the environment and artifacts, based on the physiological properties of the embodied brain, is seen in Postcognitivism as the main driving force of cognitive development. The rapid growth of Cognitive Science led to the emergence of Human-centered design and Human-Computer Interaction. These domains help adjust the artificial systems to the human cognitive models, improving the perceived usability and ease of use. The two factors are crucial for the users' acceptance of new technologies. Affects can also influence this process, so the interaction and experience should be designed to take into account the possible emotional states of the users.

The implementation of these concepts in Human-Computer Interaction and Human-Interaction domains is examined in more detail in the next section.

1.3. Human-AI Interaction

This section aims to provide an understanding of the main topics in the Human-AI Interaction domain, its interconnection with the general Human-Computer Interaction field, and examine existing guidelines in the HAI area.

1.3.1. Human-Computer Interaction as a Root

As was already mentioned in the Introduction, many problems that Human-AI interaction is facing nowadays were valid for HCI back in the 1980-90s. In 1973, James Martin wrote in "*Design of Man-Computer Dialogues"* that the computer industry "will be forced" to focus on people and usage rather than on hardware. Though until the middle of 1980s the computers were primarily used in business tasks, the comfort of the operators was still a concern for managers, as the employees could avoid using the new systems. The technology acceptance model introduced in 1989 by Fred Davis aimed to help companies increase the adoption rate.

Starting from 1980, when IBM and later Macintosh developed a home computer for individual users, cognitive psychologists became the core of the fields' research groups (Grudin, 2005). At the first HCI conference in 1983, IBM researchers led by John Gould presented a paper focused on user-centered, iterative design based on prototyping (Gould and Lewis, 1983). This less scientifically strict and faster approach to empirical studies became the framework of modern UX research.

Macintosh's success with home computers and graphical user interface (GUI) started a new era in HCI. The spreading of the Internet in the late 1990s emphasized a focus on interface development and visual design even more. As Don Norman pointed out in his "Emotional design", the product should have been not only helpful and learnable but enjoyable, as users were sensitive to the visceral design level as well.

Cognitive science supported the HCI domain in these times, providing theoretical grounds, methods, and tools - from the foundations of human psychology and physiology to the 4E approaches, comprehensively describing the interaction of users with modern computer systems. At the same time, though developed alongside the cognitive revolution, the field of AI had a strong focus on mathematics and engineering.

In his essay "Shifting viewpoints: Artificial intelligence and human-computer interaction" (Winograd, 2006), Terry Winograd distinguishes two main approaches to the development of computer systems: "rationalistic" and "design" approach. Rationalistic approach is based more on the understanding of people as "cognitive machines" that can be easily modeled and predicted, and the design approach is focused on the interaction between the person and the environment. HCI, embracing the design approach, considered human behavior too complex for modeling. New interfaces and applications were developed through iterative prototype testing and improving, where embodied human thinking was recognized and accounted for in the interaction.

As Grudin highlights in his comparison review (Grudin, 2009), the rapid growth of HCI was caused by the margin interest of the companies selling mass-market computers. AI for decades stayed a long-term investment at a governmental level. The cloud

infrastructure development in the 2000-the 2010s reduced the cost of data storage, processing, and access to the networks. Further, the spread of smartphones and "big data" made it possible to apply AI to more widely available systems, platforms and in different application scenarios. Successful AI-powered applications in turn required better interfaces and increased the demand for an improved usability and more optimal interaction design.

1.3.2. Human-AI interaction: Main Challenges

John Launchbury (2017) introduced DARPA perspective on three waves of AI, divided by the "AI winters":

1. *Handcrafted Knowledge* – inspired by symbolism and connectionism, the first wave focused on expert systems and technological exploration. Engineers defined the rules and structured the knowledge; independent learning for the machine was impossible, and handling uncertainty was poor.

2. Statistical learning – the second wave focused on technological enhancement, where statistical models in NLP, pattern recognition, and artificial neural networks developed. Machines obtained classification and prediction capabilities; the reasoning was minimal. The challenges of the second wave were individual unreliability, strong dependence on the training data, and vulnerability.

3. *Contextual adaptation* – currently arising the third wave emphasizes ethical design, usability, interaction, and human in AI systems. The Third wave brings contextual explanatory models, where machines obtain abstracting and reasoning capabilities.

Xu et al. (2021) highlight that the industry fully considers users' needs only in the latest wave. It is explained by the fact that mass-market applications powered by AI, in which systems adaptation and popularity largely depend on their usability, have finally brought enough attention to the topic of human-centered design.

As usual, there was a time lag between industry and academia in adopting the new topics and domains. First attempts to focus more on the *human* in the human-AI interaction were made in the 1990s when the HAI domain was not yet separated from the general HCI. Billings (1991) and Frischer (1995) explored collaboration between human operators and automated systems, defining the right balance between the automation and augmentation of human intelligence. Such principles as the necessity to involve and inform the operator, control the automated system, and make it more predictable are highlighted in these works.

Dan Norman (1994) discussed the exploration of intelligent agents by software manufacturers and foreseen main challenges that would emerge, being rather social than

technical. The author pointed out that for a smooth acceptance of the new technology, such pain points as feeling in control, managing expectations, safety and privacy of the data, and transparency about the underlying operations will have to be covered.

In his formatting and influential paper about principles of mixed-initiative user interfaces, Horvitz (1999) defined the main approaches on how to improve the interaction between people and automated systems: consider uncertainty about user's goals; consider the status of user's attention; allow efficient direct invocation; provide mechanisms to refine results; maintain the memory of recent interactions; continue to learn by observing.

When principles discussed in the papers above are compared with the latest guidelines for Human-AI interaction (see Section 1.3.3), it becomes clear that they were very fundamental and provided the basis for further development of the HAI domain.

Two succeeding decades solidified the idea of AI adjusting to human needs. The industry finally adopted the topic of Human-AI interaction, bringing significant attention to its challenges.

The main challenges for the HAI, as Xu et al (2021). conclude, are:

- Explainability of machine output – typically explained in the general HCI approach, it is often overlooked in AI systems, so the user may not know how and why the AI system makes decisions.
- User-friendliness of interfaces – AI applications require interaction standards specifically developed for them.

To address these challenges, the authors propose a comprehensive Human-Centered AI (HCAI) approach covering three main aspects – human, technology, and ethics. Usable & Explainable AI is one of the focus areas for this framework.

Auernhammer (2020) uses Terry Winograd's distinction between "rationalistic" and "design" perspectives to examine various approaches to solve these central challenges. In the rationalistic perspective, the problem of explainable and trustworthy AI is addressed by governmental and research regulations. The author points out that regulations and laws do not reflect the real-world complexity and do not keep up with the development of technologies and new use cases in AI-based applications. In addition, they may not be universal enough to address ethical issues and provide explainability across various cultures. The proposed solution is the "design" human-centered approach that should clarify these issues through the fast prototyping, experiments, and empirical data collection. It is also highlighted that the Interaction design method can help to improve the explainability of the application through interface usability testing, thus connecting the two main challenges.

Lieberman (2009) examines the topic of user interfaces in AI systems. The author claims that the poor interface design of the applications is caused by teams more interested in learning algorithms under the hood than in the design itself. It leads to worse results of user testing and rejecting the usefulness of AI systems in general. He also advises to adjust the available HCI design methodologies considering the specifics of these algorithms. For example, AI requires more transparency and explanations about how it works to develop trust in users, i.e.: it is helpful to create more thorough tutorials and introduce the available features.

Yung et al. (2020) follow Lieberman's findings of the necessity of adjusting general HCI approaches when applied to AI. Researchers summarize a corpus of papers devoted to human-centered AI and define HCI practitioners' specific challenges in AI applications. One of the crucial - *difficulties in applying iterative prototyping and usability testing*. Two proposed approaches to address this challenge are:

- The "Wizard of Oz" method – testing approach where the human developer imitates AI during the interaction with a user
- Early-stage deployment of AI systems for real users.

The "Wizard of Oz" method helps to check users' behavior and scenarios, but it cannot predict or simulate AI-specific errors and failures. Early-stage deployment reveals both intended and unintended AI behavior and interactions with the end-users. Still, it is time- and effort-consuming and loses the benefits of "cheap" and rapid prototyping.

One of the solutions proposed is applying more "universal" Human-AI Interaction guidelines, which can be considered a checklist for the developing teams. These guidelines cover the already revealed in many academic and industry studies pain points of interaction, which should be addressed during the AI system development.

1.3.3. Human-AI Interaction: Design Guidelines

Jeniffer Sukis (2019) from IBM made a comprehensive overview of Human-AI Interaction design guidelines available as of 2019[2]. All the big players on the AI stage – Amazon, Facebook, Google, IBM, Microsoft, – developed and published detailed guidelines for various AI cases in open access, including text and voice chatbots. The main focus in all works proposed is user experience, explainability, and trustworthiness of the AI systems, emphasizing ethics and a human-centered approach.

It is not a goal of the present work to deeply examine all the available guidelines. For testing during the experiments, the interaction framework should be developed and

[2] https://medium.com/design-ibm/ ai-design-guidelines-e06f7e92d864 - accessed on 19.12.2021

implemented in a conversational AI model. To this end, the overlapping recommendations relevant to the current study were compiled based on two complementary and detailed guidelines. The first one, Microsoft's "Guidelines for Human-AI Interaction" (Amershi et al., 2019) provides a more industry-inspired perspective for the general HAI domain. The second one, is a comprehensive academic review of chatbot-specific guidelines, "How should my chatbot interact? A survey on social characteristics in human-chatbot interaction design" (Chaves&Gerosa, 2020). Both papers were reviewed, and guidelines for the interaction framework were chosen based on how well they matched the specific use case foreseen for tests in the experiment. Here, the following two aspects were considered:

- The experimental AI system is a text based chatbot that supports the user during a quiz game.
- The task is highly time-limited (7 minutes for the whole interaction with the chatbot) and takes place only once. Correspondingly, the long-term or repetitive interactions design principles are not applicable.

The chosen guidelines from both papers were compared, and overlapping recommendations were used as an interaction framework for the conversational AI model development. See the details in Section 1.4.

1.3.4. Human-Chatbot Interaction

Following Chaves&Gerosa (2020) definition, a chatbot is a "disembodied conversational agent that holds a natural language conversation via text-based environment to engage the user in either a general-purpose or task-oriented conversation." This definition highlights several key points about the interaction between people and chatbots:

1. The interaction is held via natural for humans interface – language.
2. The primary format of the interaction is text, and the agent is disembodied, so no additional communication channel (like mimics, body language or voice) is used.
3. In most cases, the interaction is oriented toward some goal or task.

As the authors point out, the users have high expectations regarding the chatbot's understanding of the context and the goal of the conversation. They await from the assistant that it will provide meaningful answers and handle the complexity of the task - in simple words, to be a handful when playing on the human language field.

In general, expectation management is highlighted (see Lugar and Sellen, 2016; Zamora, 2017; Kocielnik, Amershi & Bennett, 2019) as a crucial point affecting the users' satisfaction. If the expectations are not properly set and are too high, the user will be

disappointed even with the planned system behavior. On the contrary, if the user clearly understands what can be achieved by using the system, the evaluation and acceptance rate will be higher (Lubbe and Ngoma, 2021).

In the study by Zamora (2017), Google Search was used as a baseline for evaluating the chatbot interaction. The authors stress that users' evaluation of chatbot performance will depend on the standards implied by the prior experience of using alternative services. Considering that Search is based on two state-of-the-art technologies, complex Transformer neural networks and Google Knowledge Graph[3], and used widely, it can be a considerable challenge for all chatbot developers. Users perceiving Google Search as an everyday basic application may have higher standards for conversational AI.

However, if the standards are so high, should one try to imitate human conversation at all? Clark et al. (2019) point out that the promise to provide a "human-like" conversational experience can be a wrong one from the very beginning. Instead, chatbots developers, HCI, and HAI practitioners could focus on what people indeed value in the assisting agents. During the interviews with participants, researchers clarified what differentiates the conversation with a chatbot from a conversation with a human and which conversation attributes are the most valuable. The results show that people tend to perceive the interaction with AI more as a *transactional* conversation, aiming to solve some problem; chatbot itself a *tool* rather than a partner; and on the scale of social interaction – a *stranger* than a friend.

If the chatbot is just a tool, it makes sense to develop it as highly valuable; not just a FAQ chatbot that stands as a detour on the way to a human agent, but a convenient access to many personalized and secured services. Precisely these attributes participants highlight as desirable in the conversation with a chatbot – personalization, trustworthiness in terms of security, data privacy and transparency, and clear understanding of goals.

Nevertheless, both studies – by Clark et al. and by Zamora – mention one more highly human-like service that users could consider valuable for interaction with a conversational AI. This service is "*chatbot as a confidant*" – fulfilling social needs and getting emotional support. A perfectly personalized, actively listening agent could be an ideal companion when people need to relieve the emotional arousal, disclose sensitive or embarrassing issues, or "think out loud". Lucas et al. (2014) found out that patients' willingness to disclose their problems during a mental health screening is higher when they believe they speak with a fully automated virtual assistant, and not a real person. The participants explained that they have "no fear of being judged" in this case.

[3] https://www.blog.google/products/search/introducing-MUM/ - accessed on 08.01.2022

Chatbots are already on the market, addressing anxiety, mood disorders or guiding the users through cognitive-based therapy (CBT), such as Wysa[4] and Woebot[5]. The latter was tested in a randomized controlled study with 70 participants (Fitzpatrick, Darcy, & Vierhile, 2017), aiming to reveal how effective this chatbot-infused guidance through CBT would be. The study showed that subjects in the Woebot-group significantly reduced the severity of depression symptoms after two weeks compared with the control group. Among the features of the chatbot favored at most participants named:

- "Checking in" (chatbot often asked about emotions and feelings)
- Empathy (chatbot showed concern)
- Learning (chatbot provided valuable insights about emotions that people tend to feel and ways of thinking)
- Comfortable conversation (humor and sympathy incorporated in the chatbot's answers, though limited)
- The chatbot provides interactive and diverse content (videos, games, interesting suggestions, graphs).

CASA framework (Computers as social actors) assumes that computers can be perceived as equal and be attributed with human characteristics, including empathy (Nass & Moon, 2000). Liu & Sundar (2018) found out that *affective empathy* (demonstrating emotional sharing of the bad experience) and sympathy from the chatbots are perceived in a more positive way than just a *cognitive empathy* (detached confirmation that the experience may be troubling). Authors also point out that the "chatbot skeptics" evaluate such emotional expressions even higher than people with a better prior opinion about conversational AI. Still, other studies show that people may evaluate the same emotional support (following the same script) provided by the chatbot lower than by a human companion, if they know that their conversation partner was a bot (Meng & Dai, 2021). Prejudice against the chatbots due to the previous unsatisfactory experience is a problem to be solved by HAI practitioners.

1.4. Interaction Framework Development

In this section, the chosen HAI design guidelines are reviewed, aiming at defining a set of optimal recommendations and the interaction design framework for the practical part of experiments.

[4] https://www.wysa.io/ - accessed on 08.01.2022
[5] https://woebothealth.com/ - access on 08.01.2022

1.4.1. Guidelines for Human-AI Interaction, Amershi et al. 2019

Based on the review of more than 150 recommendations from academic and industry works, Microsoft researchers developed a set of design principles for HAI, which were validated and tested by 49 design practitioners against 20 AI-infused products. As a result of the validation, 18 final guidelines were formulated.

The proposed principles are temporarily organized depending on when they should be applied. There are guidelines for the beginning of the interaction, for interaction in general, recommendations on how to design for failure, and long-term interaction with users.

As researchers point out, the crucial point for creating transparency is the beginning of the interaction. Guideline 1 “Make clear what the system can do” and Guideline 2 “Make clear how well the system can do what it can” match the main principles of human-centered design, examined in Section 1.2. The more thorough user onboarding can make the UX more successful. Initial messages and hints provided by the system with the first interaction/start of the interaction should create a clear understanding of which functions are available for the user and what results can be expected.

Among the further 16 guidelines, not all were relevant for the experimental task of the current work, as many of them are aimed for the long-term or repetitive interactions or just proposed for other use cases. For example:

- G3 “Time services based on actions” – the chatbot assists the user only during the test based on the user input, so it should not provide guidance on some specific time
- G6 “Mitigate social biases” – chatbot does not appeal to a user with any gender-specific suggestions or content
- G12-G18 like “Remember recent interactions”, “Encourage granular feedback,” or “Notify users about changes” – as the interaction is only a one-time experience, these guidelines do not apply.

Overall guidelines chosen can be summed up in the following points:

- Make clear what the system can do
- Make clear how well the system can do what it can do, what are limitations of the system
- Make clear why the system did what it did, provide an explanation in case of failure
- Interact appropriately (use semi-formal tone for the interaction)
- Give the user the control over the decision to use or to ignore the recommendations provided by the system.

It is important to mention that the last guideline was implemented in the design of the whole task itself and not only in the “improved” interaction model. Users should had the possibility to reject the chatbot assistant’s suggestions and make their own choice with both models to create a more real-life experience of seeking help.

1.4.2. A Survey on Social Characteristics in Human-Chatbot Interaction Design, Chaves&Gerosa 2020.

This study focuses on chatbots specifics, as interaction based on a natural language is framed with some initial expectations from the users. As a foundation for the research, 56 chatbot-related works in different domains were reviewed and summarized in 11 social characteristics that chatbots should possess to provide a better user experience.

The social characteristics revealed are united in three groups – Conversational Intelligence, Social Intelligence, and Personification, – with the following structure:

1. Conversational Intelligence

a. Proactivity – provide additional information, inspire users, recover from a failure, leverage the conversational context

b. Conscientiousness – demonstrate understanding, provide confirmation messages, support conversational flow

c. Communicability – explain the functionality and manage users’ expectations

2. Social Intelligence

a. Damage Control – deal with the unfriendly users and appropriately respond to harassment, deal with lack of knowledge

b. Thoroughness – increase human likeness and believability

c. Manners – engage in small talks, start and end conversations gracefully

d. Moral agency – avoid stereotyping and alienation

e. Emotional intelligence – use emotional utterances to demonstrate empathy and understanding and to improve human-chatbot relationships (Li et al., 2017)

f. Personalization – learn about the user, provide unique services

3. Personification

a. Identity – design and elaborate on a chatbot persona

b. Personality – use appropriate language and possess a sense of humor.

Some social characteristics are again tailored for long-term and repetitive interactions, such as “Personalization”. “Damage Control” also did not match the current research goals and the used interaction setup (experimental task conducted in the lab settings), as the likelihood that the participants will try to harass the chatbot was low.

It is also important to note that in the RASA conversational AI model used (see Section 2.2 for more details), some of the social characteristics described are implemented by design – like “Proactivity” and “Conscientiousness”. Those were not considered a part of the Interaction framework, as these features would be implemented by default in both conversational models.

Consequently, the following guidelines were selected and used:

- Explain functionality
- Manage users’ expectations
- Keep user aware of the chatbot context, explain the decisions
- Increase human likeness, use appropriate language and manners
- Use emotional utterances to demonstrate emotional intelligence, empathy and understanding

When matching both sets of the guidelines chosen, it became apparent that most of them (four out of five) are similar and represent the same guidelines with slightly different phrasing. Besides these four, the two additional recommendations were “Give the control to the user” (by Amershi et al. 2019), which, as was already discussed, should be implemented for both of the models, and “Use emotional utterances” (by Chaves&Gerosa, 2020) – see fig.2.

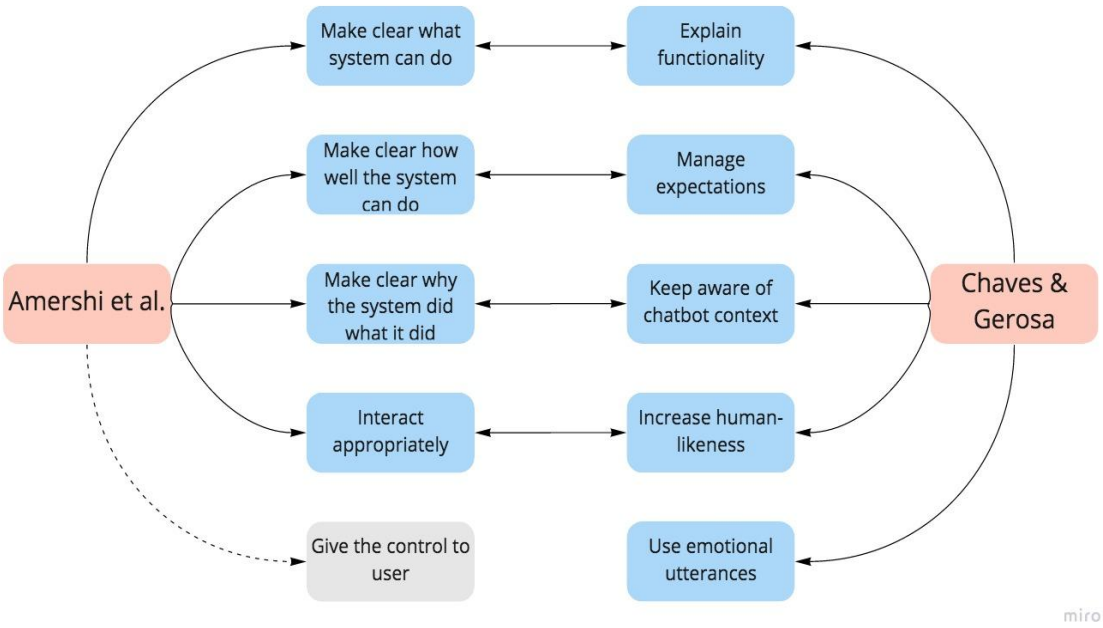

Fig. 2. Guidelines Set for Interaction Framework.

1.4.3. Interaction Framework Formulated

Based on the guidelines reviewed, the interaction framework was formed as follows:

- At the beginning of the interaction (ideally in the greeting message) – provide a clear explanation of functionality and manage the user's expectations:
  - Explain what the chatbot can do (it can help with the quiz, it has only one purpose)
  - Explain how the chatbot "understands" the meaning of the question (the question is recognized based on the training data learned)
  - Explain how to use it efficiently (it is better to use more straightforward phrases with the most meaningful words to improve the input recognition)
  - Explain the limits of the system (if the chatbot cannot answer the question several times, it makes sense to move forward to the next question)
- During the interaction – Communicate in an appropriate tone and increase human-likeness:
  - Use semi-formal language – avoid being too "robotic" and formal, but do not try to use slang; it will be unconvincing
  - Use a gender-neutral personal name
  - Use emoticons
- During the interaction – use emotional utterances:
  - Use affective utterances to demonstrate empathy and understanding and to improve the contact with the user
- "When wrong" (when system provides a wrong answer or fails to provide one) – explain why the system did what it did and keep aware of the chatbot context:
  - Explain that the chatbot cannot recognize the question and ask to rephrase it.

Implementation of the framework through conversational AI is explored in detail in Section 2, "Conversational Model Development".

1.5. Sympathetic Arousal (Stress) Measurements

In this section, subjective and objective methods of UX research are reviewed, and physiological measures applied to usability testing are described. The section also includes a short description of the method chosen for the current research (see more details in Section 3 "Method" and Section 4 "Results").

1.5.1. Objective and Subjective Methods of Usability Testing

Previous sections highlighted that UX research takes a central role in the HCI and HAI domains. It is a common approach in the modern development of new products and services. Traditional methods and metrics of user experience and usability research include[6]:

- Quantitative: A/B testing, web-analytics, time on task, error/success rate, surveys, and questionnaires,
- Qualitative: interview, questionnaires (open questions), heuristic evaluation, and observation.

These methods allow to evaluate users' response to a prototype/changes in a product or service quickly and understand the user journey and needs. The downside is that many of the approaches mentioned above, such as questionnaires, interviews, and observation, are subjective and, therefore, can be considered as less reliable (Yao et al., 2014; Zaki&Islam, 2021). Also, they do not provide direct data about the psycho-physiological state of the user, including affective states and emotions that emerged from the interaction. Still, as explored in Section 1.1, emotional aspects and user's affects are crucial for the design process, especially "designing for an error".

A field of *NeuroIS* emerged in 2007, aiming to apply neuroscience and neurophysiological tools to the research of information and communication systems (Riedl&Leger, 2016). This approach provides a quantitative objective understanding of how IT systems affect users' behavior, adding a biological level of analysis to the studies. During the empirical NeuroIS research, such neurophysiological methods as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), hormone assessments, skin conductance, heart rate measurement, eye-tracking, and facial electromyography (EMG) are used.

The HCI community and UX research followed NeuroIS by applying objective methods of evaluation of user involvement. Zaki&Islam, 2021 provided a comprehensive overview of the related studies published between 2003 and 2019, showing that the adoption of these methods is growing in the HCI community. Among measurements examined in the review are:

- EEG,
- electrocardiogram (ECG),
- EMG,
- facial expression tracking,

[6] User research methods - a comprehensive guide https://www.userzoom.com/ux-library/ux-research-methods-a-comprehensive-guide/ Acessed on 04.01.2022

- eye-tracking,
- EDA,
- heart rate variability.

Examples of industries and products tested with the physiological measurements include: gaming, advertising, social media, e-commerce, web applications, biometric identification systems, call center environment. Still, conversational AI was not covered in this review.

Concerning the physiological measurements used for testing chatbots, a few studies can be found. In the research by Ciechanowski et al. (2019), a chatbot was tested with EEG, EMG, and EDA. The study examined the physiological measures, and applied traditional questionnaires to assess if the chatbot with a human-like visual avatar causes the "uncanny valley" effect on the participants. The questionnaires showed that the avatar-based chatbot is perceived more negatively than the text-based one. The EDA and ECG data showed higher levels of emotional arousal and heart rate, providing more details about users' reactions to specific chatbot answers and interaction points.

Another example is a recent study by Yen&Chiang (2021), where the purchase activity of the customers speaking with the chatbot was controlled by EEG data. 30 participants took part in the research, which also included a traditional self-assessment.

Finally, the NeuroIS Society supported this trend with a pilot study by Carmichael et al. (2021), in which 14 participants were interacting with a chatbot online while an automated framework analyzed their facial expressions.

These few examples highlight untapped opportunities for further work in this area, which can enrich the research of Human-Chatbot interaction with objective physiological data.

1.5.2. Electrodermal Activity Measurements

EDA is a method of defining the levels of skin conductance by measuring the current flow between two skin points, between which an electrical potential was applied (Braithwaite & Watson, 2015). The EDA data includes background tonic (skin conductance level, SCL) and phasic (skin conductance response, SCR) components. These signals reflect the activity of the sympathetic nervous system, associated with emotional arousal in the following way: sweat glands, especially the ones located on the palms, are sensitive to noradrenaline. Psychological and emotional affects cause the activation of the sympathetic nervous systems by noradrenaline transmission; sweat glands answer to the increased noradrenaline levels and, in turn, increase skin conductance levels due to the produced sweat.

Colleagues from the “Competence Team for implanted devices” of the Medical University of Vienna ran several studies (Bijak et al., 2019; Deubner, 2019) proving that EDA can be used as a reliable method of stress (sympathetic arousal) monitoring. The number and amplitude of the SCR peaks can be associated with the levels of the experienced psychological arousal. Event-related SCR peaks also define the response to a specific stimulus presented. Data preprocessing is required for splitting the one signal obtained from the sensors into two components (see fig.3: red is phasic SCR component with the peaks, blue is tonic SCL component).

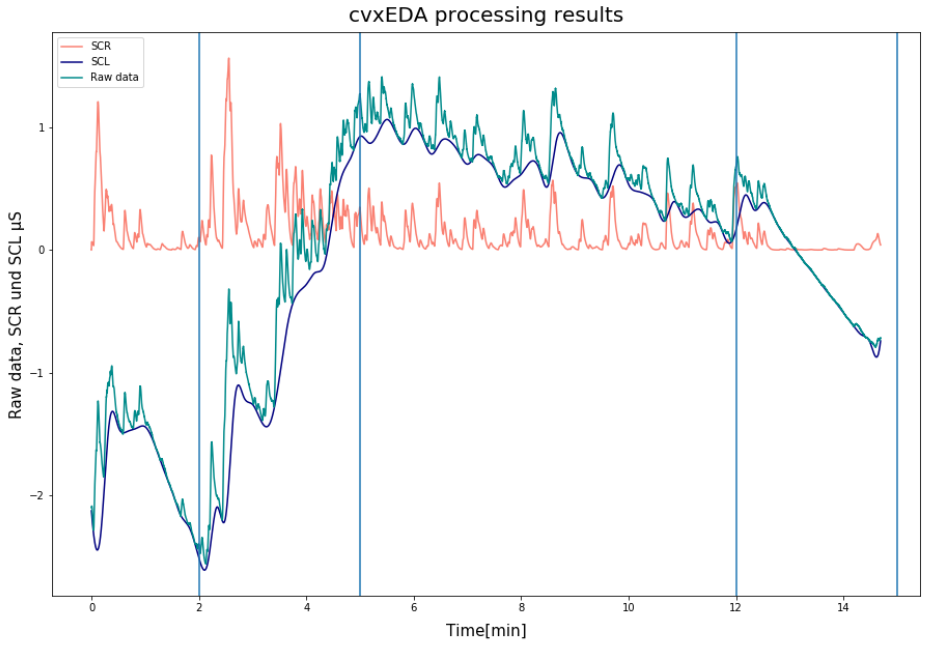

Fig.3. Raw EDA Data from Sensors and Splitted SCR and SCL Signals.

Different biometric devices to measure EDA are available on the market; in most cases, they combine several various methods of physiological measures (EDA, heart rate, ECG, EMG, others). For the present research, a medical certified Neuromaster® device was used (SOFT®med system, Insight Instruments[7]). The device measured the sympathetic nervous system activity during the experimental sessions. The obtained signal was pre-processed and split into SCL and SCR components. SCR amplitudes provided valuable objective data about the difference in the participants’ experience of dealing with the default and the improved interaction design models.

[7] https://biofeedback.co.at/international/

Summary

This section presented a foundation of the master thesis, explaining the main concepts such as Postcognitivism, Human-Centered Design, Affective States, Human-Computer and Human-AI Interaction, and physiological methods of measuring user experience, which created theoretical and methodological framework of the research. In the following sections, the empirical part of the study is described, including implementation of the interaction framework in the Conversational AI model, experimental setup, experiments, and their results.

2. Conversational Model Development

This section provides details about the conversational AI model development – how the dataset for the experiment was compiled, which natural language processing (NLP) open source framework was used as a foundation for the model, how the infrastructure and architecture of the model were built. It also describes how the interaction design formulated in Section 1.4 was implemented in the chatbot behavior, and which specific steps were fulfilled to achieve the required performance of the model. The section also presents the insights on the development of the web application used for running the experiment.

2.1. Quiz Dataset

To provide a task for the participants to solve during the experiment, a dataset consisting of 30 questions chosen from the "Jeopardy!" dataset[1] was compiled. The questions were marked as "easy," "medium," and "hard" based on the assessment of their difficulty level for participants. The assessed complexity level was validated in a preliminary survey that involved 20 participants. Six questions used for the validation were removed from a set, yielding the final set, consisting of 24 questions, was used in the experiments presented below.

During the experiments, the web application randomly retrieved ten questions from the entire database, half of which were from the "easy" group and a half from "medium"/"hard." This approach was introduced to ensure that, on the one hand, participants would not be able to answer *all* of the questions without the support of a chatbot. On the other hand, a participant could still answer at least some of the questions based on her prior knowledge – a more realistic setting of seeking external assistance.

2.2 Conversational Model

2.2.1 RASA Open Source

RASA Open Source is a Python-based learning framework for conversational AI models development and implementation (Bocklisch et al., 2017). Complete documentation can be found on https://rasa.com/docs/rasa/

RASA Open Source consists of several main modules (see Fig. 4):

- RASA NLU (Natural Language Understanding) contains an NLU pipeline, which supports intent classification and entity extraction. *Intent* is the main topic of the message like “hello” or “help me”; *entity* is a predefined language category, e.g., “DATE,” “PERSON,” etc.

[1] https://data.world/sya/200000-jeopardy-questions, last accessed 28.11.2021

● RASA Core containing Dialog Policies – set of scripts defining the chatbot behavior depending on the intents and entities recognized

● RASA SDK supporting Action Server – server for running Python scripts implementing chatbot behavior (returning the coded answers, calling API, etc.)

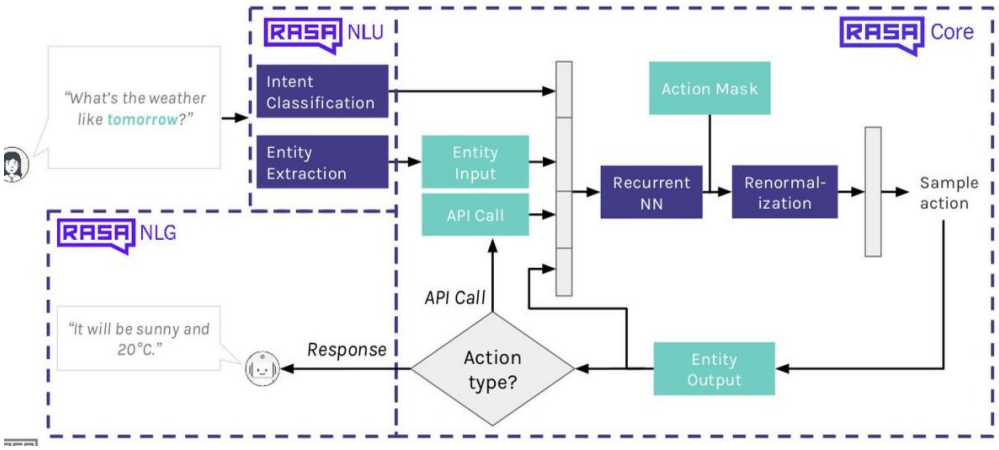

Fig. 4: RASA Open Source Modules. Source: Justina Petraitytė

RASA NLU. User inputs are passed through the NLU pipeline containing various components – pre-trained language models (see Fig.5):

- Tokenizers – split messages into tokens, natural language units, required for further processing
- Featurizers (vectorizers) – convert language tokens into vectors in a feature space of the model
- Intent classifiers – define similarity between the vectorized input and the provided in the training data examples of different intents, predicting which intent was provided by the user (e.g., “request_weather”)
- Entity extractors – extract defined language categories from the text (e.g., “DATE”).

The pipeline is defined in the *Config* file of the model and can contain various combinations of the components described above. As various tokenizers, featurizers and classifiers can perform differently on each specific dataset, the pipeline sequences can be compared during *model evaluation* and tuned (see the part *Model evaluation* below).

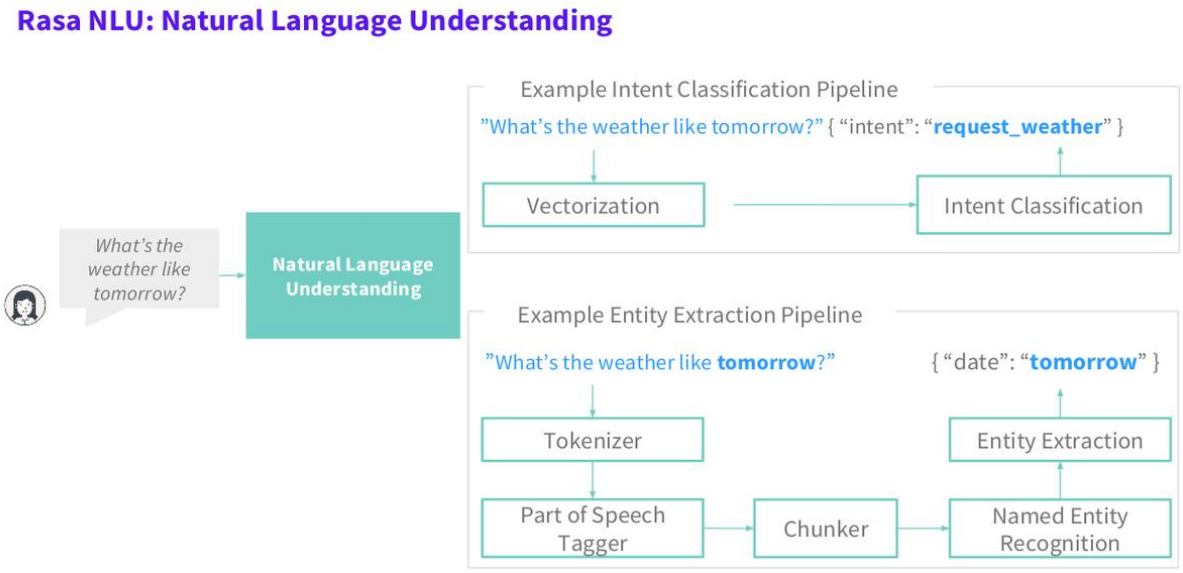

Fig.5. RASA NLU Module. Source: Justina Petraitytė

Before the training, the modeler should provide the training data. The YAML format is used for the training data as a more "user-friendly" structured alternative to the markdown. The training data in RASA framework consist of the following types, usually separated between several YAML files:

- *NLU data* – contains examples of possible user inputs marked as different intents/entities. For example, intent "greet" can be provided with examples such as "Hey", "Hi", "Hey there", "Hello", "Good afternoon", "Good morning", all of which will mean the greeting. During the training, the model learns the features of the intents described in the training data and how to distinguish them in the vector space.
- *Stories* – structured examples of the conversations between the user and AI assistant, containing a sequence of user intents and chatbot *actions* that should be returned. The actions can either contain a text response (*utterance*) or trigger some script (*custom action*) that will be processed inside the Action Server. Stories represent rather a recommendation and guidance for the chatbot than a strict algorithm. For implementing a strictly defined question-answer/action pair, *Rules* should be used.
- *Rules* - contrary to Stories, it is a fixed sequence of inputs and chatbot response actions. It is recommended to use Rules only for short and very specific conversation patterns, as they cannot generalize on unseen interaction examples, like Stories.

- *Domain* – a specific YAML file containing the whole "universe" of the chatbot, including all intents, entities, actions, and responses. Responses are also called *utterances* and are usually marked as "utter_greet" or "utter_goodbye".

After the training is finished, the trained model is saved into the directory *models* and can be evaluated and re-used later.

RASA Core. Dialogue management and Policies define how the chatbot behaves in reply to the recognized intent/entities. *Policies* are machine learning (ML) or rule-based models that predict the next chatbot's action that should be implemented. The examples of ML policies are *TED (The Transformer Embedding Dialogue)*, a multilayer neural network of transformer encoders used for both action prediction and entity recognition (Vlasov, Mosig, Nichol, 2019); *MemoizationPolicy* – a model that remembers Stories from the training data and tries to match the provided intent to the memorized stories. If successfully, the model triggers the next chatbot action specified in the Story. It is possible to configure the number of interactions for which the model should check if the story matches.

RulePolicy is an example of rule-based policies; it tries to predict the next action based on the Rules provided in the training data. If the model confidence that it "knows" the right next action is higher than a configured threshold, the predicted action will be executed; if it is lower than the threshold, the model will trigger a *fallback* action. Fallback is specified by the modeler set of actions "*designed for the error*". An example of such a fallback - chatbot, asking to rephrase the previous question or proposing to hand over the user to a human agent.

During each user-chatbot interaction turn, all policies specified in the model Config provide a certain level of confidence for the action predicted. The policy with the highest confidence level will be chosen. If two policies have the same confidence level, the action is chosen based on the priority; by default, RulePolicy has the highest priority, followed by MemoizationPolicy and finally TEDPolicy. This ensures that Rules as more strict conversation patterns will be followed in the first place.

RASA Action Server. Action Server is required for executing the actions specified in the training data. When an action is triggered, the model sends a request to the server, and receives defined events and responses. Some actions are in-built in RASA by default: *action_session_start* (start the session and reset the conversation tracker), *action_listen* (await the next input from the user), *action_default_fallback* (triggered by low prediction confidence, sends the defined by modeler fallback response), etc. *Custom actions* are

scripts that run any code required: they can send API calls to the external systems, retrieve information from the database and so on. For Action Server, a default server provided by RASA Open Source can be used (written in Python) or any custom server (that can be written in any language).

Model evaluation. There are two main levels of evaluating the model performance: evaluation of NLU model and NLU pipeline.

To evaluate the NLU model, a classical approach with splitting data into train and test sets as 80/20 can be used. It is also possible to implement a cross-validation testing, where data will be multiple times reshuffled and splitted into training-test sets. In-built in RASA rasa test method is used.

As an output, the test script returns an *intent classifier report* (containing evaluation metrics as *precision, recall* and *f-1 score* for each of the intents), *confusion matrix* (showing which intents are confused with other) and *confidence histogram* (illustrating the confidence level of the Dialogue management Policies for each of the intents).

Intent classifier report shows how good the model performs in predicting different intents, and if it is too strict and has not enough confidence (the confidence threshold is too high), too loose and overconfident (the confident level is too low), or just right.

It is also possible to compare different NLU pipelines. If several pipelines are configured, multiple configuration files can be passed as arguments to the rasa test method. It will split the data into train and test, train each of the configurations separately on a subset of the training data, and then validate it on the test dataset. The provided f-1 score graph will show the change in the performance between different NLU pipelines.

2.2.2. Model Architecture and Environment

Conversational models for the experiment were built on the RASA open-source model. To get enough computational resources and provide easy integration with external systems (such as a chatting window and a database), the infrastructure was developed as fully cloud-based. The infrastructure included the EC2 instance based on the Frankfurt AWS data center, Ubuntu OS, Ubuntu Desktop, and TightVNC installed on the virtual machine. Such a set allowed using a fully remote environment from the macOS laptop used in the experiments via VNC Viewer.

The conversational model was developed and tested locally on the virtual machine using RASA Command Line capabilities, and later tunneled via https://ngrok.com/ (a solution for secure tunneling localhost addresses to the web) to establish a public endpoint. Through the websocket channel, a connection with an open-source chat widget for RASA https://github.com/scalableminds/chatroom was created. The last required

integration was to connect an external database for Tracker Store, a RASA component that stores all chatbot logs. For the Tracker Store, a MongoDB Atlas database was used. The whole architecture and environment of the model are reflected in Fig. 6.

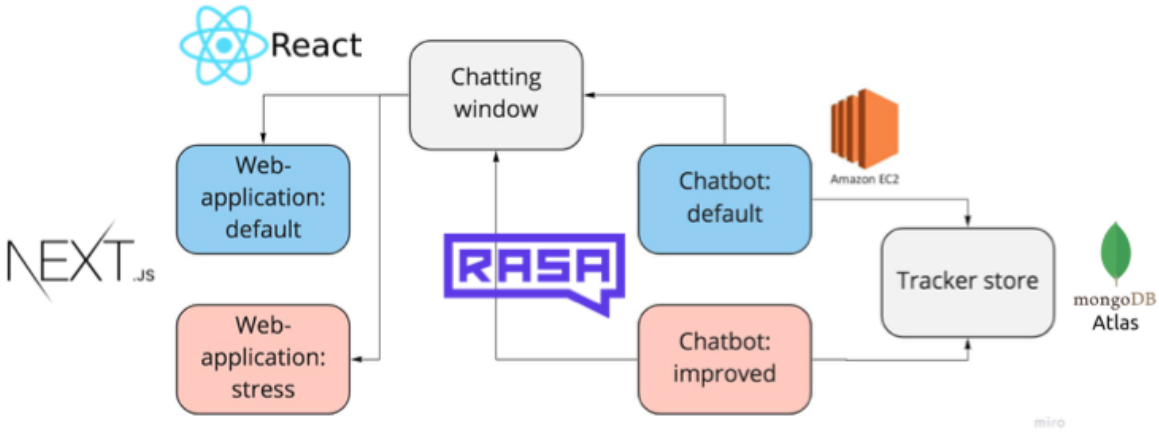

Fig. 6. Conversational Model Architecture

2.2.3. Implementation of the Interaction Framework

The model was developed taking into account the following main conditions:

- The NLU component of both models should operate the same way whether default or improved interaction design is used so that the user experience will be distinguished by the interaction design and not by NLU performance
- The difference between the default and improved interaction design should be based on how transparent and understandable the model is for the user. A variable that should be manipulated here is the chatbot responses (utterances), including the greeting message at the beginning of the interaction.
- Situations, where it does not matter if the user understands how the system works (based on the information provided by the model) or not because it consistently performs ideally, should be avoided. It will not allow distinguishing the user experience between the default and improved model.
- According to the previous condition, it should not be possible to receive a correct expected answer from the chatbot just by copy-pasting the quiz question from the system.
- Still, the model should provide a correct answer for each of the questions in the dataset.

To summarize the conditions above, the goal was to create a “not perfect” system that operates with the same level of performance for both models. In the case of the

improved interaction design, the model provides more information about how it works and why helping users interact with it more effectively.

2.2.4. Model Development

For this research, a demo RASA model "Financial Services Example Bot" was used and updated (https://github.com/RasaHQ/financial-demo). The pre-built demo model contains all the required components – project structure, training data examples, rules, stories, actions, NLU pipeline, and config file examples.

The development process looked as follows:

1. Remove all the financial-related data and actions
2. Provide new training data examples based on the actual dataset
3. Create new rules, stories, and utterances for the chatbot adhering to foreseen interaction scenarios and the experimental settings.
4. Integrate with external database and chat window
5. Train the model and evaluate the performance
6. Optimize the performance
7. Create two instances of the model with default and improved interaction design

Cleaning the data. To speed up the cleaning process, the demo model was tested against questions from the quiz dataset. The triggered financial-related intents, stories, entities, and actions were removed from the file system.

Provide new training data examples. For each of the questions, at least 2-3 different examples of phrasing the same question were provided, e.g.:

- Original question: "The strawberry is not a member of the berry family but is, in fact, a member of this garden flower family."
- Provided training examples:
    - To which garden flower family does strawberry belong?
    - If strawberry is not a berry, then what is it?
    - Strawberry is a member of which flower family?

The overall approach was to train the model based on the most meaningful words from the question. "Jeopardy!" questions are originally phrased in a tricky way. It pushes the user to interact with the chatbot following its instructions and not just copy-paste questions as they are.

Create new rules, stories, and utterances for the chatbot. Responses to intents related to the quiz game questions were specified as Rules, as the chatbot must provide an exact answer to each recognized question. For some general and chit-chat responses as greetings, answering the questions like "Are you a bot or a human?", "What can you do?", Stories were applied. This allowed to provide a more natural conversation

experience and generalize on unseen conversation examples, as each user can ask such questions differently.

Utterances were the primary means of interaction design implementation. See the comparison of variants of utterances in the “Default” and “Improved Design” models in Table 1.

Utterance	Default design model	Improved design model
utter_greet	Hello! I am your virtual assistant.	Hello! My name is Alex. I am your virtual assistant
utter_help	I can help you with the quiz	I can help you with the quiz. I predict the intent of your message based on the data I learned. Please, try to use simple phrases with the most meaningful words for the questions. If I can’t understand you, I will ask to rephrase the question. If I cannot answer it several times, maybe it is better to move forward. But I will do my best :)
utter_fail	Sorry, I didn’t get	I didn't get that. Can you please try to rephrase it?

Table 1. Utterances for Default and Improved Interaction Model

Integrate with external database and chat window. The model can be developed and tested entirely locally via the command line. Nevertheless, the external Tracker Store allows to keep and observe all the model logs conveniently. For tracking, a MongoDB Atlas database was chosen. RASA architecture provides fast and easy integration possibilities based on the database public endpoint and login-password pair.

```
27: Object
    event: "user"
    timestamp: 1622403663.3074431
    text: "Which animal has the highest blood pressure?"
  parse_data: Object
    intent: Object
        id: 601252575146453779
        name: "giraffe"
        confidence: 0.9601080417633057
    entities: Array
      text: "Which animal has the highest blood pressure?"
      message_id: "56d56228544f477888590eac49ac73ad"
    metadata: Object
    intent_ranking: Array
      0: Object
          id: 601252575146453779
          name: "giraffe"
          confidence: 0.9601080417633057
      1: Object
          id: 7332175727080623487
          name: "help"
          confidence: 0.01934649981558323
      2: Object
          id: -4720823334067000497
          name: "karaoke"
          confidence: 0.004401422571390867
```


Fig.7. Session Logs in MongoDB Database.

Streamed logs (see Fig.7) contain information about all the messages exchanged between the chatbot and the users. During the development, logs containing confidence levels for different predicted intents were especially significant. This tracking allows seeing when intents are mixed up and understanding the training process better.

Integration with the chat window was done via websockets and ngrok tunneling: ngrok allows the creation of a publicly available endpoint for the localhost address. This endpoint was saved in the chat window code snippet and inserted into the HTML header of the web application. As a result, the chat window connected with the required model was available on the web application homepage (see Fig. 8).

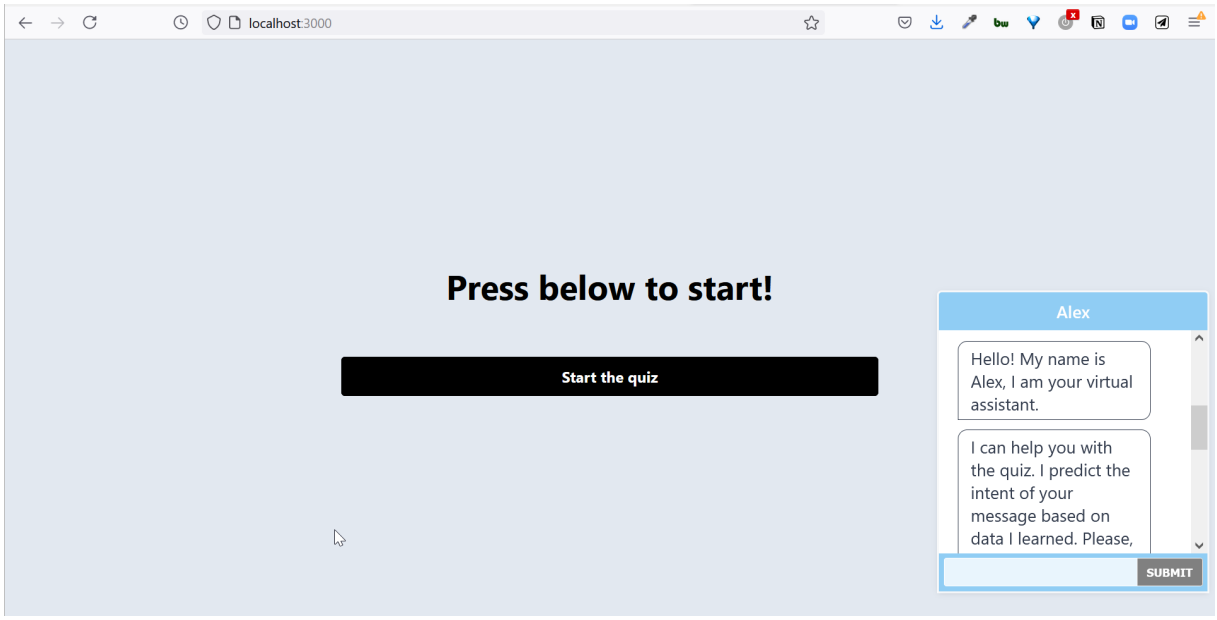

Fig. 8. Web Application Homepage with an Active Chat Window

Train the model, evaluate and optimize the performance. Initial training revealed two problems. Firstly, for some questions, the model could not provide a correct answer on both plain copy-pasting or rephrasing - these questions failed in general. Secondly, for others it always responded correctly even with the whole copy-pasted question. Both cases were against the design framework implementation conditions described in section 2.2.3, so it was essential to eliminate such undesirable behavior.

NLU model evaluation showed that in general the model performs well: for most of the intents, the precision, recall and f-1 score were close to 1.0. Only one quiz-related intent was misclassified in the NLU model, a question about Lewis Caroll: “As well as kids' books, this 19th century author wrote "Examples in Arithmetic" & other math textbooks”. At the same time, the confidence level for prediction of some intents was low: several intents were classified with confidence 0.63, 0.51 and even 0.49, all below the threshold.

To optimize the performance, the following improvements were made:

- *nlu_fallback confidence threshold* was lowered from 0.7, which is too high for a not production-ready but research chatbot model, to 0.5. This change provided a better performance for the “non-answerable” questions.
- For some questions, additional training examples helped to solve the problem and make all the questions answerable.
- To reduce the number of questions that were successfully answered even with the copy-pasting, duplicate “false” intents with similar words were created. In this case, the user had to rephrase the question and choose meaningful words from the “right” intents, which nudged users to interact with the chatbot more to get feedback from the model.

The final model had the maximum precision, recall and f1-score for all the intents, and all of them were predicted correctly. At the same time, for 30% of the dataset questions it was possible to get the proper answer from the chatbot only by rephrasing the original question into the short version.

Create two instances of the model with default and improved interaction design. For the improved interaction design model, only the chatbot responses were changed; as utterances are not included into the training data YAML files, but rather into the domain file, it didn’t require the re-training of the NLU model. Two different model versions were separated and stored in two git branches. Depending on the experimental group the participant was assigned, the corresponding git branch was activated and default or improved model was uploaded to the local server.

2.3. Web Application

The web application used for the quiz phase of the experiment was built using React and NextJS frontend development frameworks and run locally. Quiz questions were stored as a Javascript objects array, where questions and answers were implemented as keys as a JS.file in the project directory. By using the randomizing function, ten question-answer pairs were retrieved from the array and sorted. The whole application was styled with CSS using the Tailwind CSS library.

On the starting screen, the user was presented with a welcome message and the button "Start the quiz" (Fig.8). A chat window to interact with the virtual assistant was also available. The participant needed to start a conversation with the chatbot before starting the quiz to receive the model's instructions in the greeting message. The Chat window was implemented as a simple HTML iFrame pointing to the chatbot endpoint.

Two web application routes, "/timer" and "/no-timer," provided two React components. The route was chosen depending on the experimental group to which the participant was assigned. "/timer" route included a counter that started after pushing the "Start the quiz" button. Initially, the counter time was equal to five minutes, but participants reported that it was impossible to answer all ten questions during the given period during the pilot study. Therefore, the time was extended to seven minutes.

Quiz questions were displayed in a multi-step "Wizard" form implemented with the React Final Form library. The user received the questions one by one and could not return to the previous one answered. The results were shown as "correct"/" incorrect" tabs below the Wizard and were updated after each answer submission. If the user successfully answered all the ten questions, an animation with fireworks was congratulated.

The history of the conversation with the chatbot was kept during the whole quiz phase and was reset together with the quiz progress only when the page was refreshed. The user answers were stored in the PostgreSQL database.

The next section “Method” describes how the model developed was applied in the experiment and used for collecting physiological measurements.

3. Method

This section describes the design and implementation of the experiment: the sample statistics, experimental design, main variables and covariates, exclusion criteria. The experimental procedure, laboratory setup and questionnaires are also presented.

3.1. Participants

For the experiments, 33 participants have been recruited: 17 females and 16 males. Age distribution: 48% of the participants were 25-34 years old, 33% – 35-44, 18% – 18-24. 94% (31 participants) had a C1/B2 level of English proficiency, and two persons (6%) were native speakers. In terms of IT expertise, 36% described themselves as professionals, 36% reported advanced users, 21% – confident users, and 3% (one person) – beginner level. Professional chatbot experience was one of the exclusion criteria, so all the related participants were excluded from the experiment. Among others, 64% reported that they interact with the chatbots as users from time to time, 30% said that they interact often. Two persons (6%) never interacted with the chatbots before the research.

The self-reported stress level was “moderately stressed” for 64% of the participants and “mildly stressed” for 33%. 3% (one person) specified “highly stressed.” No participants were excluded based on the simplified stress evaluation (see section 3.2.3 Exclusion criteria).

3.2. Experimental Design

For the experimental design of this study, a 2x2 between-subject, *full-factorial design* was chosen, taking into account the following considerations:

● It is necessary to define the extent to which each of the two factors – improved interaction design of the chatbot and added stress – influence the levels of sympathetic arousal of the participants (target variable) and each other (two-way interaction). Using the full-factorial experimental design and a two-way ANOVA data analysis, it is possible to define these effects.

● Full-factorial experimental design allows defining the effects even on moderate samples, providing at least one observation for each of the combinations of factors (Dean et al., 2017)

3.2.1. Factors Levels, the Dependent Variable, and Experimental Groups

Interaction design model: default (-1) and improved (1)

Stress factor (timer): no added stress (-1) and added stress (1)

Skin conductance response – dependent variable

Overall there are four experimental groups: “No Stress-Default model” (control group); “No stress-Improved model”; “Stress-Default model,” and “Stress-Improved model.”

The subjects were assigned randomly to the experimental groups based on an equal distribution of participants.

Possible sources of variation, which should be considered during the data analysis or used as an exclusion factor for the study, were accounted for as described below.

3.2.2. Source of Variation

Individual differences in the skin conductance levels: due to such factors as activity of the sympathetic nervous system, number and activity of the sweat glands, and similar physiological factors, the absolute levels of the participants’ SCR can be different (Brettlecker, 2019; Deubner, 2019). To count on this covariate, it is necessary to:

1) standardize the data between subjects and

2) include baseline SCR measurements into the experiment through the block design (baseline – stress phase – baseline).

3.2.3. Exclusion Criteria

To avoid distortion of the results due to participants’ background and general nervous system conditions, the following exclusion criteria were defined:

- exposure to chronic stress;
- professional experience with the chatbots;
- age above 44 years old;
- English proficiency below B2;
- anxiety-associated disorders and heart diseases.

Exposure to chronic stress: subjects exposed to chronic stress may react abnormally to the used stress factor. The current study did not aim to fulfill a thorough stress assessment but rather a rough estimation of significant stress conditions. Therefore, a simplified evaluation of severe and moderate stressing events based on the Holmes and Rahe (Holmes & Rahe, 1967) stress scale was included in the pre-test questionnaire.

Professional experience with chatbots. People working on the chatbots will understand how the system works and why it works much better, despite the interaction design approach applied. This factor may affect the subjects' SCR and should be an exclusion factor in the pre-test questionnaire.

English language proficiency below B2 – the participants should understand the questions from the "Jeopardy!" dataset, which are intentionally formulated in a tricky way. Therefore only people with enough language proficiency should take part in the research.

From the ethical considerations, *anxiety-associated disorders and heart diseases* were also chosen as exclusion criteria: the subjects could experience mild stress during the experiment, so it was essential to ensure that the test would not trigger severe conditions.

3.3. Experimental Procedure

1. First, participants signed the informed consent. Then the Neuromaster SCR sensor was attached to the point (preferably) finger of the participant's non-dominant hand. The measurements started with a non-recorded acclimatization phase of 1-minute length, during which it was required to check the SC signal levels and make sure that the signal was appropriate. During the acclimatization phase, participants were able to find a comfortable position that would ensure a minimum possible amount of movement during the experiment.

2. The overall procedure was shortly described, then recording started, and the subject was watching the first 2-minute baseline video excerpted a footage containing natural landscapes filming and neutral instrumental music[1].

3. After the first baseline, participants became familiar with the web application and received additional instructions for the quiz phase. Participants could either answer questions themselves or ask the chatbot for help, and they were was not allowed to google the answer. To create the same conditions for all the participants, they were asked to use mouse buttons for the copy-pasting instead of the keyboard hotkeys. It was also possible to type the questions and answers. If the participant had no further questions, she was allowed to start a conversation with the chatbot assistant, read its instructions, and then start the quiz.

4. Depending on the experimental group (Stress / No stress), the participant had a visible countdown timer for 7 minutes or was asked not to bother about the time. In the second case, after 8 minutes, the researcher notified the participant that the time was over if the game was not finished.

5. After the "stress" phase, the second 2-minute baseline video was presented. At the end of this phase, the SC measurement was stopped, and the sensor was removed.

6. As the last step, the participants filled the short post-test questionnaire, after which they could ask any questions about the experiment.

[1] https://youtu.be/2OEL4P1Rz04, last accessed: 28.11.2021

3.4. Post-test Questionnaire

The post-test questionnaire was based on The Usefulness, Satisfaction, and Ease of Use Questionnaire (USE, Lund, 2001). Two questions were allocated to each dimension to check the consistency and have more data, keeping the questionnaire short. Additionally, a question about the self-evaluated levels of stress during the experiment was introduced. Post-test questionnaires were printed and provided to the participants during the experiment.

3.5. Experimental Setup and Data Acquisition

For running the experiment, two laptops were used. On the MacBook Pro 2011, a virtual machine with a running conversational model was connected, and baseline videos were played. On the Windows-based laptop with a mouse controller, the Neuromaster software and web application for participants were installed. The Neuromaster device with the attached SCR sensor was as well connected with the Windows-based laptop. The overall setup is reflected on the Fig. 9 and 10.

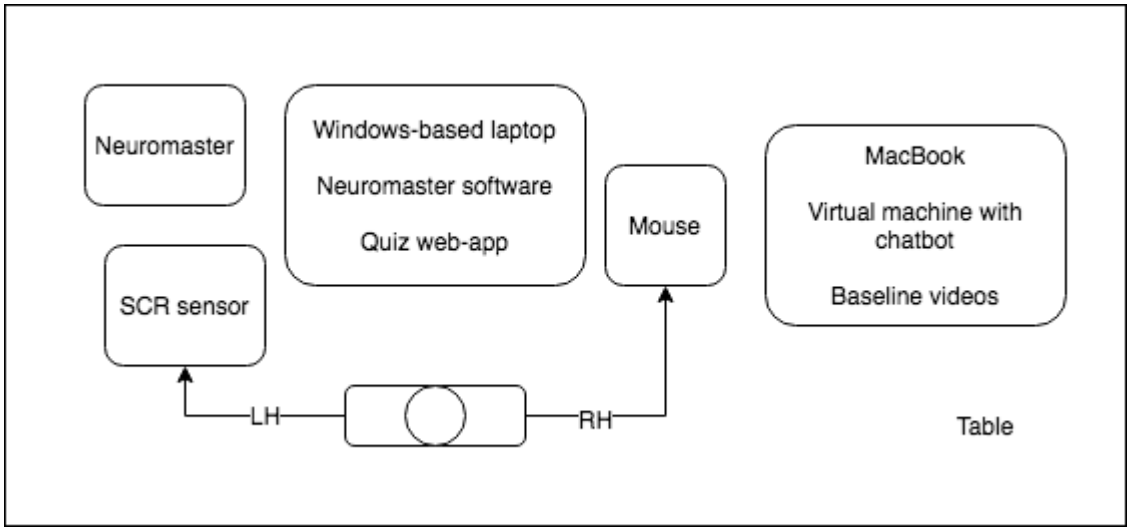

Fig. 9. Experimental Setup

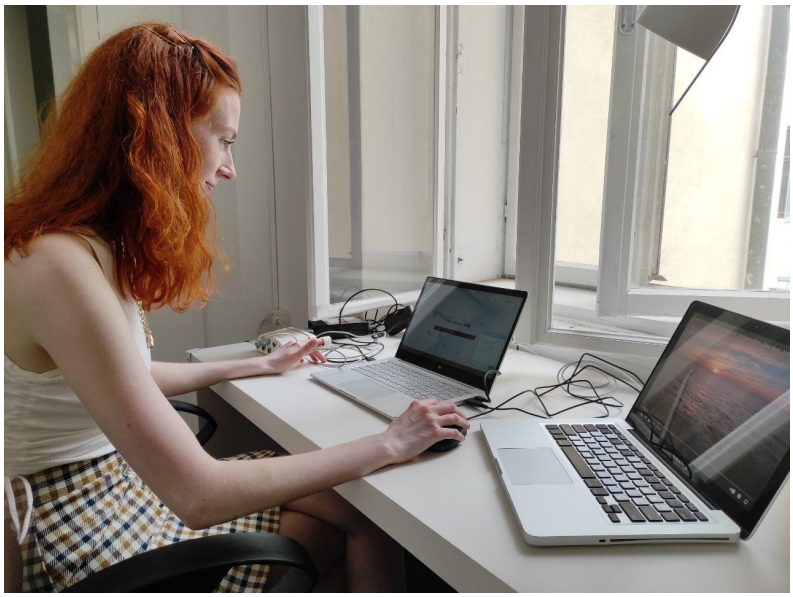

Fig. 10. Experimental Setup in the Lab

The data from the SCR sensor was sampled with a frequency of 20ms (50Hz). Raw data is recorded to text file containing timestamps in milliseconds, skin conductance signal level in microsiemens, and label of the channel (SCR/heart rate). The session data (text input from the participant and chatbot's responses) was saved in the external database.

The data analysis of the samples obtained and testing the research hypothesis are presented in the next Section 4. "Results".

4. Results

4.1. Data Pre-processing

The data pre-processing was conducted in Jupyter Notebook, using *pandas* and *numpy* packages. The overall dataset was filtered for the SCR channel. Finally, SC values were scaled to microsiemens, while timestamps were adjusted from milliseconds to minutes.

4.1.1. Motion Artifacts Removal and Data Smoothing

SC data obtained from the wearable sensor is sensitive to even slight disturbances, such as motion artifacts (see Fig.11), that have to be removed/filtered to avoid distortion of the results (Braithwaite&Watson et al., 2015).

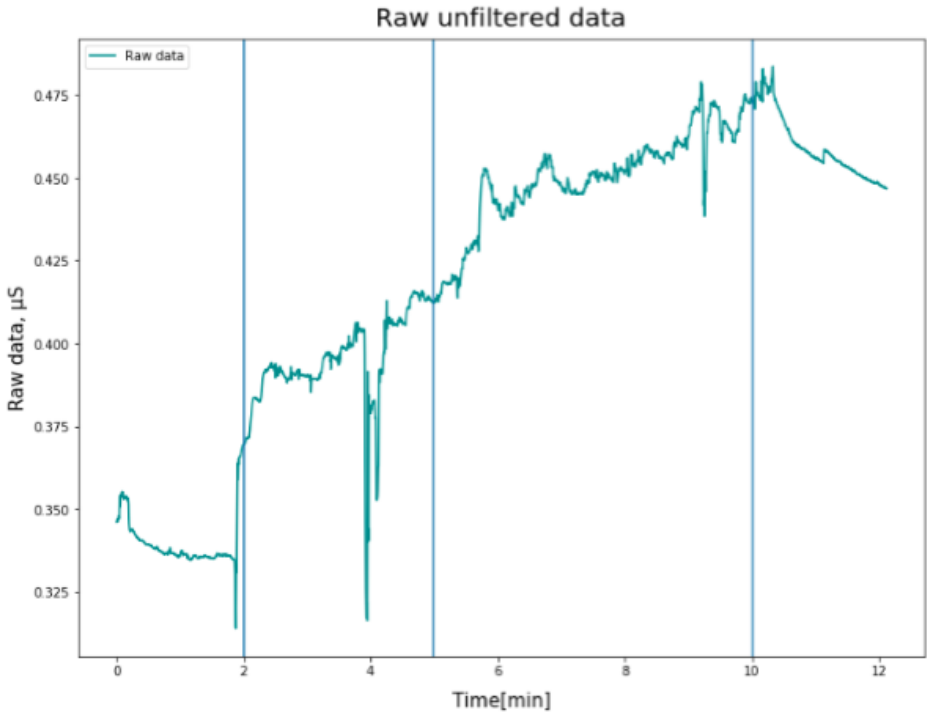

Fig. 11. Raw, Unfiltered Data of the Sample

For cleaning EDA data, different approaches are typically used. For example, the *Hampel filter* (Pearson et al., 2016; Wheeler, 2019) is an algorithm that checks if the data point lies more than the specified threshold from the dataset median and replaces it with the sliding window median value. This method shows good results for cleaning moving artifacts from SC data (Deubner, 2019). The most recent approach is applying *ML-based algorithms* that predict outliers and smoothes the curve (for example, see Taylor et al., 2015, who used Discrete Haar Wavelet Transform and Support Vector Machines

algorithm for artifact detection). However, the existing ML-based approaches are developed considering a specific data model of popular EDA-processing software, which wasn't available for the present study, so more common and straightforward filtering methods were used instead.

For this research, a *median filter* implemented in *SciPy* package was chosen as an artifact removal and smoothing solution. The median filter uses a very similar to the Hampel filter approach: the sliding window calculates the averaged value of the neighboring points to correct the median inside the window. The window size can be defined empirically, though it is recommended to use a range starting from one second and find the appropriate span (see Biopac, 2021). For the research dataset, a window size equal to three seconds was mainly used. During the investigation of two algorithms, it was revealed that the median filter provides comparable results (in some cases, it eliminates major artifacts even better) – the Fig.12 and Fig.13 represent the filtered data. At the same time, the Hampel filter is less computationally efficient (while the median filter processes the data instantly, with the Hampel filter, it took up to 14 seconds for each file).

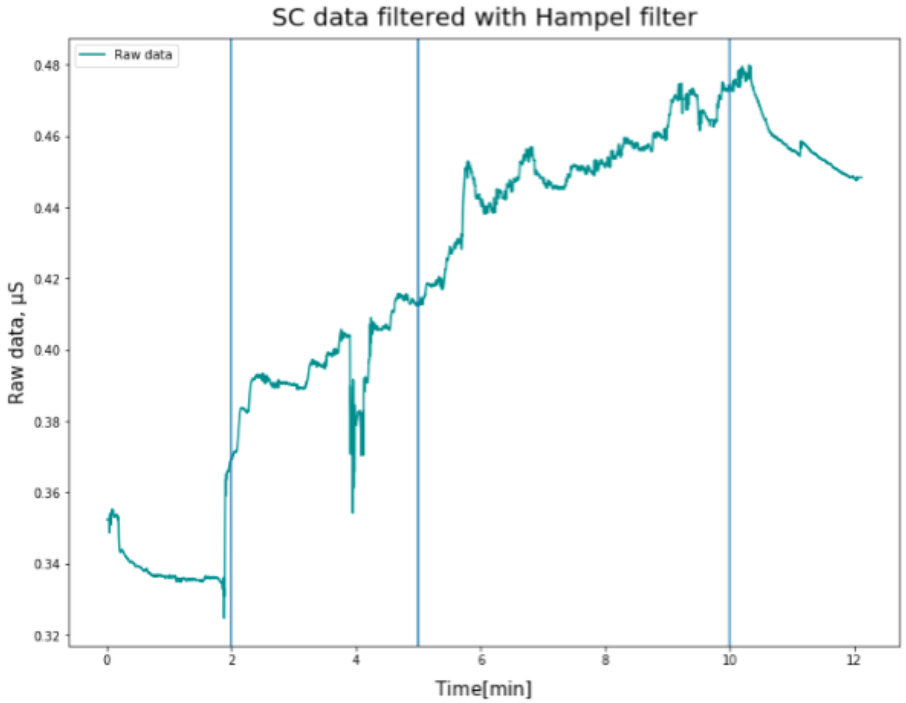

Fig. 12. The Sample, Cleaned up with Hampel Filter

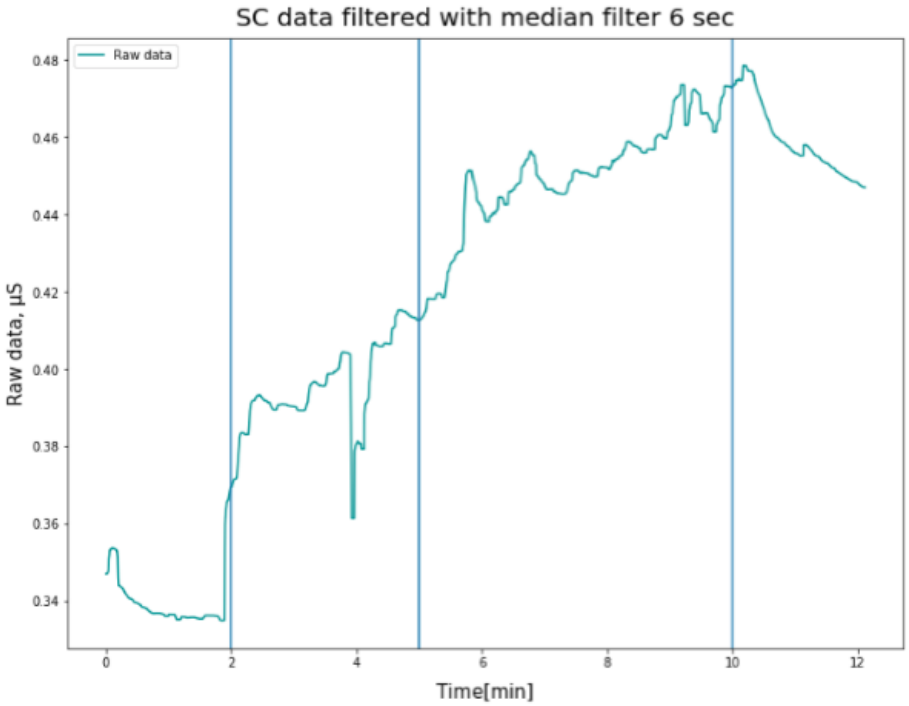

Fig. 13. Median Filter Applied to the Sample

During pre-processing, each sample was also visually examined. If the sample had problems with data acquisition (signal loss) or rough artifacts not removed even by a median filter with larger window size, it was dropped out from the further analysis. From 33 participants, 29 initial samples were obtained; 9 were dropped after the pre-processing phase.

4.2. Obtaining Sample and Dataset Statistics

After data was cleaned from the artifacts and noise, it could be split into tonic (SCL) and phasic (SCR) components. SCR data is associated with sympathetic arousal; it is possible to define the delta between the baseline and task-related levels by analyzing it. To process the data, an open-source *Neurokit2* library was used (Makowski et al., 2021). Dealing with the dependencies required for Neurokit2 installation, it is important to make sure that SciPy version is 1.2.0 or higher.

Neurokit2 package is based on the *cvxEDA* algorithm (Greco et al., 2015; Bijak et al., 2019). cvxEDA takes z-score values of the cleaned-up SC signal as an input. This approach fulfills another critical requirement to the EDA between-subject analysis – *normalization* of the data. As absolute values of the sympathetic nervous system activity can differ from one participant to another, it is crucial to normalize the whole dataset on one scale (Braithwaite & Watson et al., 2015).

As an output, Neurokit2 provides a data frame containing the SCR peaks data, including the mean amplitude of the peaks. Based on the timestamps assigned to the specific sample, dividing the whole data frame into two baselines (baseline 1, baseline 2) and the stress (task) phase is possible. As a result, the number of peaks per minute and the mean amplitude of each phase was obtained. The target variable – the delta between the stress phase and the averaged baselines' values – was also calculated.

To optimize and speed up the data analysis, all the pre-processing steps were compiled into one script. The script took the file path to raw data and timestamps of the experiment phases as an input. A number of peaks per minute, the mean amplitude of each phase, and the delta between the stress phase and the averaged baselines' values were provided as an output.

All the outputs for 20 samples kept for further analysis were combined into one final dataset. The data frame contained codes of the samples, assigned test group, all the demographic data, and covariates from the pre-test questionnaire. In addition, SCR statistics (peak number and mean amplitude), the delta between the stress phase and the baseline, and Likert scale scores from the post-test questionnaire were also included.

The *target variable* of the experiment was the delta of SCR peaks mean amplitude between the stress phase and the baseline. The data analysis showed the following results, illustrated on the Fig.14:

- For the control group (No stress, Default model), the delta = 1.2087
- Test group Stress, Default model delta = 1.5414
- Test group Stress, Improved model = 0.9921
- Test group No stress, Improved model = 0.9899

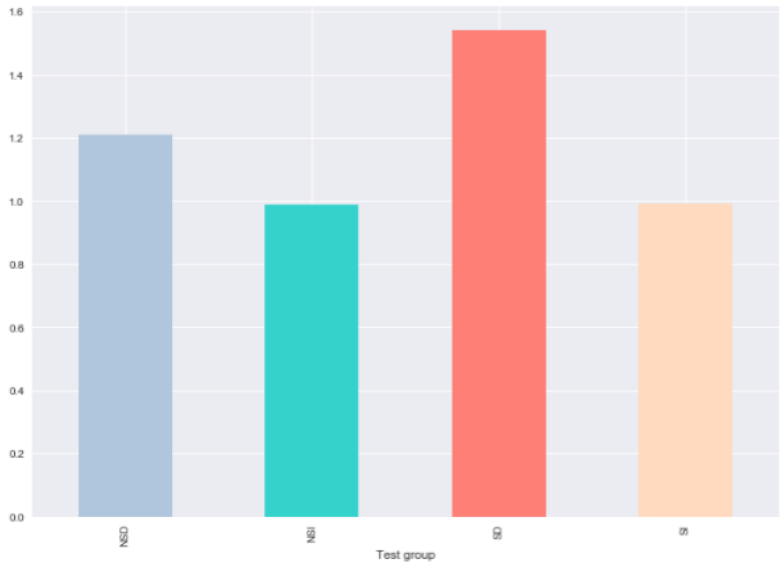

Fig.14. SCR Peaks Mean Amplitude Delta per Experimental Group

4.3. Full-factorial Analysis

For the-full factorial analysis, Python and R software was used, including such packages as *pandas, numpy* and *SciPy* (for Python).

First, the *main effects* of the independent variables on the target variable were checked. The resulting values were:

- Stress factor: 0.0242
- Model factor: -0.5307

It means that the stress factor has a weak positive correlation with the target variable, as expected (see section 5.2 Limitation for the detailed discussion of why the applied stress factor may not be a valid substitute for real-life stress). At the same time, the improved or default model factor is significantly negatively associated with the level of the target variable - the improved model reduces the values up to 53%.

Two-way interaction of the full-factorial analysis value = -0.1653. It shows that the additional stress factor weakens the effect of the improved model, though not significantly. As well it means that it still affects the target variable via indirect correlation.

The obtained coefficients for the mathematical model are:

$\hat{y} = 1.002 + 0.012\ x1 - 0.265\ x2$

Overall results explain the minimal difference between the levels of target variable in “Stress-Improved” and “No Stress-Improved” groups, considering that the difference between “Stress-Default” and “No Stress-Default” groups is much higher. The improved model on this sample size compensates for the weak stress factor, while with the default model, this compensatory effect does not occur, and the SCR peaks levels are much higher.

4.4. Hypotheses Testing and Statistical Significance

The null hypothesis H0 and alternative hypothesis are defined as:

H0: x1 = x2

H1: x1 ≠ x2,

Where x1 and x2 is the mean value of the Delta in the experimental groups with two improved interaction design factor levels -1 and 1 correspondingly. In other words, the null hypothesis assumes that *there is no statistically significant effect of the improved interaction design model.*

To check the statistical significance of the defined improved interaction effect, the ANOVA analysis in RStudio was conducted using the *car* package[1]. Before running the ANOVA analysis, it is essential to make sure that:

[1] https://cran.r-project.org/web/packages/car/index.html, last accessed: 28.11.2021

a) observations are random and independently taken from the population,

b) variance of the independent variables is equally distributed, and

c) the dependent variable is normally distributed (or close to normal).

The participants were assigned randomly to the experimental groups and took the experiment only once and independently, so the assumption a) is met. For the point b), the residual plot can be used to check the variety across the independent variables (see Fig.15)

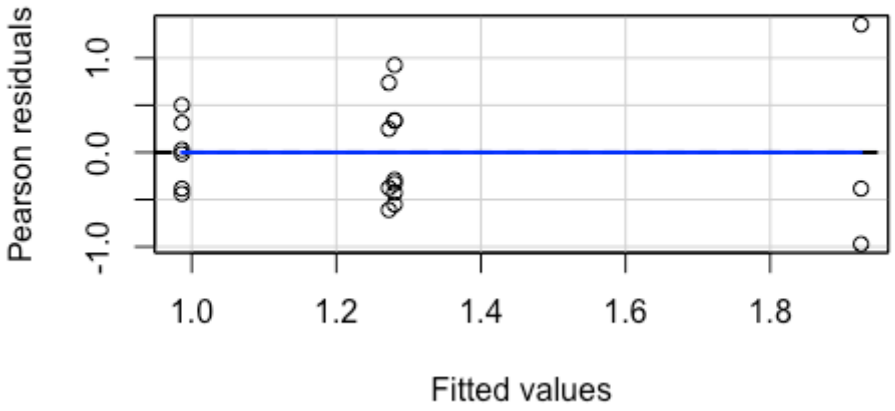

Fig.15. Residual Plot

There is a noticeable pattern in the residual plot, so the assumption is met as well.

For point c), a q-q plot can be used to check if the dependent variable is normally distributed (Fig. 16).

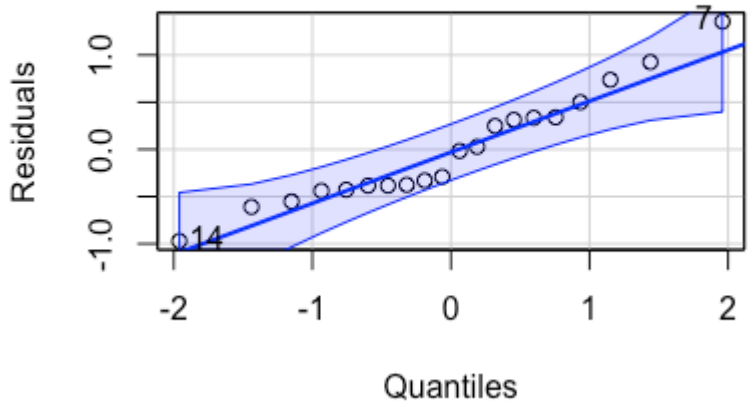

Fig. 16. Q-q pPot of the Dependent Variable Distribution.

All the values fit inside the normal distribution, so this assumption is also met.

After the dropping of the distorted samples, the distribution of the participants between the experimental groups became not equal. As non-balanced full-factorial design can affect the results, it was essential to take it into account. Therefore, ANOVA Type II squares sum was used, as recommended in the literature (Langsrud, 2003).

The ANOVA analysis was performed in the *car* R package. The F-value for the Model effect = 1.8699, p-value for the Model effect = 0.1904. To decide if the null hypothesis can be rejected, it is also important to consider a proper α value. Though the most common threshold for α value is usually set to 0.05 or 0.01, with the small sample size, it can affect the test power and make it weak. In literature (for example, see Kim, 2015), it is advised to choose a balanced α value that will increase the test power based on the sample size. With the fixed α value = 0.05, for the sample size n=20, the test power will be around 0.29. With the decreasing α value (for this sample size, the recommended value = 0.31), the test power will already be 0.69.

Based on these considerations, the obtained p-value of the Model effect by the ANOVA test = 0.1904 is lower than F-value and the chosen α value = 0.31. Therefore the null hypothesis can be rejected. **Thus, the improved interaction design model has a statistically significant effect on the participants' stress levels.**

For the Stress factor, the F-value = 0.0142, p-value = 0.9066, so the factor has no statistically significant effect on the target variable.

4.5. Analysis of the Post-test Questionnaire Likert Scale Data

The post-test questionnaires provide Likert-scale data for the self-reported level of the perceived Usefulness, Easiness of use, Learnability, and overall Satisfaction from the interaction. The additional variable introduced is the self-reported level of stress experienced during the experiment. The goal of the analysis was to define if there would be a significant difference between the reported values depending on the experimental group. In other words, the analysis should determine if the improved interaction design and the additional stress factor affected the evaluation of the participants.

For the analysis, RStudio and *likert* package was used[2]. The package supports building an ordinal regression model – one of the most common approaches to the Likert-scale data analysis (Schweinberger, 2020; Golicher, 2017).

The regression uses the factor levels and the target variable values to build the model. As an output, the factor coefficient, t-value, and p-value are provided.

[2] https://cran.r-project.org/web/packages/likert/, last accessed 28.11.2021

The Stress factor positively correlates with the self-reported level of stress (the coefficient value = 1.0647). The t-value = 1.2662, p-value = 0.2054, which is statistically significant for the thresholds chosen. For the Model factor, the effect on the self-reported levels of stress is not significant: t-value = 0.4492, p-value = 0.6533.

Another positive and statistically significant correlation is between the Stress factor and the perceived level of Usefulness: for the experimental groups with the added stress, the t-value = 1.5790, p-value = 0.1143. But again, no statistically significant correlation between the improved Model factor and the evaluated Usefulness can be reported. The same relation can be observed for the Satisfaction variable and Stress factor with t-value = 1.2675, p-value = 0.2050.

The interpretation of the results is presented in Section 5. “Discussion”.

5. Discussion

5.1. Implications of the Results

The present research examined to which extent the improved Human-Chatbot interaction design can decrease the level of sympathetic arousal in the users. The experimental data showed that the SCR peaks amplitude levels were 22% lower for the groups without the additional stress factor and 55% lower for those with the timer. This effect was also tested as statistically significant.

The analysis revealed that the stress factor itself did not significantly affect the target variable (stress levels). It can be explained by a mild severity of the factor (a timer in a quiz game), which cannot be compared with a real-life stress situation (banking account fraud or a credit card loss, problems with a traveling reservation, and similar). Still, it negatively affected the conversational model in the two-way interaction analysis. These findings confirm the importance of the human-centered design approach "designing for an error" – even mild stress could decrease the facilitating effect of the improved interaction, though not eliminate it.

The results of SCR data analysis prove that **implementation of the HAI guidelines indeed improves the interaction between humans and chatbots, increasing the quality of user experience.** As mentioned in Section 1.5, just a few studies evaluate human-chatbot interaction with objective data from physiological measures, so these findings are important empirical evidence in favor of HAI methods.

Another interesting point is the results of questionnaires Likert scale analysis. First, the statistically significant correlation was found only between the Stress factor and self-evaluated stress level and between the Stress factor and perceived Usefulness and Satisfaction dimensions. Secondly, there was no significant correlation between the conversational model and subjective evaluation of the chatbot/stress level. Finally, there was no clear trend in the questionnaire data: in all four dimensions (perceived Usefulness, Easiness, Learnability and Satisfaction) median score of evaluation provided does not change consistently depending on the experimental group. The explanation could be that participants subjectively considered themselves more stressed in the presence of the timer. Because of this, they evaluated the chatbot as more useful and satisfactory in case of successfully passing a quiz.

The absence of a clear trend in the questionnaire data could be explained by two assumptions:

1) the final sample size (20 participants), was insufficient for the Likert scale analysis, and the trend could be more visible with a larger sample,

2) As discussed in Section 1.5, there is a general problem with the reliability of subjectively reported data and self-assessments used in UX research, more pronounced in smaller samples.

The discussion of self-report validity for measuring emotions has been held in literature for more than decade (Picard& Daily, 2005; Boehner et al., 2007). In this perspective, self-report incorporates such problems as forgetting, normalization, and willful misinterpretation of the affective experience by the participants. Riedl and Leger investigated this approach in “Foundations of NeuroIS” (Riedl&Leger, 2016) and pointed out that data provided in self-report questionnaires is limited with conscious perception and thoughts and can be affected by memory distortion. Therefore, basic unconscious processes such as pleasure and stress cannot be reflected accurately. The authors assumed that application of physiological methods, such as EEG, ECG, EDA, EMG, eye-tracking, and other measures, may positively affect the reliability of UX research data.

The results of the current research with validated EDA data and unclear questionnaire trend can be considered as an additional proof that it is not sufficient to use only surveys, observations, and self-assessment for usability research, especially when it is impossible to get a large enough sample, as these methods are highly subjective.

5.2. Novelty and Further Research

The research was fulfilled in a highly interdisciplinary framework, aiming to empirically test the Human-Chatbot Interaction Design's effect with objective physiological data. As presented above, there are just a few similar studies in both HAI and NeuroIS domains, so this master thesis can be considered as possessing a pronounced novelty.

In a recent paper by Følstad et al. (2021), authors make a comprehensive overview of the current research in the area of Human-Chatbot Interaction and propose the key directions for further interdisciplinary studies. Measuring and assessing of UX with chatbots and empirical evaluation of the design models and approaches are highlighted as one of the main challenges. Considering that EDA data provides a reliable evaluation of users’ emotional experience, the current research contributes to the most actual problems in this area.

Original experimental procedure, approach to implementation of conversational AI models and data analysis pipeline for examination of EDA data in full-factorial design are another contribution of this research.

The study creates a basis for various themes that can be examined in subsequent studies. First, it can be valuable to run A/B testing of similar conversational models in a real-life setup and compare obtained survey and interview results with the physiological

laboratory data. This could provide insights about the capability of the interaction framework applied to cope with stress in daily scenarios.

Secondly, it is possible to vary the sets of the design guidelines to test which have the main effect on the arousal levels. Considering the number of the overall guidelines proposed by the HAI community, which could be hard to implement, it would be beneficial to define the critical ones that can drastically improve UX and increase the chatbots acceptance rate.

Next, in the current study it was not a goal to include all existing covariates (age, IT-proficiency level, and so on) into the data analysis to examine how they correlate with the target variable. However, it could be possible to do so with a larger sample in further research. Such findings could help practitioners adjust how chatbots interact with the needs of a specific audience.

Finally, in the scope of this thesis, only the EDA sensor was used as one of the simplest ways to obtain physiological data; the subsequent research could include more objective data sources such as heart rate variability, electrocardiogram, electromyography, and others. Cross-analysis of the several channels could help better understand the physiological response toward the improved interaction design.

5.3. Limitations

As discussed above, the Stress factor showed a weak effect on the target variable. The arousal induced during the experiment cannot be compared with facing problems with bank accounts, technical issues, or traveling bookings, which was one of the study's limitations. This limitation, while grounded in the ethically adherent experimental design used, should still be noted. Nevertheless, as the research question was aimed not on absolute but rather on relative levels of SCR during the test phase, the study is enabled to address the central research question, and provide evidence on the effectiveness of improved interaction model.

The sample size was acceptable for the full-factorial analysis, as was discussed in Section 4. Considering the experimental groups with the smallest number of samples, at least three test runs for each combination were available. Still, the larger sample could 1) make the trend in questionnaire Likert data more pronounced and, 2) increase the power of the test for statistical significance.

5.4. Conclusion

The current research contributes to the Human-AI Interaction domain using an interdisciplinary toolkit of cognitive science, empirical psychology, computer science, and neuroscience. The research is based on a comprehensive theoretical basis of Post-

Cognitivist theories, HCI, HAI, human-centered design, affective design, and technology acceptance. The interaction framework derived from the examined HAI guidelines was implemented in two conversational AI models in a cloud infrastructure. The framework was experimentally tested with physiological measures to obtain objective signals from sympathetic nervous systems besides subjective questionnaires. The results show statistical significance of the effect that Interaction design guidelines bring and provide proof of how they can improve the user experience.

Applying these recommendations to a daily practice of chatbots development can help to increase their acceptance among the users and bring the technology closer to their needs. The research also draws a perspective of further studies in the field.

References

1. Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, *31*(2), 427-445.
2. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019, May). Guidelines for human-AI interaction. *In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-13).*
3. Auernhammer, J. (2020). Human-centered AI: The role of Human-centered Design Research in the development of AI.
4. Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019, October). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7, No. 1, pp. 2-11).
5. Bijak, M. (2019) "Skin impedance is reliable parameter for arousal monitoring (stress monitoring)", in *13th Vienna International Workshop on Functional Electrical Stimulation*, 2019.
6. Billings, C. E. (1991). *Human-centered aircraft automation: A concept and guidelines* (Vol. 103885). National Aeronautics and Space Administration, Ames Research Center.
7. Biopac, 2021. EDA data analysis & correction. https://www.biopac.com/eda-faq-data/ (https://www.biopac.com/eda-faq-data/)
8. Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.
9. Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies*, *65*(4), 275-291.
10. Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2015). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, *49*(1), 1017-1034.
11. Brettlecker, S. (2019) Evaluation and classification of skin conductance profiles. Bachelor Thesis, University of Applied Sciences Technikum Wien
12. Card, S. K., Moran, T. P., & Newell, A. (2018). *The psychology of human-computer interaction*. CRC Press.
13. Carmichael L., Poirier SM., Coursaris C., Léger PM., Sénécal S. (2021) Does Media Richness Influence the User Experience of Chatbots: A Pilot Study. In: Davis F.D., Riedl R., vom Brocke J., Léger PM., Randolph A.B., Müller-Putz G. (eds) Information Systems and Neuroscience. NeuroIS 2021. Lecture Notes in Information Systems and Organisation, vol 52. Springer, Cham.
14. Chaves, A. P., & Gerosa, M. A. (2021). How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, *37*(8), 729-758.

15. Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, *92*, 539-548.
16. Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., ... & Cowan, B. R. (2019, May). What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).
17. Cooley's, M. (1982). Architect or Bee? South End Press Boston
18. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.
19. Dean, A., Voss, D., & Draguljić, D. (1999). *Design and analysis of experiments* (Vol. 1). New York: Springer.
20. Deubner, O. 2019. Bilateral Cataract Surgery Patients: Evaluation of Emotional Arousal using Skin Conductance. Master's thesis, University of Applied Sciences FH Technikum, Vienna
21. Di Paolo, E., & Thompson, E. (2014). The enactive approach. In *The Routledge handbook of embodied cognition* (pp. 86-96). Routledge.
22. Fischer, G. (1995, October). Rethinking and reinventing artificial intelligence from the perspective of human-centered computational artifacts. In *Brazilian Symposium on Artificial Intelligence* (pp. 1-11). Springer, Berlin, Heidelberg.
23. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health*, *4*(2), e7785.
24. Ford, K. M., Hayes, P. J., Glymour, C., & Allen, J. (2015). Cognitive orthoses: toward human-centered AI. *AI Magazine*, *36*(4), 5-8.
25. Følstad, A., Araujo, T., Law, E. L. C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., ... & Luger, E. (2021). Future directions for chatbot research: an interdisciplinary research agenda. *Computing*, *103*(12), 2915-2942.
26. Gibson, J. J., & Carmichael, L. (1966). *The senses considered as perceptual systems* (Vol. 2, No. 1, pp. 44-73). Boston: Houghton Mifflin.
27. Gibson, J.J. (1979) The ecological approach to visual perception: classic edition. Psychology Press
28. Golicher, D. (2017) Analysing Likert scale satisfaction scores https://rpubs.com/dgolicher/Limert_scale_analysis
29. Gould, J. D., & Lewis, C. (1985). Designing for usability: key principles and what designers think. *Communications of the ACM*, *28*(3), 300-311.
30. Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., & Citi, L. (2015). cvxEDA: A convex optimization approach to electrodermal activity processing. IEEE Transactions on Biomedical Engineering, 63(4), 797-804.
31. Grudin, J. (2005). Three faces of human-computer interaction. *IEEE Annals of the History of Computing*, *27*(4), 46-62.

32. Grudin, J. (2009). AI and HCI: Two fields divided by a common focus. *Ai Magazine*, *30*(4), 48-48.
33. Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000, December). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241-250).
34. Holmes, T. H., & Rahe, R. H. (1967). The social readjustment rating scale. *Journal of psychosomatic research*.
35. Hughes, G., Sachdev, S., Curtis, M., Bolze, J. (Accenture) (2021) The future of customer conversation: More than words, more than AI.
36. Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive science*, *19*(3), 265-288.
37. Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.
38. Johnson, M., & Vera, A. (2019). No AI is an island: the case for teaming intelligence. *AI Magazine*, *40*(1), 16-28.
39. Kim, J. (2015). How to choose the level of significance: A pedagogical note.
40. Kocielnik, R., Amershi, S., & Bennett, P. N. (2019, May). Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).
41. Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I. and Wong, W-K (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC, pp. 3-10. DOI: 10.1109/VLHCC.2013.6645235
42. Kuutti, K., & Bannon, L. J. (1993, May). Searching for unity among diversity: Exploring the “interface” concept. In *Proceedings of the INTERACT'93 and CHI'93 conference on human factors in computing systems* (pp. 263-268).
43. Kuutti, K. (1996). Activity theory as a potential framework for human-computer interaction research. *Context and consciousness: Activity theory and human-computer interaction*, *1744*.
44. Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing*, *13*(2), 163-167.
45. Launchbury, J. (2017). *A DARPA Perspective on Artificial Intelligence*, http://www.darpa.mil/about-us/darpa-perspective-on-ai.
46. Lieberman, H. (2009). User interface goals, AI opportunities. *AI Magazine*, *30*(4), 16-16.
47. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.
48. Liu, B., & Sundar, S. S. (2018). Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, *21*(10), 625-636.
49. Lubbe, I., & Ngoma, N. (2021). Useful chatbot experience provides technological satisfaction: An emerging market perspective. *SA Journal of Information Management*, *23*(1), 8.

50. Lucas, G. M., Gratch, J., King, A., & Morency, L. P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94-100.
51. Luger, E., & Sellen, A. (2016, May). " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5286-5297).
52. Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., ... & Chen, S. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 1-8.
53. Maturana, H. R., & Varela, F. J. (1972). *Autopoiesis and cognition: The realization of the living* (Vol. 42). Springer Science & Business Media.
54. Menary, R. (2010). Introduction to the special issue on 4E cognition. *Phenomenology and the Cognitive Sciences*, *9*(4), 459-463.
55. Meng, J., & Dai, Y. N. (2021). Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not?. *Journal of Computer-Mediated Communication*.
56. Nardi, B. A. (Ed.). (1996). *Context and consciousness: Activity theory and human-computer interaction*. mit Press.
57. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), 81-103.
58. Norman, D. A. (1986). *User-centered system design: New perspectives on human-computer interaction*. CRC Press.
59. Norman, D. A. (1994). How might people interact with agents. *Communications of the ACM*, *37*(7), 68-71.
60. Norman, D. A. (2004). *Emotional design: Why we love (or hate) everyday things*. Basic Civitas Books.
61. Norman, D.A. (1989; 2013 - the revised version). *The Design of Everyday Things*. CurrencyDoubleday, New York.
62. Pearson, R. K., Neuvo, Y., Astola, J., & Gabbouj, M. (2016). Generalized hampel filters. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 1-18.
63. Picard, R. W., & Daily, S. B. (2005, April). Evaluating affective interactions: Alternatives to asking what users feel. In *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches* (Vol. 10, No. 1056808.1057115, pp. 2119-2122). New York, NY: ACM.
64. Riedl, R., & Léger, P. M. (2016). Fundamentals of NeuroIS. *Studies in neuroscience, psychology and behavioral economics*, 127.
65. Rowlands, M. J. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. MIT Press.
66. Russell, J. A. (1980). A circumplex model of emotion. *Journal of Personality and Social Psychology*, *39*, 1161-1178.
67. Schweinberger, Martin. 2020. Questionnaires and Surveys: Analyses with R. Brisbane: The University of Queensland. url: https://slcladal.github.io/surveys.html (Version 2020.12.11).

68. Segars, A. H., & Grover, V. (1993). Re-examining perceived ease of use and usefulness: A confirmatory factor analysis. *MIS quarterly*, 517-525.
69. Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, *12*(3), 109-124.
70. Skjuve, M., Haugstveit, I. M., Følstad, A., & Brandtzaeg, P. B. (2019). HELP! IS MY CHATBOT FALLING INTO THE UNCANNY VALLEY? AN EMPIRICAL STUDY OF USER EXPERIENCE IN HUMAN-CHATBOT INTERACTION. *Human Technology*, *15*(1).
71. Sukis, J. (2019). AI Design & Practices Guidelines (A Review). https://medium.com/design-ibm/ ai-design-guidelines-e06f7e92d864.
72. Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015, August). Automatic identification of artifacts in electrodermal activity data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1934-1937). IEEE.
73. Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.
74. Tsai, W. H. S., Liu, Y., & Chuan, C. H. (2021). How chatbots' social presence communication enhances consumer engagement: the mediating role of parasocial interaction and dialogue. *Journal of Research in Interactive Marketing*.
75. Vlasov, V., Mosig, J. E., & Nichol, A. (2019). Dialogue transformers. *arXiv preprint arXiv:1910.00486*.
76. Weber, P., & Ludwig, T. (2020, September). (Non-) Interacting with conversational agents: perceptions and motivations of using chatbots and voice assistants. In *Proceedings of the Conference on Mensch und Computer* (pp. 321-331).
77. Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625-636.
78. Winograd, T. (2006). Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artificial intelligence*, *170*(18), 1256-1258.
79. Wheeler, W. (2019) Clean up your time series data with a Hampel filter. Medium, 27.05.2019
80. Xu, W., & Dainoff, M. J. (2021). Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI.
81. Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020, April). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1-13).
82. Yao, L., Liu, Y., Li, W., Zhou, L., Ge, Y., Chai, J., & Sun, X. (2014, June). Using physiological measures to evaluate user experience of mobile applications. In *International conference on engineering psychology and cognitive ergonomics* (pp. 301-310). Springer, Cham.
83. Yen, C., & Chiang, M. C. (2021). Trust me, if you can: a study on the factors that influence consumers' purchase intention triggered by chatbots based on brain image evidence and self-reported assessments. *Behaviour & Information Technology*, *40*(11), 1177-1194.

84. Zaki, T., & Islam, M. N. (2021). Neurological and physiological measures to evaluate the usability and user-experience (UX) of information systems: A systematic literature review. *Computer Science Review*, *40*, 100375.
85. Zamora, J. (2017). I'm sorry, Dave, I'm afraid I can't do that: Chatbot perception and expectations. In *Proceedings of the 5th international conference on human agent interaction* (pp. 253-260).

Appendix 1: Abstrakt

Diese Masterarbeit untersucht das Thema der Mensch-Chatbot-Interaktion aus der interdisziplinären Perspektive der Kognitionswissenschaft, der empirischen Psychologie, der Informatik und der Neurowissenschaften. Im Rahmen der Masterarbeit wurde ein Interaktionsframework auf Basis der überprüften HAI-Guidelines entwickelt. Das Framework wurde empirisch in einem vollfaktoriellen interindividuellen Experiment mit 33 Teilnehmern getestet. Während des Experiments wurden physiologische Messungen der elektrodermalen Aktivität verwendet, um objektive Signale von sympathischen Nervensystemen zu erhalten. Zusätzlich wurde ein subjektiver Fragebogen durchgeführt, um die Selbstberichterstattung über Stress und wahrgenommene Nützlichkeit, Leichtigkeit, Lernfähigkeit und Zufriedenheit aus der Interaktion zu messen. Die Ergebnisse zeigen die statistische Signifikanz des Effekts, den Interaktionsdesign auf die Benutzererfahrung hat. Die Hauptbeiträge der Arbeit bilden i) Erdung der Mensch-Chatbot-Interaktionsansätze in einer umfassenden theoretischen Grundlage; ii) Entwicklung eines originellen Versuchplans; iii) empirische Prüfung der HAI- Guidelines durch Erhebung objektiver physiologischer Daten; iv) Entwicklung einer neuartigen Datenanalyse-Pipeline für elektrodermale Daten in einem vollfaktoriellen Versuchplan; v) empirische Beiträge zur Diskussion über subjektive und objektive Methoden der Usability-Forschung; vi) Vorschläge für die zukünftige Forschung im Bereich der empirischen Untersuchung der Mensch-Chatbot-Interaktion.