CENTRIC 2020

The Thirteenth International Conference on Advances in Human oriented and Personalized Mechanisms, Technologies, and Services

ISBN: 978-1-61208-829-7

October 18 -22, 2020

CENTRIC 2020 Editors

Stephan Böhm, RheinMain University of Applied Sciences, Germany

Oana Dini, IARIA, USA/EU

CENTRIC 2020

Forward

The Thirteenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2020), held on October 18 - 22, 2020, addressed topics on human-oriented and personalized mechanisms, technologies, and services, commonly known as I-centric.

There is a cohort of technologies that favored the so called “user-centric” services and applications. While some of them reached some maturity, others are to prove their economics (WiMax, IPTV, RFID, etc). The human-oriented and personalized technologies and services rely on a key set of features, some to be deployed, others getting more mature (personal profiles, preferences, identity, proximity, personal devices, etc.). Following, advanced applications covering human related activities benefit from personalized and human-oriented networks and services, especially preventive and personalized medicine, body networks and devices, or anticipative systems.

The conference provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The conference sought contributions presenting novel result and future research in all aspects of user-centric mechanisms, technologies, and services.

Similar to the previous editions, this event continued to be very competitive in its selection process and very well perceived by the international community. As such, it attracted excellent contributions and active participation from all over the world. We were very pleased to receive a large amount of top quality contributions.

We take here the opportunity to warmly thank all the members of the CENTRIC 2020 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to the CENTRIC 2020. We truly believe that thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the CENTRIC 2020 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success.

We hope the CENTRIC 2020 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in personalization research.

CENTRIC 2020 Steering Committee

Stephan Böhm, RheinMain University of Applied Sciences - Wiesbaden, Germany

CENTRIC 2020 Publicity Chair

Lorena Parra, Universitat Politecnica de Valencia, Spain

CENTRIC 2020

Committee

CENTRIC 2020 Steering Committee
Stephan Böhm, RheinMain University of Applied Sciences - Wiesbaden, Germany

CENTRIC 2020 Publicity Chair

Lorena Parra, Universitat Politecnica de Valencia, Spain

CENTRIC 2020 Technical Program Committee

Stefania Bandini, RCAST - Research Center for Advanced Science & Technology | The University of Tokyo, Japan
Lasse Berntzen, University of South-Eastern Norway, Norway
Stephan Böhm, RheinMain University of Applied Sciences - Wiesbaden, Germany
Daniel B.-W. Chen, Monash University, Australia
Sabine Coquillart, INRIA, France
Ângela Cristina Marques de Oliveira, Instituto Politécnico de Castelo Branco, Portugal
Carlos Cunha, Polytechnic Institute of Viseu, Portugal
Pradipta De, Georgia Southern University, USA
Luciane Fadel, Federal University of Santa Catarina, Brazil
Rainer Falk, Siemens AG Corporate Technology, Germany
Filipe Fidalgo, Polytechnic Institute of Castelo Branco, Portugal
Alicia García-Holgado, GRIAL Research Group - University of Salamanca, Spain
Faisal Ghaffar, IBM Ireland / University College Dublin, Ireland
Till Halbach, Norwegian Computing Center, Norway
Qiang He, Swinburne University of Technology, Australia
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Takeshi Ikenaga, Kyushu Institute of Technology, Japan
Imène Jraidi, Advanced Technologies for Learning in Authentic Settings (ATLAS) Lab | McGill University, Montreal, Canada
Christos Kalloniatis, University of the Aegean, Greece
Yasushi Kambayashi, NIT - Nippon Institute of Technology, Japan
Mazaher Kianpour, Norwegian University of Science and Technology (NTNU), Norway
Boris Kovalerchuk, Central Washington University, USA
Cun Li, Eindhoven University of Technology, Netherlands
Célia Martinie, Université Paul Sabatier Toulouse III, France
José Martins, LIAAD - INESC TEC / Polytechnic of Leiria, Portugal
Thomas Marx, TH Bingen, University of Applied Sciences, Germany
Erik Massarczyk, RheinMain University of Applied Sciences Wiesbaden Rüsselsheim, Germany
Pedro Merino, ITIS Software | University of Malaga, Spain
Toshiro Minami, Kyushu Institute of Information Sciences, Japan
Eduardo Miranda, University of Plymouth, UK

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Bias – A Lurking Danger that Can Convert Algorithmic Systems into Discriminatory Entities

Towards a Framework for Bias Identification and Mitigation

Thea Gasser, Eduard Klein
Business Department
Bern University of Applied Sciences (BUAS)
Bern, Switzerland
email: thea.gasser@live.com; eduard.klein@bfh.ch

Lasse Seppänen
Business Information Technology Department
Häme University of Applied Sciences (HAMK)
Hämeenlinna, Finland
email: lasse.seppanen@hamk.fi

***Abstract*—Bias in algorithmic systems is a major cause of unfair and discriminatory decisions in the use of such systems. Cognitive bias is very likely to be reflected in algorithmic systems as humankind aims to map Human Intelligence (HI) to Artificial Intelligence (AI). An extensive literature review on the identification and mitigation of bias leads to precise measures for project teams building AI-systems. Aspects like AI-responsibility, AI-fairness and AI-safety are addressed by developing a framework that can be used as a guideline for project teams. It proposes measures in the form of checklists to identify and mitigate bias in algorithmic systems considering all steps during system design, implementation and application.**

Keywords – Bias; Algorithm; Artificial intelligence; AI-safety; Algorithmic system.

I. INTRODUCTION

Artificial intelligence is present in almost every area of our society, be it in medicine, finance, social media, education, human resource management and many more. This trend will take up a deeper part of people's lives, since according to the Accenture Trend Report [1], about 85% of the executives surveyed plan to invest widely in AI-related technologies over the next three years. Moreover, AI will play a central role in how customers perceive a company and define to a large extent how interactions with their employees and customers take place. AI will become a core competency and will reflect a large part of a company's character. In five years, more than 50% of the customers will no longer choose a service based on the brand but will focus on how much AI is offered for that service [1].

Recently, however, there has been growing concern about unfair decisions made with the help of algorithmic systems that have led to discrimination against social groups or individuals [2] [3]. As an example, Google's image search had been accused of bias indicating fewer women than the reality when searching for the term "CEO". Additionally, Google's advertising system displayed high-income jobs much less to women than to men [4]. The COMPAS algorithm was accused of predicting that "black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk" [5]. Microsoft's Tay robot held racist and inflammatory conversations with Twitter users which contained many political statements. It learned from the users' inputs and reflected it in its answers [6]. These and many more known examples show that methods to measure algorithms, recognize and mitigate bias and provide fair AI-software, especially in a highly data oriented machine learning context, are demanded [3] [7].

This article contributes to AI-safety by highlighting that bias in AI is very likely, illustrating possible sources of bias and proposing a framework which supports the identification and mitigation of bias during the design, implementation and application phases of AI-systems.

The following research questions from Gasser [8] are addressed to tackle the above-mentioned aspects: (1) What is expected from AI-systems in relation to how humans make decisions? (2) How is bias present in algorithmic systems that affect human behavior and decisions? (3) How can bias in algorithmic systems be identified? (4) What measures can be taken to mitigate bias in algorithmic systems? Questions (1) and (2) are discussed in Sections III and IV based on literature research, and the proposed framework in Section V gives advice for answering questions (3) and (4) in the context of machine learning based AI projects.

The rest of this paper is organized as follows. Section II describes the research design. In Section III, different types of bias are discussed, followed by related research in Section IV. Section V addresses the bias mitigation framework in finer detail. The conclusions in Section VI close the article.

II. RESEARCH DESIGN

Extensive and systematic literature research has been conducted and the results have been analyzed according to [9]. Systematic analysis was applied by researching specific AI and bias related topics and content, thereby identifying central sources. Based on backward search strategy, further literature was identified. In total, over 100 journal articles, collected works, reference works, books and websites were researched.

As starting points, plain web search and database searches in the scientific portals SAGE journals, ScienceDirect, Springer Link, Google Scholar and the

 ISBN: 978-1-61208-829-7

JSTOR Journal storage have been carried out. A set of search terms has been employed such as "expectations towards AI", "human intelligence", "algorithmic bias", "bias in software development", “mitigating algorithmic bias”, thereby filtering and selecting to 75 relevant sources.

Based on the findings of the literature research, sources of bias and methods for identifying and mitigating bias in algorithmic systems were identified and structured and are systematically presented in Sections III and IV. The findings led to a framework for use in project settings which is described in Section V, thereby identifying and mitigating bias through the use of a metamodel and a set of checklists.

III. FROM HI TO AI

With AI, terms like imitation, simulation or mimicking are repeatedly applied which implies copying something, respectively, someone as, e.g., acting, learning and reasoning like humans [10]. Therefore, if today’s AI-behavior such as Apple’s Siri is considered, it could be claimed that the voice assistant is not intelligent. Looking into details, Apple’s voice assistant is based on evaluated data and facts permitting to offer an appropriate answer [11]. An independently thinking and reasoning machine is not yet present since, amongst other things, an input is still needed. Even though AI acquires intelligence and learns through an autonomous process it lacks sentience and self-awareness and is still only a simulation of HI and nothing more [10].

Despite the expectations and efforts to map HI to AI, to date, there is no system that can be classified as "strong AI", since this would include machines that act completely autonomously and have their own intelligence and self-awareness like humans. However, "weak AI" systems working in a narrowly defined area are used successfully already [12]. Even in the case of self-learning machines, there is initial program code, a model and learning rules so that machine learning can be effective [13]. Because human traits like self-awareness or empathy are missing in today's AI-systems, there is still a gap between AI and HI. This, in turn, implies that partly intelligent systems are shaped by the influence of humankind and with it by cognitive bias which is naturally present in humans and subsequently reflected through individuals and societies in algorithmic systems [14]. Research questions (1) and (2) relate to the decision-making aspects with AI systems.

A. Lack of Transparency in AI Systems

Algorithms are penetrating more and more into people’s lives and will likely overtake even stronger parts of their daily routine so that they will depend heavily on how secure and efficient these algorithms are [15]. Considering that algorithms are becoming more and more complex, and systems may become opaque so that it becomes partly unclear even to the creators of such systems themselves how exactly the interactions in the system(s) take place [16], measures need to be taken in order to minimize undesirable ethical consequences that might arise though the use of such systems. Therefore, the focus must be on potential bias that might arise in the system design, implementation and application phases.

B. Bias and Fairness

Since the term bias is defined as “the action of supporting or opposing a particular person or a thing in an unfair way, because of allowing personal opinions to influence your judgement” (according to the Cambridge Dictionary) the topic of fairness plays a central role. A system might be viewed as fair in some circumstances and in other situations it might be considered unfair. In addition, the presence of bias in an AI system cannot be regarded as evidence of the classification of a system as unfair, which means that neutral or even desirable biases may be present in AI systems without producing undesirable results [17]. Therefore, classifying an AI-system as fair or unfair is subjective and may depend on the viewer, e.g., based on the application context's cultural setting.

Based on these factors, it is important to identify bias and consider whether there is a need for action for reducing it or whether bias should even be used specifically to prevent other in a different part of the system that would have more undesired consequences [17].

The question of whether recognized bias needs to be reduced at all should always be assessed in the individual system context since mitigating bias can be a major effort. On the one hand, several associations demonstrate differences in how and which values are put in the foreground and which seem less important. On the other hand, the situation can reach a level of complexity that no matter what perspective is adopted, some bias will always be identified from a certain point of view. In the end, technology cannot fully answer questions about social and individual values. It is therefore up to humans to make sure that the particular situation is always evaluated in a comprehensive context, meaning taking into account the whole ecosystem around the machine [17].

C. Sources of Bias

Different authors identified various sources of bias in AI-systems. Barfield & Bagallo [18] consider what we call *direct bias* whose sources are related to the core of AI systems: (1) *Input bias* where the source data is biased due to absence of specific information, nonrepresentativeness or reflecting historical biases; (2) *Training bias* which arises when the baseline data is categorized, or the output is assessed; (3) *Programming bias* which emerges in the design phase or when an algorithm modifies itself through a self-learning process.

In [19], sources are identified of what we call *indirect bias*, which are not located in the core of AI systems but in the ecosystem around it: (1) *pre-existing bias* which often emerges through social institutions, practices and attitudes even before a system is designed; (2) *Technical bias*, emerging from technical constraints, e.g., by favoring data (combinations) due to the order or size of screens and visual results presentation; (3) *Emergent bias* arising when using a system outside its intended context of operation.

 ISBN: 978-1-61208-829-7

IV. RELATED RESEARCH

Recently, human aspects of AI have attracted a lot of attention. Not only private companies, research institutions and nonprofit organizations, but also public sector organizations and governments have issued policies and guidelines on human aspects of AI. Many recent publications cite or build on the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems called "Ethically Aligned Design" (EAD), where methodologies to guide ethical research are presented with the aim of promoting a public debate on how these intelligent and autonomous technologies can be aligned with moral values and ethical principles that prioritize human well-being [20].

The non-profit research organization AlgorithmWatch is developing an "AI Ethics Guideline Global Inventory" [21] to address the question of how automated decision-making systems should be regulated. At the time of writing, more than 80 movements are listed, ranging from a few private companies (e.g. Google, Microsoft, IBM) to organizations (e.g. IEEE, ACM, Bitkom) and government-related organizations (e.g. China, European Commission, Canada, Singapore).

Several metastudies presented the state of the art in human aspects of AI at the time of writing. In [22], an extended list is supplemented by a geographical distribution displayed on a world map. A global convergence of ethical aspects is revealed, emerging around five ethical principles: transparency, fairness, nonmaleficence, responsibility and privacy. It highlights the importance of integrating efforts to develop guidelines and its implementation strategies.

In [23], a comprehensive literature review is presented based on key publications and proceedings complementing existing surveys of psychological, social and legal discussions on the subject with recent advances in technical solutions for AI governance. Based on the literature research, a taxonomy is proposed that divides the field into areas for each of which the most important techniques for the successful use of ethical AI systems are discussed.

All publications mentioned present principles and guidelines for the consideration of ethical aspects in AI systems, thereby addressing research questions (1) and (2). However, they are general and generic and could be used as high-level recommendations only, which are not sufficiently specific for AI projects. The framework presented in Section V further develops these ideas and therefore points the way to the next step in incorporating ethical aspects in a project-oriented environment. Based on a metamodel and a set of checklists, it allows to identify and mitigate bias in AI systems in a project-oriented setting, thereby addressing research questions (3) and (4). The integration of ethical aspects into all project phases during the conception, development and use of a system guarantees a high level of awareness among all project stakeholders.

V. THE BIAS MITIGATION FRAMEWORK

Awareness of the topic is the first step towards addressing bias in algorithmic systems. According to [24], 92% of AI-leaders make sure their technologists receive ethics training and 74% of the leaders assess AI-outcomes every week. However, it is not enough to just dispose ethics codes that prevent harm. Therefore, establishing usage and technical guidelines and an appropriate mindset among the stakeholders are suggested.

To address bias in algorithmic systems appropriately, an overarching and comprehensive governance must be in place in companies. Using the proposed framework, the project members should be committed to the framework, considering it as a binding standard.

In literature, many possibilities are described to identify bias such as (1) monitoring and auditing an AI system's creation process [25], (2) Applying rapid prototyping, formative evaluation and field testing [19], (3) manipulating test data purposefully in order to determine whether the results are an indication of existing bias in the system [26], (4) using the Socratic method promoting critical thinking and challenging assumptions through answering questions, where scrutiny and reformulation play a central role in the identification and reduction of bias [27].

Tool-based approaches such as IBM's "AI Fairness 360" offer metrics to check for unwanted bias in datasets and machine learning models [28]. Google's "What-If" tool enables visualization of inference results, e.g., for exploring the effect of a certain algorithmic feature and also testing algorithmic fairness constraints [29].

Despite the many approaches that have been suggested in literature and the tools that are available focusing on specific topics in ethical aspects, justification for the proposed framework is in incorporating aspects for all members involved in the process of creating an algorithmic system and all relevant aspects researched.

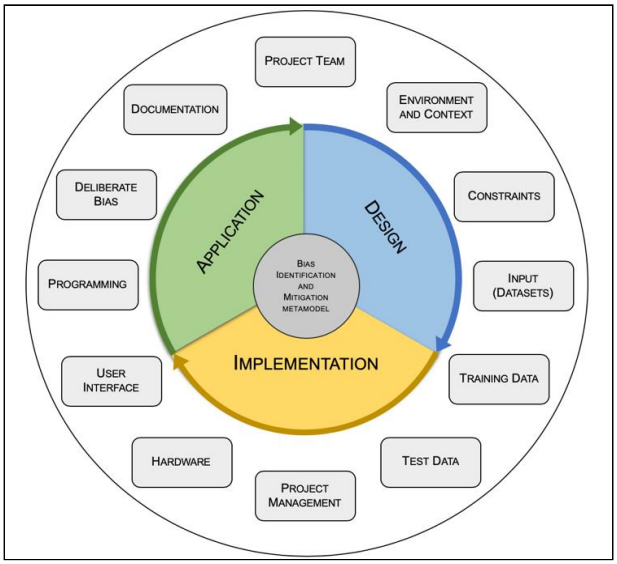

Figure 1. Metamodel for the Bias Mitigation Framework.

The framework consists of a metamodel (see Fig. 1) which is completed by checklists for areas covering the whole software life cycle around design, implementation and application. The areas (e.g. Project Team, Environment,

Content) are illustrated as rectangles in Fig. 1. The elements of each checklist consist of statements and questions that need to be addressed by the project team. The checklists are derived from the findings of the research described in Sections II, III and IV and relate to the research questions (3) and (4).

As an example, the area *Project Team* is subsequently described and detailed in Fig. 2. Knowledge, views and attitudes of individual team members cannot be deleted or hidden, as these are usually unconscious factors, due to everyone's different background and experiences.

Element	Description/Comments	Y/N
Project Team		
All project members have had ethical training	- Members have a confirmation that they have completed courses or workshops or similar - The minimum requirements to consider this element as fulfilled must be defined in the company	
All project members are aware of the topic of bias that exists in the human decision-making process	- Members took part in courses or workshops or similar - The minimum requirements to consider this element as fulfilled must be defined on a project or company level	
All project members know about the fact that human bias can be reflected in an algorithmic system	- Members took part in courses or workshops or similar - The minimum requirements to consider this element as fulfilled must be defined on a project or company level	
All project members consider the same attributes and factors as most relevant in the system context.	- A workshop is held where members share their views. Discrepancies are pointed out and a common understanding is developed. The workshops aim to share views, ideas and openly in order to reveal conflicts and misunderstandings - Due to cultural and background dissimilarities members might (unconsciously) weight attributes differently	
The project team represents stakeholders of all possible end user groups	- Stakeholder analysis comprehensively identifies end user groups with a focus on identifying users who might be disadvantaged through the system outcomes - Stakeholder analysis should be carried out with a change of perspective, where the worst scenario, i.e. if the system behaves discriminatory, identifies the groups that would be disadvantaged. (see area Project Management)	
The project team is a cross-functional team including diversity in ethnicity, gender, culture, education, age and socioeconomic status	- The inputs of all the diverse individuals have to be taken into consideration.	
The project team has representatives from the public and private sector	- Exclusions need to be avoided	
Independent consultants are included for comparison with competing products	- Pre-existing bias in the context of the company's culture, attitude and values can be revealed. - Independent consultants are needed because they are not biased by the companies' views	

Figure 2. Checklist for the metamodel area *Project Team*.

The resulting bias is likely to be transferred into the algorithmic system. Therefore, measures must be taken to ensure the neutrality of the system as far as appropriate. It is necessary that there is an exchange among project members where everyone shares their views and concerns openly, fully and transparently before creating the system. Misunderstandings, ideas of conflict, too much euphoria and unconscious assumptions or invisible aspects might get revealed this way. The checklist in Fig. 2 proposes the following concrete measures for addressing the above-mentioned issues: All project members (1) have had ethical training, (2) are aware of the topic of bias that exists in the human decision-making process, (3) know about the fact that bias can be reflected in an algorithmic system, and (4) consider the same attributes and factors as most relevant in the system context.

Ideally, the project team (1) represents stakeholders of all possible end user groups, (2) is a cross-functional team including diversity in ethnicity, gender, culture, education, age and socioeconomic status, (3) has representatives from the public as well as the private sector. Moreover, independent consultants are included for comparison with competing products.

A. *Checklists*

The metamodel in Fig. 1 illustrates 12 areas of interest, where the project team area was detailed already in Section V. This subsection gives an overview of the 11 remaining areas. For each area the checklist is presented, and the corresponding literature references are explained.

In [19], the different cultural values and attitudes of individuals are emphasised that could collide as they incorporate those into the project work. These aspects are covered by the areas *Environment and Context* and *Constraints* (Fig. 3) in the Framework. In [13] [17] [26] [30], the influence of direct bias is discussed (see "sources of bias" in Section III), leading to the basis for the areas in Fig. 4.

Element	Description/Comments	Y/N
Environment and Context		
All end user groups are included in the testing phase	- The behaviour of end users can only be reliably recorded if they test directly on the live system. Hidden behaviour can thus be detected	
End user groups have been evaluated	- End user groups' behaviour is monitored and evaluated from different perspectives (surveys, interviews, recording behaviour, letting them explain what they do and think while testing)	
Consequences and intentions have been considered	- For what and with what intentions was the system created for? - What is the worst thing that can happen in this algorithm if it starts interacting with others?	
Context is faithful to the original source	- Does the current context represent the one, for which the system was originally created?	

Element	Description/Comments	Y/N
Constraints		
Business aspect reviewed	- Under what circumstances will the system be developed?	
Scope reviewed	- The requirements for the scope of the data set and the diversity are to be determined in the respective project	
Technical aspect reviewed	- Do technical constraints affect the way the system is designed?	
Legal aspect reviewed	- Do regulatory/law constraints affect the way the system is designed?	

Figure 3. Checklist for areas *Environment and Context* and *Constraints*.

Element	Description/Comments	Y/N
Input (Datasets)		
The data set is fully understood	- The meaning of each attribute is understood and its purpose in the system context is clear	
Data is transparent	- Data must be reliable, accurate and kept up to date	
It is ensured that the data set represents the correct scope (enough data representing a population resp. a target group)	- Enough data and diversity are available - The requirements for the scope of the data set and the diversity are to be determined in the respective project.	
The source of the data is known and verified	- Unknown source of the data might lead to that the data is used in a context it was originally not intended to	
The quality of the data is ensured	- Data with low quality will cause even worse outputs since AI-systems might reinforce errors in data sets	
It is clarified which attributes can legally be used	- Use of illegal attribute leads to a system becoming biased even though the attribute itself is not cause for bias	

Element	Description/Comments	Y/N
Training Data		
The training data set is still as representative as the original data set	- Adjusting source data to training data can bear exclusion which needs to be prevented	
Added or omitted attributes are carefully chosen and justified	- One attribute can influence different areas in a system. Interconnectedness needs to be considered	

Element	Description/Comments	Y/N
Test Data		
Test data is independent	- The system uses test data it has never seen before	
Test data is defined	- Test scenarios are defined which are designed to detect bias which could be caused by a certain attribute	
Test data is reviewed	- Tests include omission and addition of attributes to test how system output changes	

Figure 4. Checklists for the areas concerning *direct bias*, derived from "sources of bias" in Section III.

 ISBN: 978-1-61208-829-7

It is suggested that the complete algorithmic system lifecycle is accompanied and controlled through all phases with a project management approach. The classical element "risk analysis" must be expanded with a focus on risk factors that could favour bias and the effects recognised bias could have. Isele [27] suggests that critical questions should be asked, critical thinking adopted, assumptions challenged, and the results of the system evaluated. Aspects on the Project Management area are gathered in Fig. 5.

Element	Description/Comments	Y/N
Project Management		
Project management process includes methods that focus on bias issues	- Stakeholder analysis is adjusted for disadvantaged group identification in worst case	
Risks concerning bias are assessed and known to each team member	- Risk analysis is adjusted for additional focus on bias and worst-case scenarios provoking to bias	
Critical thinking is promoted and demanded at every stage of the system creation process	- How would changes to a data point affect the model's prediction? - Does it perform differently for various groups? For example, historically marginalised people? - How diverse is the dataset I am testing my model on? - Is the system context the one the system was intended to? - Can the outcome/result/system recommendation be justified? - How diverse is the dataset I am testing my model on? - Does it perform differently for various groups–for example, historically marginalized people? - How would changes to a data point affect my model's prediction?	
Perspectives are changed continuously to challenge assumptions	- Different points of views ensure identification of hidden assumptions	
Monitoring measures are defined, communicated and applied	- End user groups' behaviour is monitored and evaluated from different perspectives (surveys, interviews, recording behaviour, letting them explain what they do and think while testing)	
Auditing measures are defined, communicated and applied	-	
Workshops / meetings are set frequently which address upcoming doubts of team members	- Critical thinking is continuously fostered in workshops and outside	
Scenario thinking is fostered	-	
Freedom of expression is guaranteed and desired	- Every input of any team member can reveal hidden bias	

Figure 5. Checklist for the area *Project Management.*

Hardware limitations, such as screen size or performance bottlenecks, could influence system output [19]. The design of visual representations of objects could also be a source of bias, requiring a careful design of the graphical user interface [31]. Checklists for hardware limitations and Graphical User Interface (GUI) design are detailed in Fig. 6.

Element	Description/Comments	Y/N
Hardware		
Hardware limitations	- Do hardware limitations exist?	
Influence on creation process	- Do these limitations influence the system creation process?	
Influence on production environment	- Do these limitations influence the system's functionality in the production environment?	

Element	Description/Comments	Y/N
User Interface		
Visual aspects are determined appropriately	- The font-style, font-size, font-colour and placement of text are justified and reflect the intention of the system's functionality - Colour, size and placement of forms and graphics are justified and reflect the intention of the system's functionality	
Visual result	- Does visual result representation (alphabetically or random) make any difference (user always choses the results displayed first?)	
Navigation	- Does a change in navigation representation lead the user to favour different results?	
Graphical User Interface	- Is graphical UI limiting/favouring data over other data?	
Language Aspects	- How does the chosen language influence the user's perception and interpretation in different contexts and circumstances? - Is a translation of data/information necessary? - Do the information and results become distorted through the application of translation? - How is the translation interpreted by the end users?	
Alternative GUI	- The system features are changed, and end users are monitored once more in order to see how their behaviour changes - Several features may need to be changed various times in order to reveal hidden assumptions of end users	

Figure 6. Checklist for areas *Hardware* and *UI.*

Sources of bias in programming and documentation and discussion on deliberate bias [17] are given in Fig. 7.

Element	Description/Comments	Y/N
Programming		
Code reviews take place	- Measures aim to understand adapted or reused code fully	
Independent code audits are conducted	- Independent audits foster considering the code from a different point of view and reveal unconscious assumptions	
Possible user behaviour is analysed beforehand to keep a learning system from adopting discriminatory behaviour	- Thinking outside the box is fostered especially considering word and language usage in the system context - The system can handle discriminatory user behaviour	

Element	Description/Comments	Y/N
Deliberate Bias		
Bias is identified and categorized	- Are the identified biases considered as good, neutral or bad ones? - Is there any bias which was implemented on purpose in order to mitigate other?	
It is ensured that all the identified biases are monitored during the whole system creation process	- Bias needs to be tracked and changes identified as well as recorded throughout every stage of the project	

Element	Description/Comments	Y/N
Documentation		
Availability of relevant information	- Traceability, justification and business continuity is ensured	
Comprehensible documentation	- The language may only contain such a high degree of complexity and technical language that every project member understands it - Prevention of misunderstandings is ensured	
Documentation has been reviewed and approved	- The documentation needs to be reviewed by several project members and stakeholders	

Figure 7. Checklist for areas *Programming, Documentation and Deliberate Bias.*

The presence of deliberate bias might be surprising at first, however, is applied in some cases to prevent bias from arising in another, more important area of a system. As an example, a statistically biased estimator in an algorithm might exhibit significant reduced variance on small sample sizes, thereby greatly increasing reliability and robustness in future use [17].

B. Application of the Framework

Based on the outcome of the above-mentioned literature research, the approach presented is intended to be an initial framework that can be adapted to specific needs within a given project context. It comes in shape of a guideline complemented with checklists, e.g., for the members of a project team.

The adjustments could be made based on an adapted understanding of system neutrality which may be specific for the respective application or application domain. If the proposed framework is used in a mandatory manor within a project, it is very likely that the developed application reflects the neutrality defined by the project team or company.

Verifying that the framework has been applied and the requirements have been met will help to determine the extent to which the system is neutral and the need for appropriate action.

VI. CONCLUSION

Since currently there are only weak AI-systems which lack self-awareness and depend on human advice in shape of created models and selected training data, human bias is naturally and unintentionally reflected in crafted algorithmic systems. A framework has been proposed which helps to

identify and mitigate bias in algorithmic systems, covering aspects of the complete life cycle of such software systems.

Since the framework in its current state is a synthesis of desk research, future research should implement the approach in realistic software project situations such that its added value could be observed, evaluated, validated and subsequently adapted based on the project experiences. During validation, each metamodel area would require separately assessing the priority of the questions and requirements in the checklists and ensuring useful answers.

In addition, it would be useful to investigate to what extent automation of the use of the framework could mitigate subjective opinions and views of the stakeholders involved. As an example, the following scenario could be realized: Information about the adapted framework (metamodel areas and checklist elements), which is considered standard for ensuring system neutrality up to a certain point in the respective project, could be supported by a software system. During the project, the checklists are continuously filled with data by the project team, thereby enabling analysis of the process, comparison of different implementations of the framework and revealing indications where the recommendations were complied with and where it was not.

On the one hand, a specific project team would always be aware when creating an algorithmic system, which of the specified areas would not be adhered to and could exhibit potential bias. On the other hand, this mechanism could also be used for end users. They could more easily assess how reliable the results of the AI-system are and which areas need more attention regarding bias. The impact of decisions taken through the AI-system's suggestions can be better analyzed by knowing which areas do not comply with the elaborated standard.

However, to reach this point, there are several aspects that need to be considered. Elements from the checklists would have to be detailed at micro level to define, for example, what a stakeholder is or how it can be verified that the test user belongs to a respective gender. Instead of a yes/no check mark in the check lists, there could be more detailed measures, e.g., indication of the level to which a team member has received ethical training. In addition, mechanisms could be integrated to take account of the truthfulness of the answers in the checklists.

REFERENCES

[1] Accenture, "AI is the new UI – Experience Above All," Accenture Technology Vision, 2017. https://www.accenture.com/_acnmedia/accenture/next-gen-4/tech-vision-2017/pdf/accenture-tv17-full.pdf. [retrieved: 18-Aug-2020].

[2] A. Koene, "Algorithmic Bias: Addressing Growing Concerns," IEEE Technol. Soc. Mag., vol. 36, no. 2, pp. 31–32, Jun. 2017.

[3] M. Veale and R. Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data," Big Data Soc., vol. 4, no. 2, pp. 1-17, Dec. 2017.

[4] D. Cossins, "Discriminating algorithms: 5 times AI showed prejudice," New Scientist, 2018. https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/. [retrieved: 18-Aug-2020].

[5] S. Feuerriegel, M. Dolata, and G. Schwabe, "Fair AI," Bus. Inf. Syst. Eng., pp. 379-384, 2020.

[6] E. Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter," The Guardian, 24-Mar-2016. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter. [retrieved: 18-Aug-2020].

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems, pp. 1-9, 2016.

[8] T. Gasser, "Bias – A lurking danger that can convert algorithmic systems into discriminatory entities," HAMK University, Finland, 2019.

[9] M. Kornmeier, Wissenschaftlich schreiben leicht gemacht (academic writing made easy), 8th ed. Haupt Verlag, 2018.

[10] S. Holder, "What is AI, really? And what does it mean to my business?," 2018. https://www.sas.com/en_ca/insights/articles/analytics/local/what-is-artificial-intelligence-business.html. [retrieved: 18-Aug-2020].

[11] A. Goel, "How Does Siri Work? The Science Behind Siri," Magoosh Data Science Blog, 2018. https://magoosh.com/data-science/siri-work-science-behind-siri/. [retrieved: 18-Aug-2020].

[12] J. R. Searle, "Minds, brains, and programs," Behav. Brain Sci., vol. 3, no. 3, pp. 417–457, 1980.

[13] E. Alpaydın, Introduction to Machine Learning, 2nd ed. 2012.

[14] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," Big Data Soc., vol. 3, no. 1, pp. 1-12, 2016.

[15] A. Smith, "Franken-algorithms: the deadly consequences of unpredictable code," The Guardian, 2018. https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger. [retrieved: 18-Aug-2020].

[16] C. Smith, B. McGuire, T. Huang, and G. Yang, "The History of Artificial Intelligence," Univ. of Washington, 2006.

[17] D. Danks and A. J. London, "Algorithmic Bias in Autonomous Systems," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017, pp. 4691–4697.

[18] W. Barfield and U. Pagallo, "Research Handbook on the Law of Artificial Intelligence. Edited by Woodrow Barfield and Ugo Pagallo. Cheltenham, UK," Int. J. Leg. Inf., vol. 47, no. 02, pp. 122–123, Sep. 2019.

[19] B. Friedman and H. Nissenbaum, "Bias in computer systems," ACM Trans. Inf. Syst., vol. 14, no. 3, pp. 330–347, Jul. 1996.

[20] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design," 2019.

[21] AlgorithmWatch, "AI Ethics Guidelines Global Inventory," 2019. https://inventory.algorithmwatch.org/about. [retrieved: 18-Aug-2020].

[22] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nat. Mach. Intell., vol. 1, no. 9, pp. 389–399, Sep. 2019.

[23] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, "Building Ethics into Artificial Intelligence," in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 5527–5533.

[24] SAS, "Organizations Are Gearing Up for More Ethical and Responsible Use of Artificial Intelligence, Finds Study," 2018. https://www.sas.com/en_id/news/press-releases/2018/september/artificial-intelligence-survey-ax-san-diego.html. [retrieved: 18-Aug-2020].

 ISBN: 978-1-61208-829-7

[25] A. Raymond, “The Dilemma of Private Justice Systems: Big Data Sources, the Cloud and Predictive Analytics,” Northwest. J. Int. Law Bus., vol. 22, pp. 1-44, 2014.

[26] I. Žliobaitė, “Measuring discrimination in algorithmic decision making,” Data Min. Knowl. Discov., vol. 31, no. 4, pp. 1060–1089, 2017.

[27] E. Isele, “The Human Factor Is Essential to Eliminating Bias in Artificial Intelligence,” Chatham House, 2018. https://www.chathamhouse.org/expert/comment/human-factor-essential-eliminating-bias-artificial-intelligence. [retrieved: 18-Aug-2020].

[28] R. K. E. Bellamy et al., “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,” IBM J. Res. Dev., vol. 63, no. 4–5, pp. 1–15, 2019.

[29] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson, “The What-If Tool: Interactive Probing of Machine Learning Models,” IEEE Trans. Vis. Comput. Graph., vol. 26, no. 1, pp. 56-65, 2020.

[30] S. Barocas and A. D. Selbst, “Big Data’s Disparate Impact,” 104 Calif. Law Rev. pp. 671-732, 2016.

[31] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. Upper Saddle River: Prentice Hall, 2010.

 ISBN: 978-1-61208-829-7

Human-Machine Interaction: EEG Electrode and Feature Selection Exploiting Evolutionary Algorithms in Motor Imagery Tasks

Aurora Saibene

Multi Media Signal Processing Laboratory
Department of Informatics, Systems and Communications
University of Milano-Bicocca
Milan, Italy
Email: a.saibene2@campus.unimib.it

Francesca Gasparini

Multi Media Signal Processing Laboratory
Department of Informatics, Systems and Communications
University of Milano-Bicocca
Milan, Italy
Email: francesca.gasparini@unimib.it

***Abstract*—Brain Computer Interfaces (BCIs) based on the recording of electroencephalographic signals have revolutionized the human-machine interaction. Being in presence of heterogeneous electrophysiological data, that come with a low number of instances and a great number of features, it is necessary to find a solution that can achieve good performances with respect to all the subjects, having as input a restricted feature subset. Firstly, we propose a population-based approach that allows to mitigate the data heterogeneity. Secondly, not wanting to make assumptions on the feature types, we propose the application of genetic algorithm, particle swarm optimization and simulated annealing as evolutionary feature selection techniques. We present the results of our proposal on a motor movement/imagery experiment. From these results, we verified that each feature type contributes differently depending on the task and on the sensor it was computed on, thus giving a broader view of the different type of analyses that can be performed to allow a better interaction between a user-centric system like a BCI based on motor imagery and its human user.**

Keywords–Brain Computer Interface; Electroencephalography; Evolutionary Feature Selection.

I. INTRODUCTION

The combination of Brain Computer Interfaces (BCIs) and Electroencephalography (EEG) has allowed the development of a plethora of applications directly based on the translation of human brain responses into machine understandable instructions. These responses are usually due to natural neurological processes or elicited by external stimuli and interactions and can be easily recorded in a non-invasive way by placing electrodes (sensors) on a volunteer's scalp.

Each electrode returns an electroencephalographic signal of a peculiar brain area, deputed to specific brain activities and functions [1]. Thus, the EEG signal representing the responses has a spatial connotation in addition to its temporal resolution. It is also characterized by different frequency bands [2], each of which is associated to a peculiar set of brain states, summarized in Table I.

Moreover, the EEG signal is easily affected by noise and is heterogeneous, having variations inter-volunteers, but also intra-volunteer. In fact, depending on the volunteer's physiological and psychological conditions, on external factors like the environmental temperature or on the type of recording that is performed (e.g., clinical analysis, experimental setting and so on) the EEG signal could drastically change.

The described characteristics must be taken into account when the recorded EEG signals are used as inputs to a BCI, which provides a user-centric system able to recognize the brain activity patterns coming from the EEG signals and consequently allows a human-machine interaction [3], following two steps [4]: (1) offline training for system calibration and (2) online translation of brain responses.

One of the most widely studied BCI applications is based on Motor Imagery (MI), i.e., the imagination of movement, mainly for rehabilitation purposes: from moving a prosthetic arm to controlling a wheelchair. Focusing on left/right-handed MI tasks, it has been proved that the brain activation coming from the imagination of the left/right hand movement mimics the one necessary to perform a real movement of the left/right hand. This activation involves a specific brain area, called motor cortex, which in a modified 10/10 electrode configuration is covered by the sensors highlighted (light-blue) in Figure 1 [5]. However, most of the literature works reduces the analyses on the electrodes enclosed by the red line (Figure 1), being this choice bounded to experimental design or made a priori. The sensors placed on the right hemisphere records the motor left-handed movement/imagination, while on the left, the motor right-handed movement/imagination.

Notice that mainly two frequency bands are involved during a motor imagery task [6]: the power spectrum in the α band (also called μ band when observed in the motor cortex) decreases, while in the β band increases.

Having this field knowledge in mind, the aim of the various researches conducted on this topic is to discriminate the left/right-handed MI tasks in order to have reliable and efficient brain computer interface systems. Different classification techniques have been applied to the electroencephalographic signals [4] and recently deep learning models [7] [8] have been developed to move from hand-crafted features, i.e., custom computed signal characteristics, to learned features.

Most of the state-of-the-art works dealing with standard classification techniques (e.g., support vector machines, neural networks and so on), have mainly concentrated their efforts in refining the classification performances. They also compute hand-crafted features limited to the previously described field knowledge. Moreover, the computed features usually take into account only a subset of the various combinations that could be made, especially using the spatial information, power spectra on the frequency bands of interest or some statistical measures.

 ISBN: 978-1-61208-829-7

Having a limited amount of instances per task, the proposed approaches could be considered a good compromise to maintain a low computational complexity while obtaining good accuracy values without incurring in the overfitting and curse-of-dimensionality issues.

However, some information that could have an impact on the interpretation of the recorded brain responses could be ignored and the various contribution of different electrodes and feature types being lost.

A solution to this loss of information could be the computation of different feature types and a subsequent feature selection, unbiased by a priori knowledge.

Therefore, in our work we compute heterogeneous features on the signal obtained by each available electrode and a set of Evolutionary Feature Selection (EFS) methods, based on Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Simulated Annealing (SA). We compare the performances obtained by applying different Support Vector Machines (SVMs) models on the resulting subset of features against the classical a priori selection and Principal Component Analysis (PCA) computation.

Our aim is to provide a benchmark that will highlight the contribution given by the spatial (i.e., electrode) and feature type information and that will be exploited for the future development of more complex and possibly efficient classification models for a better interaction between a MI-based BCI and its human user.

To this hand our contributions can be summarized as follows:

1) analysis of the motor left/right hand movement/imagination tasks with a population-based approach instead of limiting the analysis on a single-subject;
2) consider a combination of heterogeneous features in the time, frequency and time-frequency domains, through statistical measures, not wanting to be limited by the field knowledge;
3) apply different feature selection techniques in order to verify the efficacy of methods that do not make assumptions on the features, passing from a priori knowledge selection and extracting dimensions with PCA, to the original application of EFS algorithms;
4) original analyses on the agreement between the EFS resulting feature subset, considering both the electrode and feature type contributions.

TABLE I. FREQUENCY BAND BRAIN ACTIVITY ASSOCIATION [2].

Name	*Range (Hz)*	*Association*
δ	0.5 - 4	present during sleep
θ	4 - 7	present during sleep
α	8 - 13	present in a relaxed state while awake
β	13 - 30	present in a focused/alert state
γ	30 - $<$ 100	present during insight/problem solving phenomena

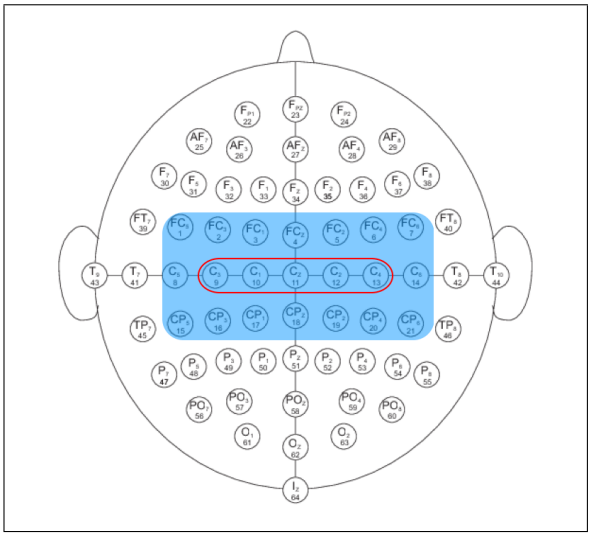

Figure 1. Electrodes covering the motor cortex brain area.

The rest of the contribution is organized as follows. Section II provides the background information on the state-of-the-art and the exploited characteristics of EFS algorithms. Also, the used dataset is described. In Section III, we provide a detailed explanation of the proposed approach, while in Section IV, we discuss the obtained results from different tests. Section V concludes the paper highlighting our contributions, some notes on the obtained results and the future work.

II. BACKGROUND

The core of our proposal is the feature selection performed by evolutionary computation algorithms. This process consists in the search of a relevant subset of features with a multi-objective approach: find the minimum number of features needed to obtain the maximum classification accuracy.

Generally, an evolutionary feature selection method starts with the initialization of its parameters and a random selection of the features. Afterwards, the feature subset search and the evaluation of its quality are performed until a stopping criterion is met. The evaluation step, represented by the fitness function, could follow different approaches [9], i.e., the wrapper and filter approaches. In particular, we use a wrapper approach [10], which includes a classification algorithm for the evaluation of the feature subset. Therefore, we discarded the filter approach, which ignores the classification performance, being it not suitable for our purpose. As a final step, the obtained results are validated.

The EFS algorithms are appealing due to the fact that they do not require field knowledge and can return different solutions in a single execution, without making any assumption on the features [11].

We exploit these advantages in an offline configuration knowing that these techniques have as major drawbacks the high computational complexity and cost. Moreover, there could be a stability issue due to the random nature of the processes [11], which we mitigate by investigating the agreement between the applied EFS algorithms on the selected features.

To our knowledge, we are the first to apply three different EFS techniques, i.e., genetic algorithm, particle swarm optimization and simulated annealing, and analyze their agreement on the selected features considering both the electrode and feature type contributions.

In fact, some works have proposed the usage of evolutionary computation for feature selection in the context of BCIs, focusing their attention on one aspect and technique at a time. Also, the concept of feature subset is mostly considered as an electrode set reduction.

On this topic, *Atyabi et al.* [12] propose the separate usage

of GA, PSO and random search to find the best electrode locations that guarantee the maximum classification accuracy using a sigmoid extreme learning machine. *Amarasinghe et al.* [13] also apply the GA technique on their BCI data for robot control to obtain the best classification accuracy from a support vector machine by selecting the minimum number of electrodes. Instead, *Gonzalez et al.* [14] apply the NSGA-II optimization technique and change the fitness function using a combination of Kappa index and error distribution. They also propose a feature ranking procedure to address the stability issue, but they make an a priori choice on the sensors set.

Even though these researches have relevance in the field of EFS, the authors test their techniques with a subject-based approach and with a small amount of instances per class. Here, we propose a population-based approach to assess the possibility of having a generalized procedure, that we can apply on a greater number of instances and subsequently use to make some assumptions when analyzing the data coming from a new single volunteer.

To this hand, we test our methods on the PhysioNet *EEG Motor Movement/Imagery Dataset* (https://physionet.org/content/eegmmidb/1.0.0/) [15] [16] dealing with it in three different ways: Non-Normalizing the Data (NN-DS), performing a Min-Max score normalization (MM-DS) and applying the Z-Score normalization (ZS-DS).

In this dataset are collected the EEG recordings of 109 subjects, who performed an experiment consisting of real and imagined movements of hands and feet. We focus our attention on the motor movement/imagery tasks of the left/right hand. The signal was recorded from the electrodes presented in Figure 1, with a sampling rate of 160 Hz. We reorganized the data, obtaining 4924 instances (2469 for the left hand) for the Motor Movement Task (MM-T) and 4915 (2479 for the left hand) for the Motor Imagery Task (MI-T).

III. Proposed Approach

The proposed pipeline (Figure 2) is divided in three main modules: feature computation, feature selection and classification through different support vector machine models.

All the data that are passed to the first module have been

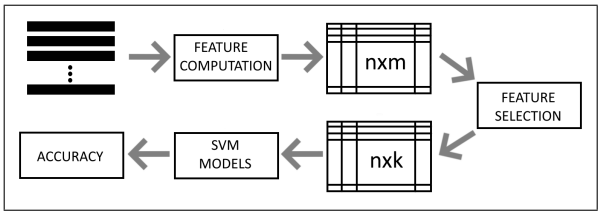

Figure 2. Proposed pipeline scheme.

pre-processed with a notch filter (50 Hz) to remove the direct current interference and with a finite impulse response filter in the range 7 - 31 Hz. In this specific case, we used the field knowledge to retain only the frequency bands of interest (μ and β), trying to have as less noise as possible without applying other noise removal techniques.

Different tests are conducted on the considered PhysioNet dataset with the previously cited configurations: NN-DS, MM-DS, ZS-DS.

In the following, we describe the main modules in more details.

A. Feature Computation

Our proposal includes features computed on each electrode in the time, frequency and time-frequency domains and also some statistical measures, in order to access the contributions given by different type of analyses. All the procedures are developed in MATLAB.

The Hjorth activity, mobility and complexity parameters [17] represent respectively the EEG signal power, the proportion of power spectrum standard deviation and the signal similarity to a sine wave [18]. Thus, they characterize the EEG signal in the time domain and they also have low computational cost. We developed the parameters following the formulae reported by *Oh et al.* [18].

Wanting to also have a representation of the data in the frequency and time-frequency domains, we estimate the power spectral density (PSD) using the Welch's method [19] and the complex Morlet wavelet [20] on the previously cited frequency bands of interest, i.e., μ and β. The Welch's method divides the signal into windows on which are computed the periodograms. Their average represents the PSD estimation. Instead, the complex Morlet wavelet is convolved with the EEG signal, obtaining the data power and phase. The idea underlying the development proposed by *Cohen* [20] is about being able to control the trade-off between time and frequency precision with the *cycle* parameter; thus, we exploit this characteristic and perform the feature computation with a better time-precision (3 cycles), a better frequency precision (7 cycles) and a trade-off between the two modalities (3 - 7 cycles).

Finally, we use the function provided by the MATLAB tool EEGLAB [21] to compute the mean, standard deviation, skewness, excess kurtosis, median, low/high percentile and trimmed mean/standard deviation on the signal obtained from each electrode.

As a final result, we obtain a vector of 1280 features: 64 electrodes $\times$ [2 frequency bands $\times$ (PSD estimate through Welch's method + 3 modalities $\times$ PSD extraction through Morlet wavelets) + statistical measures].

B. Feature Selection

As described in the previous sections, the evolutionary computation algorithms applied for feature selection have the advantage of being decoupled from the field knowledge and do not make any assumption on the features.

The developed EFS techniques are the genetic algorithm, particle swarm optimization and simulated annealing. They are Python coded and while the first two are modified versions of pre-existing codes, we developed the SA procedure following the pseudo-code provided by *Jeong et al.* [22]. Also, notice that we update the PSO velocity and position at each iteration exploiting the cognitive, social and inertia parameters presented by *Clerc et al.* [23]. All the parameters, reported in Table II, are empirically adapted to the presented problem, starting from the default values.

The EFS algorithms follow a wrapper approach, thus a SVM classification with radial basis kernel [24] and gamma scaled to $1/(N_{features} * variance(data))$ is applied on the dataset divided in training (80%) and test (20%) set. Moreover, we developed two different fitness functions for the evaluation step. The first one takes into consideration only the accuracy obtained by applying a SVM as the classifier, while the second one is meant to find the best trade-off between the number of

TABLE II. EVOLUTIONARY FEATURE SELECTION ALGORITHM PARAMETERS AND CONSTRUCTION COEFFICIENTS [23].

Algorithm	*Parameters and construction coefficients*
GA	iterations = 100; population size = 8; # parents = 4; # mutations = 3
PSO	construction coefficients: kappa = 1; $\phi_1 = \phi_2 = 2.05$; $\phi = \phi_1 + \phi 2$; $\chi = 2 \times \frac{kappa}{\lvert 2-\phi-\sqrt{\phi^2-4\phi}\rvert}$ parameters: iterations = 100; # particles = 30; # neighbors = 5; cognitive = $\chi \times \phi_1$; social = $\chi \times \phi_2$; inertia = χ; euclidean distance = 2
SA	initial temperature = 100000; temperature reduction = 0.9

selected features and the SVM accuracy, following the function presented by *Vieira et al.* [25]:

$$f(x) = \alpha(1 - acc) + (1 - \alpha)\left(1 - \frac{N_{sf}}{N_{if}}\right) \quad (1)$$

where $\alpha \in [0, 1]$ is a constant weighting the feature number/accuracy trade-off and is set to 0.88 after verifying that with lower/higher values, α does not provide better results; acc is the accuracy obtained by the SVM model; N_{sf} corresponds to the number of selected features, while N_{if} corresponds to the initial number of features.

The final result obtained by the EFS algorithms is a binary vector $1 \times N_{if}$, where 1s represent the selected features, 0s otherwise.

C. SVM Classifiers

The last main module is the one that performs the binary classification of left/right hand motor movement/imagination by applying the SVM models (Linear, Quadratic, Cubic, Fine/Medium/Coarse Gaussian) provided by the Classification Learner MATLAB application [26]. The SVMs are trained using a 5-fold cross-validation.

We use as benchmarks the models performed on the dataset retaining (1) all the computed features, (2) the feature subset selected a priori, consisting of the feature computed on the electrodes *C5, C3, Cz, C2, C4* (highlighted by the red line in Figure 1) and (3) the dimensions explaining at least the 95% of the data variance obtained by the principal component analysis, a standard procedure to reduce the feature number.

Afterwards, we apply the SVM models on the datasets obtained by the EFS techniques and using the feature subset representing the agreement between the evolutionary computation algorithms.

Finally, we compute the accuracy obtained by the various models.

IV. DISCUSSION

Following the previously described pipeline, we conducted 11 tests, whose best results are summarized in Table III and in Table IV. The benchmark corresponds to the first three entries of the tables, while the tests on the proposed EFS methods are reported in the remaining rows.

The benchmark consists of the results obtained by applying the Classification Learner SVM models to all the dataset types, i.e., non-normalized (NN-DS), min-max score (MM-DS) and z-score (ZS-DS) normalized.

Notice that the best result achieved by the benchmark (67.8% of accuracy) for the motor left/right hand movement task (from now on called MM-T) is obtained by the cubic SVM on the ZS-DS retaining all the computed features. The dataset consisting of the features selected a priori and of the dimension extracted by the PCA do not have comparable results.

Comparable accuracy values are achieved by some of the evolutionary feature selection models, which are computed only on the normalized dataset, noticing that the best results achieved by the benchmark tests are on ZS-DS and MM-DS. In particular, the GA and SA algorithms obtained the same result as the best benchmark test using as fitness function the trade-off between the feature number and accuracy value, while the PSO using the trade-off function exceeds the best result with the 68.0% of accuracy. Even though the SA with the only accuracy fitness function achieves the best result (68.3%) on MM-T, we highlight the fact that the technique retains 1117 of the 1280 total features. Thus, we consider the SA result not fitting our purpose, i.e., having a minimum feature subset that can guarantee a comparable/better accuracy on the original dataset. Therefore, we elect the GA and PSO with the trade-off function as the best methods.

As a final remark on MM-T, we highlight that the results obtained by the SVM models applied on the feature subset generated by the EFS algorithms approach the best accuracy values returned by the previous tests. Also, notice that the ZS-DS is the most present dataset in Table III, suggesting that in a population-based analysis a z-score normalization seems to be the best approach.

Concerning the motor left/right hand imagination task (from now on called MI-T), surprisingly the best accuracy (64.3%) is obtained by the linear SVM on the NN-DS retaining all the computed features. The observations on the a priori feature selection and PCA tests for MI-T are the same reported for the MM-T.

A comparable result is achieved by PSO with the trade-off function (64.0% of accuracy), which selects 714 of the 1280 features on the ZS-DS. On the feature agreement, the quadratic SVM model obtains 63.3% of accuracy when applied on the ZS-DS with the trade-off fitness function. The z-score normalization seems to be confirmed as the best approach in a population-based analysis.

A possible reason behind the decrease in all the accuracy values on MI-T compared with MM-T, is represented by the inability of accessing if the subject performed correctly the imagination of the left/right hand movement, thus causing an uncontrolled introduction of outliers.

However, notice that there are numerous similarities in the results obtained on both tasks. Focusing on the benchmark, the a priori feature selection and the PCA dimensions are unable to provide an accuracy comparable to the result

TABLE III. BEST RESULTS OBTAINED IN EACH TEST ON MOTOR LEFT/RIGHT HAND MOVEMENT (MM-T).

Test	*SVM model*	*Dataset*	*# features*	*Accuracy (%)*
all features	cubic	ZS-DS	1280	67.8
a priori	mean Gaussian	ZS-DS	100	62.7
PCA	quadratic	MM-DS	43	62.3
GA accuracy	cubic	ZS-DS	662	67.2
GA trade-off	cubic	ZS-DS	646	67.8
PSO accuracy	cubic	ZS-DS	620	67.3
PSO trade-off	quadratic	ZS-DS	675	68.0
SA accuracy	cubic	ZS-DS	1117	68.3
SA trade-off	cubic	ZS-DS	1116	67.8
agreement accuracy	quadratic	ZS-DS	264	66.4
agreement trade-off	cubic	ZS-DS	308	67.5

 ISBN: 978-1-61208-829-7

TABLE IV. BEST RESULTS OBTAINED IN EACH TEST ON MOTOR LEFT/RIGHT HAND IMAGINATION (MI-T).

Test	*SVM model*	*Dataset*	*# features*	*Accuracy (%)*
all features	linear	NN-DS	1280	64.3
a priori	linear	ZS-DS	100	59.7
PCA	quadratic	MM-DS	41	59.5
GA accuracy	cubic	ZS-DS	641	63.8
GA trade-off	quadratic	ZS-DS	608	63.7
PSO accuracy	cubic	MM-DS	622	61.7
PSO trade-off	quadratic	ZS-DS	714	64.0
SA accuracy	cubic	ZS-DS	1114	63.6
SA trade-off	cubic	ZS-DS	1117	63.8
agreement accuracy	cubic	ZS-DS	272	62.4
agreement trade-off	quadratic	ZS-DS	313	63.3

obtained by the test retaining all the features. Moving to the proposed EFS techniques, observe that: (1) the various models generally achieve the best accuracy on the ZS-DS; (2) the activation function exploiting the trade-off between the number of selected features and the accuracy, allows the EFS methods to achieve a better accuracy compared to the accuracy-only activation function; (3) the best methods are GA and PSO, which maintain good performances with a restricted feature subset; (4) the SA technique retains about the 87% of the original features on average, thus not representing a good solution for the feature minimization and accuracy maximization problem.

Having a general description of the best results, we now focus our attention on the EFS agreement; in particular on the feature subset selected through the trade-off fitness function.

Figure 3 and Figure 4 report the number of features selected for each electrode on MM-T and MI-T respectively. Table V summarizes how frequently a specific feature type is selected in the EFS feature selection agreement on the motor left/right hand movement/imagination tasks.

Observe that the set of electrodes that are usually selected a priori (*C5, C3, Cz, C2, C4*) contribute minimally to the classification in both tasks. Their contribution is also not symmetrical, i.e., if one of these electrodes is selected in the left hemisphere of the brain, most probably the corresponding one in the right hemisphere is not selected. This could actually be an optimal configuration, knowing that each of these electrodes is at least coupled with a feature in the time-frequency domain. As stated in the introduction, we know that the power spectrum decreases on the μ band, while it increases in the β band when dealing with a motor related task and also it has a spatial connotation depending on the fact that the movement/imagination is intended for the left/right hand.

Concerning the other electrodes related to the motor cortex (light-blue highlighted in Figure 1), the number of contributions is greater for MI-T than for MM-T, which reports some specifically localized contributions.

The frontal (*Fp, AF, F*) sensors are involved in the motor tasks, probably due to the experimental settings. In fact, a subject had to perform the motor left/right hand movement/imagination following a visual cue, thus involving the specific brain area coupled with the previously cited electrodes.

The parietal (*P*) sensors make some contributions, especially with the statistical features. This brain area is deputed to sensory information and thus could be involved in the motor tasks.

Finally, the temporal, parieto-occipital and occipital (*T, PO, O*) electrodes give some information, mostly through Hjorth parameters and statistical measures. This could be due to their brain area activities concerning memory and visual processing.

We finally report the selection frequency of the various feature types (Table V).

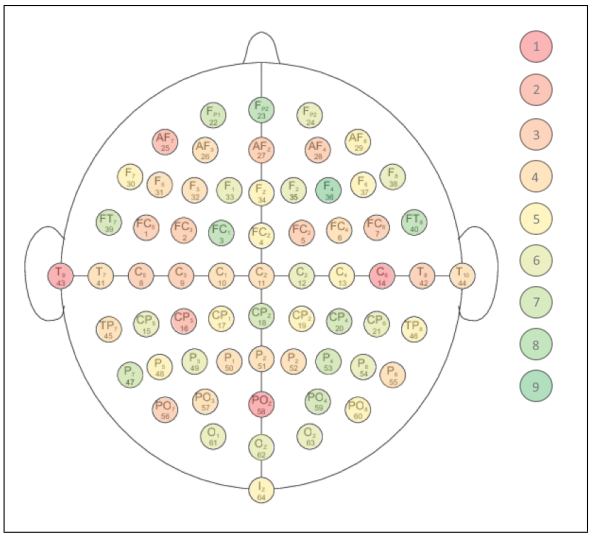

Figure 3. Agreement electrode contributions for the motor left/right hand movement (MM-T) task.

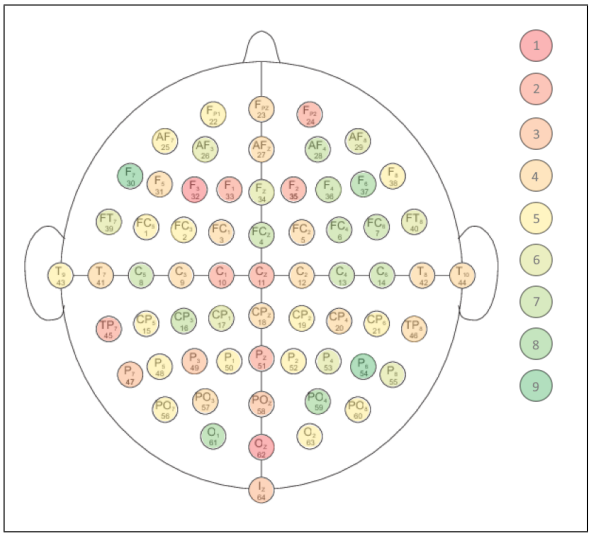

Figure 4. Agreement electrode contributions for the motor left/right hand imagination (MI-T) task.

In MM-T there is a balanced selection of the features involving the information in the frequency domain. Between the Morlet wavelet related features, the most selected is the power spectral density extracted using this technique on the μ band with a time-frequency trade-off. There is also a good balance in the same type of features selected for MI-T.

The Hjorth parameters have a big impact especially for MI-T, where the activity parameter appears 27 times and thus with the highest frequency in respect to the other feature types.

The standard deviation and median features give a great contribution especially for MM-T.

The rest of the feature types are balanced for both tasks.

 ISBN: 978-1-61208-829-7

TABLE V. FREQUENCY OF FEATURE TYPE SELECTION FOR MM-T AND MI-T.

Feature type	*Frequency on MM-T*	*Frequency on MI-T*
Hjorth activity	13	27
Hjorth mobility	13	14
Hjorth complexity	15	8
PSD Welch on μ	17	17
PSD Welch on β	14	12
PSD Morlet time-prec on μ	13	11
PSD Morlet time-prec on β	14	15
PSD Morlet freq-prec on μ	15	16
PSD Morlet freq-prec on β	14	16
PSD Morlet trade-off on μ	17	13
PSD Morlet trade-off on β	10	16
Mean	16	15
Standard deviation	20	18
Skewness	15	19
Excess kurtosis	16	17
Median	25	17
Low percentile	17	16
High percentile	14	18
Trimmed mean	18	15
Trimmed standard deviation	11	12

V. CONCLUSION AND FUTURE WORK

In this work we investigated the possibility of conducting a preliminary analysis on the data provided by a MI-based BCI, to improve the interaction between this peculiar kind of system and its users. In particular, we concentrated our efforts on the PhysioNet *EEG Motor Movement/Imagery Dataset* tasks concerning the motor left/right hand movement/imagination.

Firstly, we noticed that most of the data normalized by the z-score normalization technique achieves better results and thus allow a population-based analysis. We can assume that by applying the data normalization, the data heterogeneity due to inter- and intra-subject variability is mitigated. Therefore, we can exploit this results to have a higher number of instances per class when dealing with a MI-based BCI.

Secondly, we computed different feature types in the time, frequency and time-frequency domains and also as statistical measures, wanting to have a broad insight on their contributions.

We verify on this specific dataset that the feature types contribute in the task discrimination depending on the electrodes on which they are computed. Thus, the brain area localization is an important information.

We notice that not only the electrodes covering the motor cortex are involved in the motor tasks with their time-frequency related features, but also the other brain areas contribute with different types of features, especially the Hjorth parameters and the statistical measures.

The evolutionary feature selection algorithms represented great allies in the optimal feature subset search. In particular, the genetic and particle swarm optimization algorithm obtained the best results, having the support vector machine models (applied on the reduced datasets) obtained comparable or better results in respect to the benchmark ones.

Finally, the agreement of the EFS techniques on the selected features has highlighted the various contribution from each electrode and from each feature type, without decreasing drastically the accuracy values.

As future work, we would like to test the EFS techniques with different fitness functions and on different datasets. Using data obtained by experimental protocols that do not only involve the motor imagery, but also cognitive workload, emotion recognition and so on, we could simulate a real-life scenario and verify if our approach is generalizable.

We would also like to define our own experimental protocol for a live MI-based BCI, taking into account the ergonomic issues that could be involved in this user-centric system modeling. In fact, as stated by *Baek et al.* [27], most of the BCI related works do not consider the importance of having user-friendly, flexible and accessible systems, which could allow a better EEG recording in absence of stress and discomfort.

REFERENCES

[1] S. Ackerman et al., Discovering the brain. National Academies Press, 1992.

[2] A. Zani, A. Proverbio, G. Mangun, E. Fletcher, E. Brattico, C. Olcese, M. Tervaniemi, R. Näätänen, E. Wilding, K. Federmeier, M. Kutas, R. Knight, D. Scabini, P. Luu, and D. Tucker, Metodi Strumentali nelle Neuroscienze Cognitive. EEG ed ERP- Instrumental Methods in Cognitive Neuroscience. EEG and ERP. Aracne Editrice, 11 2013.

[3] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," sensors, vol. 12, no. 2, 2012, pp. 1211–1279.

[4] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," Journal of neural engineering, vol. 15, no. 3, 2018, p. 031005.

[5] P. Szczuko, "Real and imaginary motion classification based on rough set analysis of EEG signals for multimedia applications," Multimedia Tools and Applications, vol. 76, no. 24, 2017, pp. 25 697–25 711.

[6] M. Dai, D. Zheng, R. Na, S. Wang, and S. Zhang, "EEG classification of motor imagery using a novel deep learning framework," Sensors, vol. 19, no. 3, 2019, p. 551.

[7] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," Journal of neural engineering, vol. 16, no. 3, 2019, p. 031001.

[8] B. Rim, N.-J. Sung, S. Min, and M. Hong, "Deep learning in physiological signal data: A survey," Sensors, vol. 20, no. 4, 2020, p. 969.

[9] B. Xue, M. Zhang, and W. N. Browne, "New fitness functions in binary particle swarm optimisation for feature selection," in 2012 ieee congress on evolutionary computation. IEEE, 2012, pp. 1–8.

[10] R. Kohavi and G. H. John, "The wrapper approach," in Feature extraction, construction and selection. Springer, 1998, pp. 33–50.

[11] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," IEEE Transactions on Evolutionary Computation, vol. 20, no. 4, 2015, pp. 606–626.

[12] A. Atyabi, M. Luerssen, S. Fitzgibbon, and D. M. Powers, "Evolutionary feature selection and electrode reduction for EEG classification," in 2012 IEEE congress on evolutionary computation. IEEE, 2012, pp. 1–8.

[13] K. Amarasinghe, P. Sivils, and M. Manic, "EEG feature selection for thought driven robots using evolutionary algorithms," in 2016 9th International Conference on Human System Interactions (HSI). IEEE, 2016, pp. 355–361.

[14] J. González, J. Ortega, M. Damas, P. Martín-Smith, and J. Q. Gan, "A new multi-objective wrapper method for feature selection–Accuracy and stability analysis for BCI," Neurocomputing, vol. 333, 2019, pp. 407–418.

[15] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," circulation, vol. 101, no. 23, 2000, pp. e215–e220.

[16] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: a general-purpose brain-computer interface (BCI) system," IEEE Transactions on biomedical engineering, vol. 51, no. 6, 2004, pp. 1034–1043.

[17] B. Hjorth, "EEG analysis based on time domain properties," Electroencephalography and clinical neurophysiology, vol. 29, no. 3, 1970, pp. 306–310.

 ISBN: 978-1-61208-829-7

[18] S.-H. Oh, Y.-R. Lee, and H.-N. Kim, "A novel EEG feature extraction method using Hjorth parameter," International Journal of Electronics and Electrical Engineering, vol. 2, no. 2, 2014, pp. 106–110.

[19] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," IEEE Transactions on audio and electroacoustics, vol. 15, no. 2, 1967, pp. 70–73.

[20] M. X. Cohen, "A better way to define and describe Morlet wavelets for time-frequency analysis," NeuroImage, vol. 199, 2019, pp. 81–86.

[21] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," Journal of neuroscience methods, vol. 134, no. 1, 2004, pp. 9–21.

[22] I.-S. Jeong, H.-K. Kim, T.-H. Kim, D. H. Lee, K. J. Kim, and S.-H. Kang, "A feature selection approach based on simulated annealing for detecting various denial of service attacks," Software Networking, vol. 2018, no. 1, 2018, pp. 173–190.

[23] M. Clerc and J. Kennedy, "The particle swarm-explosion, stability, and convergence in a multidimensional complex space," IEEE transactions on Evolutionary Computation, vol. 6, no. 1, 2002, pp. 58–73.

[24] R. Bousseta, S. Tayeb, I. El Ouakouak, M. Gharbi, F. Regragui, and M. M. Himmi, "EEG efficient classification of imagined hand movement using RBF kernel SVM," in 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA). IEEE, 2016, pp. 1–6.

[25] S. M. Vieira, L. F. Mendonça, G. J. Farinha, and J. M. Sousa, "Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients," Applied Soft Computing, vol. 13, no. 8, 2013, pp. 3494–3504.

[26] MATLAB, 9.5.0.1033004 (R2018b) Update 2. Natick, Massachusetts: The MathWorks Inc., 2018.

[27] H. J. Baek, M. H. Chang, J. Heo, and K. S. Park, "Enhancing the usability of brain-computer interface systems," Computational intelligence and neuroscience, vol. 2019, 2019.

 ISBN: 978-1-61208-829-7

A Preliminary Analysis of the Physiological Response Generated by Negative Thoughts

Nagore Sagastibeltza,
Asier Salazar-Ramirez
and Raquel Martinez

Faculty of Engineering in Bilbao
University of the Basque Country (UPV/EHU)
Bilbao, Spain 48013
Email: nagore.sagastibeltza@ehu.eus,
asier.salazar@ehu.eus,
raquel.martinez@ehu.eus

Ainhoa Yera
and Javier Muguerza

Faculty of Informatics
University of the Basque Country (UPV/EHU)
Donostia-San Sebastián, Spain 20018
Email: ainhoa.yera@ehu.eus,
j.muguerza@ehu.eus

***Abstract*—Positive and negative emotions have a great impact on the human organism. The main purpose of this work is to analyze the physiological effects of negative thoughts using affective computing techniques. With this aim, we carried out an experiment in which participants had to recall past negative experiences while their physiological signals were being collected. Then, using two algorithms based on biosignal analysis, we assessed their levels of stress and relaxation at each time. According to the results, 100% of the participants who had negative thoughts produced a physiological stress response. This is promising, since it could enable affective computing based systems to adapt to the emotional state of the users through the real-time monitoring of the physiological signals of the human body.**

Keywords–Affective computing; Emotions; ECG; GSR; Negative thoughts.

I. INTRODUCTION

Feeling emotions, either positive or negative, is common in every human being. Nowadays, one of the most demanded skills in different social and/or labor environments [1] is knowing how to control emotions and thoughts generated as a reaction to particular situations. Affective computing provides useful tools for emotion recognition by developing algorithms and devices that can interpret human emotions [2].

We ourselves create our emotions, based on the interpretation we have of the information we receive [3] [4]. In our everyday life, we tend to generate thoughts, emotions and actions unconsciously [5]. Controlling and educating our thoughts is of paramount importance in order to have increasingly positive emotions and consequently, behaviors and actions that provide a greater well-being to our lives. In contrast, negative thoughts cause negative and frustrating emotions and actions. Because the physiological response of our organism is in sync with our thoughts, stressful thoughts or painful memories increase the cortisol, lower the defenses, deteriorate the physiology of the arteries, activate the nervous system, and accelerate the heart rate [6] [7].

To this regard, affective computing provides methods that recognize human emotions and help to identify what everyone is feeling and living [8]. This paper aims to analyze the effects of negative thoughts in the physiological response, more precisely, in those relating to the heart and sweat. To do so, Section II explains the experimental design used for collecting a physiological signal database related to stressful memories. Section II also presents the collected biosignals, as well as the studied population. Then, the analysis of the database is covered in Section III, providing explanations of the results obtained from the application of two affective computing algorithms. Finally, the conclusions and future lines of the work are presented in Section IV.

II. METHODOLOGY

In this section, we describe the methodology of experimentation used to obtain the psychological database analyzed. This procedure has been certified by the corresponding ethical committee CEISH-UPV/EHU, BOPV 32 (M10_2016_189).

A. Experimental setup

In order to analyze the physiological response to negative thoughts, we have designed an experiment where participants had to recall a problematic or stressful situation from the past using visualization techniques, while their electrodermal and heart activities were being captured. Several researches have shown that visualization techniques bring the imagination closer to the reality [9] [10].

The experiment had several stages and was conducted individually for each participant (see Fig. 1). In the first stage, we ensured the participants were comfortable in their position, described to them the experimental procedure and gave them a consent form to sign. Then, we connected the sensors to the participants and provided some time so that they could choose a stressful situation to use during the experiment.

In the second stage, at the start off of the experiment, the participants watched a relaxing video with the lights off. Thereon, we asked the users to close their eyes and let themselves be guided by the words of the experimenter while listening to quiet music in the background. In this first part of the visualization, the participants were induced into a deeper level of relaxation, and once they reached this state, they were asked to bring back the previously chosen stressful situation. Finally, using neurolinguistic programming techniques, the subjects were guided to solve that conflict [11]. Once the visualization was completed, a new relaxing

 ISBN: 978-1-61208-829-7

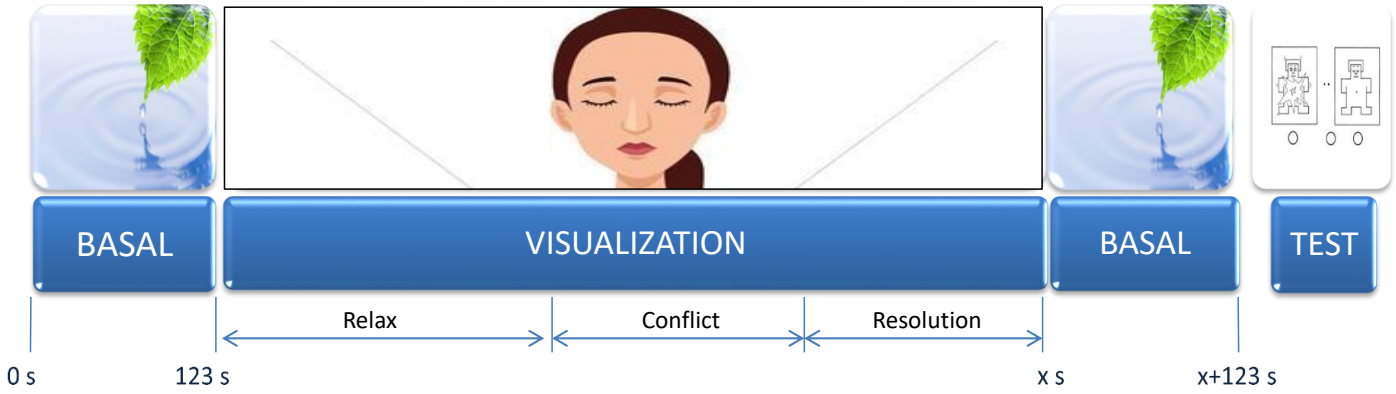

Figure 1. Stages of the experiment carried out.

video was projected to bring the participants back to a basal emotional situation.

The last stage of the experiment ended with the emotional evaluation using the Self-Assessment Manikin questionnaire [12] and a personal interview. Regarding the stages of the experiment represented in Fig. 1, it should be noted that although users had similar visualization duration records, these were not equal, since they were adapted to the participant's non-verbal communication at each moment.

Fourteen volunteer students from the Faculty of Engineering in Bilbao of the University of the Basque Country (UPV/EHU) were recruited for the experiment. However, due to technical problems, one of the participants was removed from the study. Thus, a total of 13 participants (9 males and 4 females), 19-22 years old (mean=20.30, standard deviation=0.94) were taken into account for the analysis.

B. Biosignals and data acquisition system

This work is based on the study of two particular biosignals, the Heart Rate Variability (HRV) and the Galvanic Skin Response (GSR). The HRV is a signal derived from the electrocardiogram (ECG) and it is representative of the heart's activity. This signal provides information of the variation in time of the heartbeat, with high variability indicating a healthy heart. On the other hand, the GSR signal is representative of the conductive capacity of the skin's surface. This signal has been widely used in the area of electrophysiology, since besides providing information on the body's thermoregulatory activity, its variations are indicative of different psychological phenomena: nerves, surprise, anxiety, etc.

Regarding the regulation of the signals mentioned in the previous paragraph, this is managed by the autonomous nervous system and most of these regulatory functions take place in the hypothalamus. In turn, the hypothalamus is strongly connected with the amygdala, that is the part of the brain responsible for, among other functions, the emotional responses. Considering this connection, it is not surprising that emotional changes produce changes in the physiological balance of the organism, with the HRV and GSR signals being affected, among others, hence the reason we selected them for this study.

Concerning the materials used in the experiment, the capture of the aforementioned signals was done using the commercial hardware Biopac MP36 and its associated software Studentlab, which are considered a reference experimental equipment for gathering physiological signals. The acquisition was configured at a sampling frequency of 1000 Hz and was performed with the corresponding electrodes to collect the ECG and GSR signals. In addition, the room in which the experiment took place was equipped with a PC to store the signals and with the audiovisual material necessary to project the two relaxing videos (projector, screen and speakers).

III. Results and Analysis

After collecting the aforementioned physiological signals, the next step of this study was to analyze the participants' physiological reactions to the experiment. Different algorithms can be used depending on what the study looks for. For instance, for continuous online analysis, some researchers have used algorithms such as fuzzy logic [13], support vector machines [14] or artificial neural networks [15] to compute physiological signals. However, in this case, the target of the experiment was to analyse whether the participants' negative memories had an impact on their physiology. Therefore, the results belong to the discrete domain, and thus, for the sake of continuity, we decided to use two algorithms previously developed by the research team, which assess both stress and relaxation in a discrete manner: algorithm 1 [16] and algorithm 2 [17], respectively, named ALG 1 and ALG 2 hereinafter.

ALG 1 was designed for detecting arousal of the Autonomic Nervous System (ANS) caused by a stressful experience. The output of ALG 1 varies discretely from 0 to 6 according to the intensity and duration of the arousal, output 0 meaning that there is no stressful arousal. Then, output levels 1, 3 and 5, respectively, correspond to the detection of a short-duration arousing alert of low, medium and high-intensity. Finally, output levels with even numbers (2, 4 and 6) correspond to the detection of ANS activations that last for longer than 30s. For instance, ALG 1 will trigger output level 1 if it detects an ANS activation of low intensity. If the activation lasts over 30s, then ALG 1 will trigger output level 2. Accordingly, output levels 4 and 6 would be triggered when 3 and 5 activation levels lasted for longer than 30s. Therefore, even-number outputs stand for sustained stressful activations.

On the other hand, ALG 2 is used for detecting the opposite type of reaction: the inhibition of the ANS or a relaxing

 ISBN: 978-1-61208-829-7

response. In the case of ALG 2, its output is bounded to the [-3, 0] range and, as it happened with ALG 1, level 0 is related to the absence of any relaxing response. Then, similarly to ALG 1, output levels -1, -2 and -3 correspond to low, medium and high-intensity relaxation responses, respectively. However, unlike ALG 1, the outputs of ALG 2 do not take into account the time-length inhibition of the ANS.

After applying both algorithms to the physiological signals of the participants, we obtained similar results to the ones shown in Fig. 2. The three a) graphs of the figure show the data of the subject 12 of the experiment, whereas the three b) graphs correspond to the subject 5. The first two graphs of each participant correspond to the HRV and GSR signals respectively, plotting green the raw signal and black the low-pass filtered signal. Then, the third graph shows the outputs of both ALG 1 (in red) and ALG 2 (in blue). Although only data from two participants is depicted in Fig. 2, these physiological patterns are also representative physiological reaction patterns of the other volunteers. The first main pattern, shown in charts a) of Fig 2., represents those subjects that could recall a negative memory which had a physiological impact on the organism (subject 12, for instance). On the other hand, the patterns shown in charts b) of Fig. 2, correspond to subject 5 and are representative of the cases in which the participant did not get involved or that could not recall a negative memory.

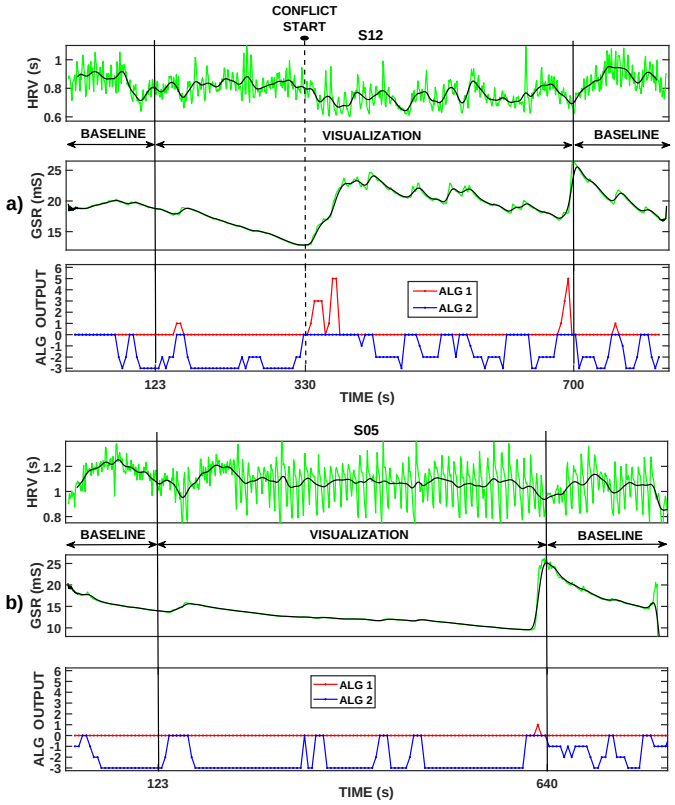

Figure 2. Physiological signals (HRV and GSR) and algorithm outputs. The charts of a) corresponds to subject 12 and the charts of b) to subject 5.

According to Fig. 2 a), subject 12 was able to relax during the beginning of the visualization stage. However, his ANS activated when he started remembering his conflict and bad memories came to his mind. This activation produced an acceleration of the cardiac rhythm, along with the subsequent decrease of the HRV and an increase in the sweating. Looking at the third graph, it is possible to see how the two types of physiological patterns were correctly detected by both ALG 1 and ALG 2.

On the contrary, Fig. 2 b) presents another type of situation in which subject 5 did not feel stressed at all during the experiment. According to the personal interview, subject 5 could not manage to remember any personal conflict and so, no negative memories were recalled. This absence of negative memories were clearly reflected in the outputs of the algorithms, which indicated high levels of relaxation during the conflict evoking stage.

As seen in Fig. 2, there are significant differences in how participants reacted during the experiment. Most of them mentally recalled a personal conflict, but three of them did not do so, either because they could not manage to do it, they did not want to face again such a delicate experience, or because they did not want be involved in the experiment. This information is summarized in Table I, where the first row gives the identification code of the subjects and the second (S.E.) gives the information on whether they recalled the personal conflict (Y) or not (N). Finally, the last row (S_{max}) presents the maximum ANS activation level detected by ALG 1.

TABLE I. SUMMARY OF THE EXPERIMENTAL DATA AND ALGORITHM OUTPUTS.

	Subjects												
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
S.E.	N	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y	Y
S_{max}	0	3	3	3	0	3	5	5	1	0	3	5	5

The content of Table I shows how ALG 1 gives an output value that is coherent with the information provided by the participants in their respective interviews. For all the subjects who stated that they had felt stressed, the algorithm detected all those ANS activations with different intensity levels (see Fig. 2 and Table I). Besides, ALG 1 gave a 0 stress level output for the three participants that were unable to either evoke the conflict or get stressed, whereas ALG 2 detected high relaxation levels for those subjects (output level -3).

Hence, this preliminary study corroborates the hypothesis that negative thoughts generate similar physiological variations to the ones that stress produces on the organism: all the subjects that thought about and recalled a conflictive situation suffered from ANS activations (getting a 100% accuracy for the used detection algorithms).

IV. CONCLUSIONS

The study presented in this work shows the steps of an analysis relating physiological changes to negative thoughts. For this initial analysis stage, we designed an experimental setup in which 13 participants had to recall a personal conflict using guided visualization techniques. As mentioned in Section III, it has been possible to confirm that not all the participants could recall a negative situation of such characteristics. This work also corroborates Schachter's and Singer's cognitive theory [18] that stated that it is not the stimulus the one which produces the emotional reaction on the organism, but the person's cognitive perception of the stimulus. Besides, in the case of all the participants that could evoke a past conflict, their

 ISBN: 978-1-61208-829-7

organism reacted in the same manner as if they were going through a stressful situation. These results help to provide adaptations according to the emotional state of the users based on the physiological information of the human body.

As future lines, we would like to widen the study by adding new biomarkers related to stress, such as cortisol or electroencephalographic signals. By doing this, we aim to clarify to what extent the negative thoughts can affect the organism. Besides, this initial study is limited to a very reduced population. Therefore, we plan to expand the cases of study to a larger population. Finally, we are designing a new set of experiments in which bad thoughts and feelings are elicited with other types of techniques from the field of psychology and also using audio-visual stimulation.

ACKNOWLEDGMENT

This work has been funded by the following units: First, by the research group ADIAN which is supported by the Department of Education, Universities and Research of the Basque Government, (grant IT980-16). Second, by the Ministry of Economy and Competitiveness of the Spanish Government, co-founded by the ERDF (PhysComp project, TIN2017-85409-P).

REFERENCES

[1] P. Lopes, D. Grewal, J. Kadis, M. Gall, and P. Salovey, "Evidence that emotional intelligence is related to job performance and affect and attitudes at work," Psicothema, vol. 18, no. Suplemento, 2006, pp. 132–138.

[2] E. Cambria, "Affective computing and sentiment analysis," IEEE Intelligent Systems, vol. 31, no. 2, 2016, pp. 102–107.

[3] C. Casado and R. Colomo, "Un breve recorrido por la concepción de las emociones en la filosofía occidental," A parte Rei, vol. 47, no. 10, 2006.

[4] R. Lazarus, "The cognition-emotion debate: A bit of history," in Handbook of cognition and emotion, T. Dalgleish and M. Power, Eds., 1999, vol. 5, no. 6, pp. 3–19.

[5] W. Wood, J. M. Quinn, and D. A. Kashy, "Habits in everyday life: Thought, emotion, and action." Journal of personality and social psychology, vol. 83, no. 6, 2002, p. 1281.

[6] L. Capobianco, A. P. Morrison, and A. Wells, "The effect of thought importance on stress responses: a test of the metacognitive model," Stress, vol. 21, no. 2, 2018, pp. 128–135.

[7] H. Gu, C. Tang, and Y. Yang, "Psychological stress, immune response, and atherosclerosis," Atherosclerosis, vol. 223, no. 1, 2012, pp. 69–77.

[8] S. Wang, P. Phillips, Z. Dong, and Y. Zhang, "Intelligent facial emotion recognition based on stationary wavelet entropy and jaya algorithm," Neurocomputing, vol. 272, 2018, pp. 668–676.

[9] M. Hoffart and E. Keene, "Body-mind-spirit: the benefits of visualization," AJN The American Journal of Nursing, vol. 98, no. 12, 1998, pp. 44–47.

[10] R. Burkhard, "Strategy visualization: A new research focus in knowledge visualization and a case study," in Proceedings of I-KNOW, vol. 5, 2005, pp. 1–8.

[11] J. Lee, "Teaching NLP for conflict resolution," The law teacher, vol. 34, no. 1, 2000, pp. 58–76.

[12] P. Lang, "Self-assessment manikin," Gainesville, FL: The Center for Research in Psychophysiology, University of Florida, 1980.

[13] A. de Santos Sierra, C. Ávila, J. Casanova, and G. B. del Pozo, "A stress-detection system based on physiological signals and fuzzy logic." IEEE Transactions on Industrial Electronics, vol. 58, no. 10, 2011, pp. 4857–4865.

[14] G. Sakr, I. Elhajj, and H. Abu-Saad, "Support vector machines to define and detect agitation transition." IEEE Transactions on Affective Computing, vol. 1, no. 2, 2010, pp. 98–108.

[15] N. Sharma and T. Gedeon, "Artificial neural network classification models for stress in reading." in International Conference on Neural Information Processing (ICONIP 2012), vol. 7666. Springer, 2012, pp. 388–395.

[16] R. Martinez, E. Irigoyen, A. Arruti, J. Martín, and J. Muguerza, "A real-time stress classification system based on arousal analysis of the nervous system by an f-state machine," Computer methods and programs in biomedicine, vol. 148, 2017, pp. 81–90.

[17] R. Martinez et al., "A self-paced relaxation response detection system based on galvanic skin response analysis," IEEE Access, vol. 7, 2019, pp. 43 730–43 741.

[18] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state." Psychological review, vol. 69, no. 5, 1962, p. 379.

 ISBN: 978-1-61208-829-7

Estimation of Body Part Acceleration While Walking Using Frequency Analysis

Estimating head acceleration from movement of upper trunk

Shohei Hontama, Kyoko Shibata, Yoshio Inoue
Kochi University of Technology
Miyanokuchi 185, Tosayamada, Kami, Kochi, Japan
e-mail: hontama.kut@gmail.com
e-mail: shibata.kyoko@kochi-tech.ac.jp
e-mail: inoue.yoshio@kochi-tech.ac.jp

Hironobu Satoh
National Institute of Information and Communications Technology
Nukui-kitamachi, Koganei, Tokyo, 187-8795, Japan
e-mail: satoh.hironobu@nict.go.jp

***Abstract*— This study group is developing a mobile system that can easily estimate the floor reaction force with a small number of wearable sensors for reduced user burden. In this study, we propose a method to measure accelerations by wearable inertial sensors and we estimate the floor reaction force based on these measurements. In previous works, the number of sensors was reduced from 15 body parts to 5 selected body parts, in order to decrease burden on a user. However, the estimation accuracy also decreased. Therefore, in this paper, we consider reducing the number of sensors without sacrificing accuracy. In the previous report, the relations between the acceleration of each part of the body were quantified by analyzing the acceleration of each part in the Fourier analysis and expressing it in the frequency domain. This paper quantifies the relations between the accelerations of the head and the upper trunk using previously reported methods, then estimates the head acceleration from the upper trunk acceleration. As a result, it was possible to capture the characteristics of the head acceleration in two directions while walking and it was also possible to estimate it accurately.**

Keywords- Fourier analysis; gait analysis; motion mode function.

I. INTRODUCTION

Walking is one of the most familiar activities performed by many people. Gait analysis data are important information in the fields of healthcare, clinical medicine and sports, so they are effective for health promotion, rehabilitation, improving athletic function in athletes, and so on. General gait analysis is performed by combining an optical Motion Capture (MC) system and force plates. The advantage of this measurement system is that it enables detailed analysis of gait, such as calculation of joint moments. However, this measurement system is very expensive, moreover there is a limit to the measurement range for using installation type equipment. To address this problem, our study group proposed a method for estimating floor reaction forces using wearable inertial sensors and succeeded in making it mobile [1]. This method divided the body into 15 parts, as reported by Ae et al. [2], then it was shown that the sum of the inertial forces and gravity obtained from each of these parts is balanced with the measured values of the force plate. To accurately estimate the floor reaction force, it is achievable to use each inertial force derived from the acceleration measured from 15 parts of the whole body. However, for a simple system to reduce the burden on the user, the idea is to reduce the number of sensors. Attempts to estimate using a small number of inertial sensors selected so far resulted in poor accuracy [1].

Therefore, in this paper, we consider reducing the number of sensors without sacrificing accuracy. One method is to estimate the acceleration at the unmeasured part from the measured part, which reduces the number of sensors and the loss of accuracy caused by the reduction in the number of sensors. For this purpose, it is necessary to understand the relation between the movement of the measured part and the unmeasured part. However, it is not easy to describe the relation quantitatively using the time history waveforms as they are. Hence, in the previous report [3], walking was regarded as a periodic motion and Fourier analysis was performed on the acceleration of each part. As a result, using the acceleration expressed in the frequency domain, it was possible to extract the characteristics of each part of the movement. Based on this, the unmeasured part was divided by the measured part for each frequency component. This is called "motion mode function" in this paper. This motion mode function enabled us to quantify the relation between the acceleration of the two target parts. In this paper, the acceleration at the unmeasured part is estimated in the frequency domain using the obtained motion mode function and is converted into an acceleration in the time domain by inverse Fourier analysis. The usefulness of the method is investigated by comparing the estimated acceleration with the measured acceleration.

This report shows the estimation results of the acceleration in the vertical direction and walking direction, with the head and the upper trunk as target parts, where the difference in movement is relatively large.

II. METHOD

A. *Method for obtaining acceleration data for Fourier analysis*

The development of a simple system using inertial sensors is our goal. However, in this paper, acceleration data

 ISBN: 978-1-61208-829-7

for analysis are acquired using MC as a basic study to estimate acceleration at the unmeasured part.

In the experiment, MC (manufactured by Motion Analysis Co., Ltd.), force plate 3 units (manufactured by Tec Gihan Co., Ltd., TF-6090-C 1 unit, TF-4060-D 2 units), metronome were used. Force plates are strain-gauge transducer that can measure forces, moments, and centers of pressure. Acceleration is obtained by attaching recurrent markers to the head and upper trunk, as defined by Ae et al. [2], as shown in Figure 1.

Figure 1. Position of recursive markers.

Two healthy subjects (male: age 22±0, height 1.75±0.05 [m], weight 65±5 [kg]) are measured and 10 steps are taken from the start of walking. The acceleration data to be examined are for one gait cycle of two steps (1.2 seconds) of the fifth step and the sixth step, which are steady walking. The cadence is set to 100 BPM (0.6 seconds per walking cycle) with a metronome. Each subject measure 15 times of trial data, according to the rhythm of the metronome. The accuracy of the analysis is improved if both ends of the acceleration data of one gait cycle are the same. Therefore, measurements were taken after sufficient walking exercises by following per the metronome so that both ends of the data were aligned as much as possible. The acceleration data used for the analysis were obtained by sampling at a sampling frequency of 100 Hz and smoothed by low-pass processing with a cutoff frequency of 9 Hz.

B. Estimation of acceleration at the unmeasured part

First, the magnitude and phase of each frequency component are calculated by the Fourier analysis of the acceleration data of the head and the upper trunk, respectively. This result shows which frequency components are important in estimating the acceleration.

Next, the motion mode function for estimating the acceleration of the unmeasured part is obtained. In this paper, the head acceleration is estimated from the upper trunk acceleration, assuming the upper trunk as the measured part and the head as the unmeasured part. The motion mode function is derived by dividing the head by the upper trunk for each frequency component using the Fourier analysis results of the head and the upper trunk. Finally, the head acceleration is estimated by multiplying the obtained motion mode function by the Fourier analysis result of the upper trunk. In this paper, the motion mode function is obtained as an average motion mode function using 14 times of trial data, and the head acceleration is estimated from the remaining one trial data.

III. RESULT

Since similar results were obtained in two subjects, only the results for subject A are shown.

First, the acceleration data of the head and the upper trunk obtained from the experiment are shown. Figure 2 and Figure 3 show the vertical direction and the walking direction, respectively. The blue line is the head acceleration, and the orange line is the upper trunk acceleration. In the vertical direction shown in Figure 2, the acceleration waveform showed a similar trend. In the walking direction shown in Figure 3, the upper trunk acceleration was higher than the head acceleration. However, the relation between the movement of the head and the upper trunk in both the vertical and walking directions could not be quantified, and it is not possible to estimate the head acceleration from the upper trunk acceleration by correcting the constants.

Next, the acceleration data of the head and the upper trunk are decomposed into frequency components by a Fourier analysis, and the magnitude and phase are obtained for each direction. From the results, the important frequency bands for estimating the acceleration of the unmeasured part of the head were identified. The magnitude and phase in the vertical direction are shown in Figure 4 for the head and in Figure 5 for the upper trunk. The magnitude and phase in the walking direction are shown in Figure 6 for the head and in Figure 7 for the upper trunk. The magnitude and phase are shown as the mean and standard deviation calculated from 14 times of trial data.

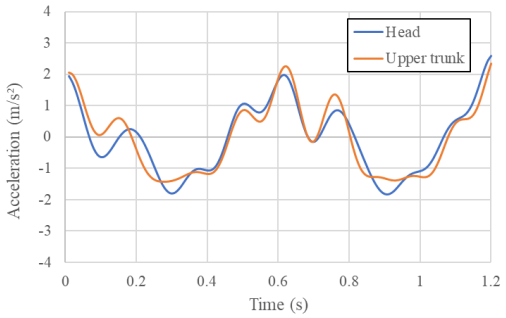

Figure 2. Acceleration data of the head and upper trunk in the vertical direction.

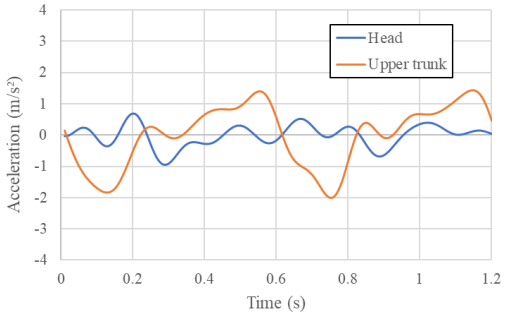

Figure 3. Acceleration data of the head and the upper trunk in the walking direction.

 ISBN: 978-1-61208-829-7

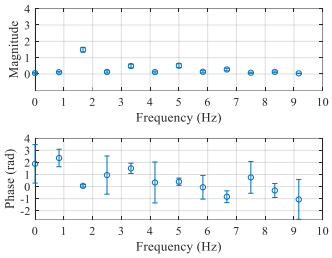

Figure 4. Result of frequency analysis of the head acceleration in the vertical direction.

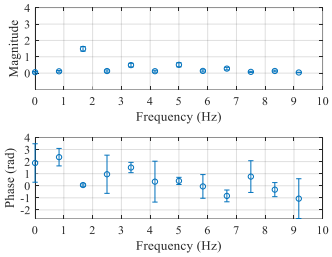

Figure 5. Result of frequency analysis of the upper trunk acceleration in the vertical direction.

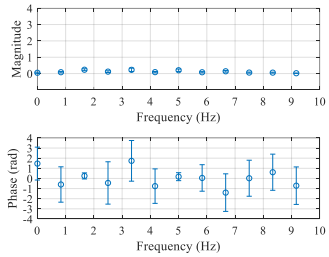

Figure 6. Result of frequency analysis of the head acceleration in the walking direction.

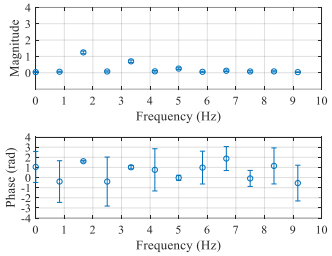

Figure 7. Result of frequency analysis of the upper trunk acceleration in the walking direction.

From Figure 4 to Figure 7, the gait frequency component (1.667Hz) and its integer multiple components have a magnitude in both the vertical direction and the walking direction. In particular, the magnitude of the gait frequency component is extremely large in the vertical direction. In the walking direction, there is a magnitude in the second-order component of the gait frequency in the upper trunk, and the magnitude in the head is small from the low-frequency bands to the high-frequency bands.

Hence, the movement in the vertical direction was found to be highly dependent on the gait frequency component. In the walking direction, the movement of the head was small, and the movement of the upper trunk was found to be significantly involved up to the second-order component of the gait frequency.

Next, the motion mode functions of the head and the upper trunk are obtained. The magnitude and phase difference of the calculated average motion mode function are shown in the vertical direction in Figure 8 and the walking direction in Figure 9. Gains and phase differences of the results are shown as means and standard deviations obtained from 14 times of trial data.

Next, the motion mode functions of the head and the upper trunk are obtained. In this paper, the motion mode function is defined as the average of the motion mode functions of 14 times of trial data. The magnitude and phase difference of the calculated motion mode function are shown in the vertical direction in Figure 8 and the walking direction in Figure 9. Gains and phase differences of the results are shown as means and standard deviations obtained from 14 times of trial data.

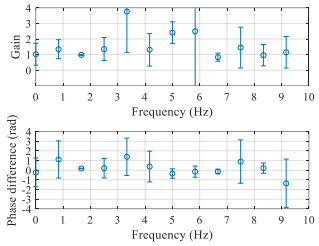

Figure 8. The average motion mode function in the vertical direction with the input as the upper trunk and the output as the head.

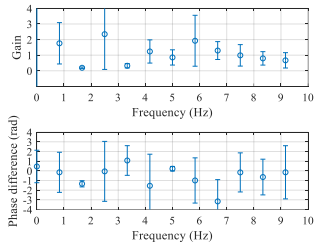

Figure 9. The average motion mode function in the walking direction with the input as the upper trunk and the output as the head.

In this paper, the relation between the head acceleration and the upper trunk acceleration, focusing on the gait frequency component and its second-order component indicated. Figure 8 shows that the second-order component has a large gain, while the most important gait frequency component has gain around 1 and phase difference around 0. From this, it can be seen that in the vertical direction, the head and the upper trunk have similar movements in the most important gait frequency component. From Figure 9, it was found that in the walking direction, the gain was small and the amplitude of the head was smaller than that of the

 ISBN: 978-1-61208-829-7

upper trunk in the low order components, which are main components.

Finally, using the motion mode function obtained from the 14 times of trial data shown in Figure 8 and Figure 9, the head acceleration was estimated from the measured upper trunk acceleration for the remaining one trial data. From the results in Figure 4 to Figure 7, it was found that the gait frequency (1.667Hz) and the second-order component are greatly involved in the movement. However, in this paper, we also focus on high-frequency components that are less involved in motion but have a magnitude. Therefore, the inverse Fourier transform is performed using a total of 5 points of the gait frequency component and its integer multiple components to convert them into a waveform in the time domain. The measured and estimated head accelerations are shown in Figure 10 in the vertical direction and Figure 11 in the traveling direction. The red line shows the estimated acceleration and the blue line shows the measured acceleration.

Since similar results were obtained in two subjects, only the results for subject A are shown. The estimation accuracy of the estimated acceleration compared to the measured acceleration was considered using the correlation coefficient, and the results are shown in TABLE I.

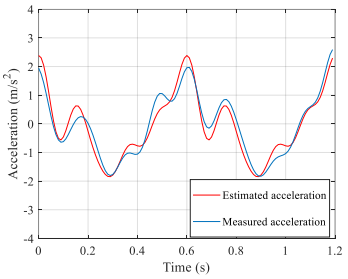

Figure 10. Comparison of head vertical acceleration estimated using the mean vertical motion mode function with measured values.

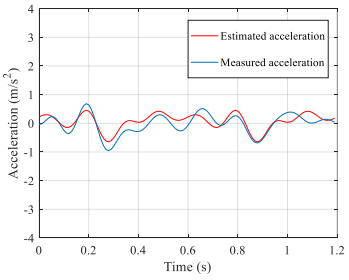

Figure 11. Comparison of head acceleration in the walking direction estimated using the mean vertical motion mode function with measured values.

TABLE I. CORRELATION COEFFICIENT BETWEEN ESTIMATED ACCELERATION AND MEASURED ACCELERATION

	Vertical direction	Walking direction
Correlation Coefficient	0.961	0.822

It can be seen from TABLE I that the correlation is strong in the vertical direction and that the estimation can be performed with high accuracy. In the walking direction, the correlation coefficient is smaller than that in the vertical direction, but the correlation is stronger. Consequently, it is found that the estimated the head acceleration derived from the average motion mode function shows the same tendency as the measured acceleration.

IV. CONCLUSION AND FUTURE WORK

In this paper, as a basic study to reduce the number of sensors used in estimating the floor reaction force, the acceleration at the unmeasured part was estimated using a motion mode function that describes the relation between the motion of each part. In this paper, the head acceleration is estimated from the upper trunk acceleration, with the upper trunk as the measured part and the head as the unmeasured part. As a result of Fourier analysis of the head acceleration and the upper trunk acceleration, it was found that the gait frequency and its integer multiple components have magnitude and are important for estimating the acceleration. Therefore, the inverse Fourier transform is performed using a total of 5 points of the gait frequency component and its integer multiple components to estimate the waveform of the head in the time domain. As a result, the correlation between the measured acceleration and the estimated acceleration at the head was high in both the vertical direction and the walking direction, and it was possible to make an accurate estimation.

In this paper, the acceleration of the head is estimated from the upper part of the trunk as an example to demonstrate the usefulness of the proposed method. In the future, by applying the proposed method to the entire body, the number of sensors used to accurately estimate the floor reaction force will be reduced. This will build a mobile system for healthcare wearable sensing.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP18K11106.

REFERENCES

[1] A. Isshiki, Y. Inoue, K. Shibata, and M. Sonobe, “Estimation of Floor reaction force during walking using physical Inertial force by Wireless motion sensor,” 19th International Conference on Human-Computer Interaction, HCI International 2017, Communications in Computer and Information Science, vol. 714, pp. 249-254, doi:10.1007/978-3-319-58753-0_37, 2017, pp.22-33, ISSN:1348-7116.

[2] M. Ae, H. Tang, and T. Yokoi, ”Estimation of Inertia properties of the Body Segments in Japanese Athletes,” Soc. Biomechanisms Jpn., vol. 11, 1992, pp. 22-33 (in Japanese).

[3] S. Hontama, Y. Inoue, and K. Shibata,” Characteristics of walking motion by using Frequency analysis: Transfer function for Upper body,” The Japan Society of Mechanical Engineers Chugoku-Shikoku Bracch , The 50th Student Graduation Research Presentation Lecture, No.06b3, 2020 (in Japanese).

 ISBN: 978-1-61208-829-7

Interactive Wiki for Special-Purpose Machines

Thomas Herpich

Institute of Information Systems
Hof University of Applied Sciences
95028 Hof, Germany
Email: thomas.herpich.2@iisys.de

Valentin Plenk

Institute of Information Systems
Hof University of Applied Sciences
95028 Hof, Germany
Email: valentin.plenk@iisys.de

***Abstract*—Machines in processing plants are frequently generating failures that must be manually fixed by the operators. Machines often lack advanced assistance systems for to address these failure cases. While some new developments try to use only machine data, in many applications, the human knowledge of the operators can be very useful. In this paper, we propose a new assistance system used to merge machine data with the operator's knowledge. This system is tested with an industry partner. The test results are used to create design considerations, compare different reasoning algorithms, and check the influence on the machine downtime.**

Keywords–Human Machine Interfaces; User Assistance Systems; Special-purpose Machines.

I. Introduction

Digitally controlled machines and systems in production are becoming increasingly intelligent. As part of in-process control, they continuously check their own status to control the production process and monitor compliance with the specifications. Maintenance and repair requirements can also be predicted in reasonable time so that consistent quality in the production process is permanently ensured. Failure times are reduced accordingly, relieving the operating personnel of numerous routinal tasks during the operation of the machines.

Despite the increasing intelligence of these computer-controlled machines, problems still occur in production processes which cause the systems to come to a standstill. In these cases, the problems are often more complex, for example because a plant consists of several components from different manufacturers, whose interaction is not regulated in any manufacturer's user manual. Despite digital technologies in the control of each individual component of the plant, manual intervention by the operator is necessary in such cases. The operator must then be able to identify the reasons for the malfunction, initiate suitable measures and put the system back into operation. They are often assisted only by the user manual, or an assistance system that is different for each manufacturer and machine [1], [2].

In [3], Oehm *et al.* created and evaluated a design for an assistance system for processing plants. In this paper, we propose an implementation for that design which was slightly adjusted for special-purpose machines used by our project partner. In [4], we described an approach to automatically extract the knowledge only from available machine data without integration of the operators. In [5], we showed that this approach is not always feasible, because the data only represents a small part of the work that must be done. To accommodate this problem, the machine operators can document information about the current situation and the appropriate steps to fix the failure in the new software. This information is then mapped with the current state and should be automatically retrieved when this situation occurs the next time. Available knowledge can also be reused by the operators.

With this approach, the software should be able to provide useful information automatically after a training phase. This should reduce the downtime, especially with inexperienced workers or infrequent failure causes. The software learns automatically, which data can be used to distinguish different failure cases and is machine independent, if the generated data can be made available to the system.

The new system is called "IISYS Machine Wiki".

In Section II, we describe the structure of all involved systems, Section III gives the description of the new assistance system. This work ends with a conclusion and future work.

II. System Description

This section gives an overview of the involved systems. Figure 1 shows the structure of the machine we are using and the new assistance system.

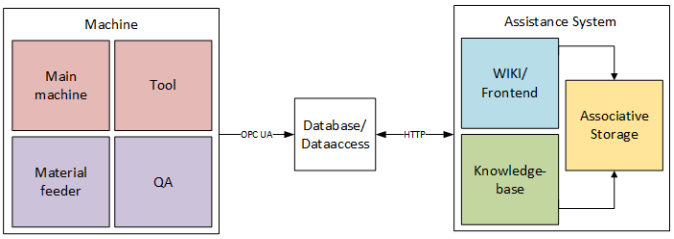

Figure 1. System structure.

a) Machine: The machine (left part of the figure) is a stamping press (Main Machine). It uses a changeable tool for the work. There are also some peripheral machines like a material feeder, an automated Quality Assurance (QA), and the commission. All machines are generating data consisting of setpoints, sensor values and machine states. This data is collected via Open Platform Communications Unified Architecture (OPC UA) and then saved in the time series database influxdb (center in figure). For one machine, there are around 7500 variables available, where some are not used. More details are provided in our previous work [5]. From there, all additional software can access all data generated since the data collection started (currently around two years).

 ISBN: 978-1-61208-829-7

b) Assistance System: The newly created assistance system (right part of the figure) consists of the following parts:

- Frontend: This is the user interface that is used by the machine operators and the foreman.
  It is created as a Web application with JavaEE, so it can be used with any computer and from any machine in the factory.
- Knowledgebase: The knowledgebase stores all information gained from the user. It also maps it with the corresponding failure. The knowledgebase is then used to train a model to find the correct information for any occurring situation.
- Associative Storage: This storage maps error states to machine data and is used to find the appropriate wiki page for any given situation.

III. Description of the Assistance System

This section describes the new system and its features.

A. Algorithm

This section describes the algorithm used in the new system. Figure 2 shows all steps done by the system. Symbols in blue use the user interface, green interact with the knowledgebase and yellow process the information from the knowledgebase.

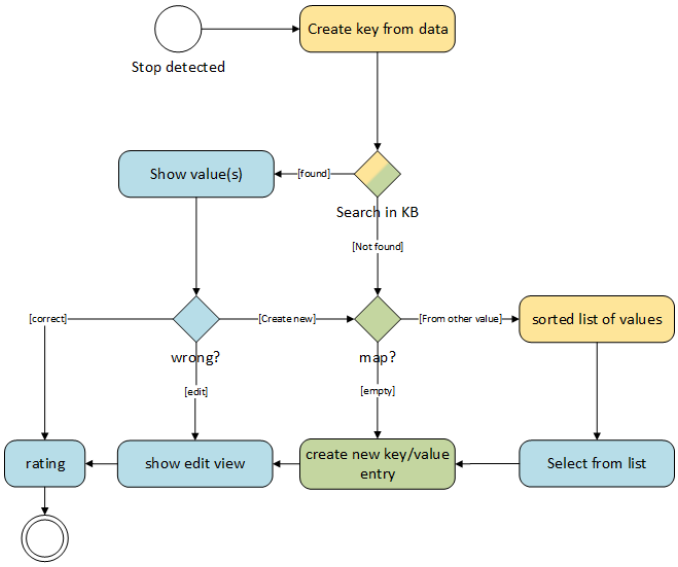

Figure 2. Activity diagram new system.

The software continuously checks incoming machine data for the stop condition, as described in Section III-A1. When a stop is found, the following sequence starts:

- Build the key representing the current machine state from the available data, as described in Section III-A1.
- Search for the key in the knowledgebase. This is described in Section III-A2.
- If the key is found, the corresponding value (wiki page) is shown to the operator.
- The operator can request to change the shown wiki page, see Section III-A4.
- If the operator proposes another page or no page could be found in the knowledgebase, a new mapping must be created, see Section III-A4. This new mapping can be filled manually with information (case "[empty]") or linked to a value from another key.
- To create this link, a list with possible values is shown. This list is sorted by the knowledgebase key, where the most equal key will be shown first. Section III-A2 describes how this works.
- When this mapping is created, the operator can edit the new page.
- At the end, the operator can rate the proposed sequence with 0 to 5 stars.

1) Stop Detection: The new system must detect if the machine is in an error state. This can be accomplished by analyzing new incoming data from the machine. This data contains one main error code (OPCGeneralInterface.ErrorMessage). This code is currently used by the machine to display an error text. While many error causes within the main machine can be distinguished by this code, every failure from the peripheral devices is only reported with one equal error code. The code contains the number 0 while the machine is running and a positive integer when in error state. The "ErrorMessage" can change during the stop while the operator fixes the error and can be used to display additional help.

An additional variable (OPCGeneralInterface.State) contains the operation mode, where 4 indicates normal operation, 0/1/2 a stopped machine, 3 manual set up and 5 an error.

The time span from error begin until the return to normal operation is called stop.

The machine sends the same main error code for every failure caused by peripheral devices, so it is not possible to distinguish all states. However, the machine data contains additional fields that can be used:

- Around 250 bit-fields (Table "Ems_db"). These bits correspond to different errors and states from the peripheral machines.
- 15 - 20 float values containing measurements like temperatures or the pressforces and some changing settings.

The distinct combinations of those fields would lead to around $2^{250} * errorcodes$ machine states. The algorithm used in this software must automatically decide which fields and corresponding values should be used to distinguish between different states.

2) Knowledgebase: The knowledgebase maps machine state information to the wiki pages that should be proposed. This information is used to find wiki pages that can be proposed to an operator in case of a machine failure.

Different methods can be used to compare the machine data of different situations and to distinguish between them. In [6], the authors use a distance-based approach where one number (called distance) is calculated from selected machine variables which is then used to compare the data. A drawback of this method is that the variables to be used must be known to generate a useful result.

An approach that decides automatically which variables to use and that generates decision rules are decision trees. A random forest uses multiple decision trees combined as a voting

 ISBN: 978-1-61208-829-7

system. These trees are trained using different combinations of the available fields. It can also leave out variables that are not helpful in distinguishing the different situations.

Any of these approaches need a training set in form of past situations mapped to the correct wiki pages. This training set will be generated during the first months of use. Until this is available, the software only uses the main error codes and the operator needs to select which proposed page is the best for the current situation.

3) Show Wiki Pages: When a machine failure is detected, our system uses the current machine data to search for appropriate wiki pages. In the best case, that is exactly one page, but if the software cannot distinguish some cases, more than one page is proposed to the operator. These pages are displayed in tabs as described in Section III-B. These tabs are sorted with the first tab being the best matching page. The operator can select the best page for the current case or create a new one. A similarly sorted list is used if the operator searches for wiki pages that can be mapped to the current situation, as used in Section III-A4. When using a random forest, the best entry is the most proposed entry from the forest. If the order for some entries cannot be determined clearly, the system uses the ratings created for this page in previous situations. An additional factor is how often the page was used and displayed in previous cases.

4) Create and Use Wiki Pages: The operator has different possibilities to change wiki pages and map them to the current situation. This section describes these possibilities.

a) Create a New Page: When the algorithm could not find any wiki page to display, the operator must create a new page, which is mapped with the current situation with a new entry in the knowledgebase. This is also available if there are pages available, but no page suits the current needs.

b) Create Page from Template: The operator can create a new page based on another page. The content of the existing page is copied to the new page, which can be changed by the operator. These changes will not apply to the base page as opposed to the operation in the next paragraph.

c) Associate an Existing Page with an Error: With this action the operator can map the current situation to an existing wiki page which was not proposed. This creates a new mapping in the knowledgebase. When this page is edited, the changes will be shown in all situations that are mapped to this page.

d) Edit Current Page: This action does not change any mappings in the knowledgebase. If the displayed page is mostly correct, but there are some misleading or missing information, the operator can edit this wiki page. The change applies to all machine states that are mapped to the same page.

e) Replace a Mapping: This action removes the current mapping with the situation to the selected page and then starts the action “Associate an existing page to this error”.

B. User Interface Design

The design for the UI was worked out with the project partner in [7]. The software is designed according to user experience guidelines given by [1], [8]–[10].

The wiki system will be displayed on a screen next to the machine with a keyboard and mouse attached. This setup is already available and used for other tasks by the operators, so there is no new hardware needed.

Figure 3 shows the main page of the application.

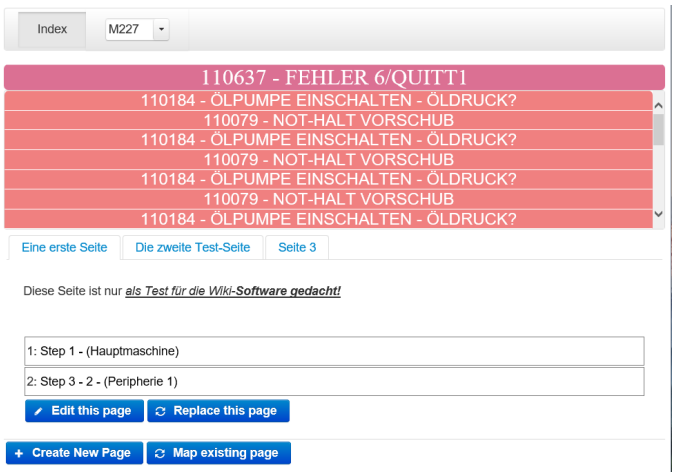

Figure 3. Screenshot main page.

The operator can see the current error that caused the failure in dark red. Below is a list of additional occurring error messages while the error is fixed, as suggested by the operators. If the machine is running with no error, the background changes to green.

All possible wiki pages that are found for the current situation are displayed as tabs below the error code. The user can select the most useful page. From within these tabs, the tasks described in III-A4 can be executed:

- “Edit this page”
  Edit the selected page. This change must be approved by a foreman before it will be shown in the main page.
- “Replace this page”
  This action removes the selected page as possibility for the current situation. After this action, the operator can select another page that suites the current failure.
- “Create New Page”
  The operator can create a new wiki page. This page is automatically mapped to the current situation.
- “Map existing page”
  The operator can select an existing wiki page that suites the current failure. This selected page will then be mapped to the current situation. As opposed to “Replace this page”, the mapping of other pages with this situation will not be removed.

The tasks “Edit this page” and “Replace this page” are only available if the algorithm has found at least one wiki page to show for the current situation. If no pages are available, the tab list remains empty and the operator can create a new page or map an existing one.

When the error is fixed, the operator has the possibility to rate the proposed page with 0 to 5 stars. This rating is then used to improve the algorithm as described in Sections III-A2 and III-A3.

Every wiki page consists of a title and an error description which can be entered through a rich text editor to enable some formatting as well as the insertion of images. The error description should contain a detailed description of the problem as well as some information how the operator identified the exact

 ISBN: 978-1-61208-829-7

problem. The operator can enter the steps that are necessary to fix the failure. These steps are structured in different categories (e.g., Main Machine, Periphery QA) where both can only be created by a foreman. This is to ensure that every operator uses the same wording for the same steps. Figure 4 shows the UI to edit the step list with the page title and error description on the top. The step list can be reordered by the buttons on the right.

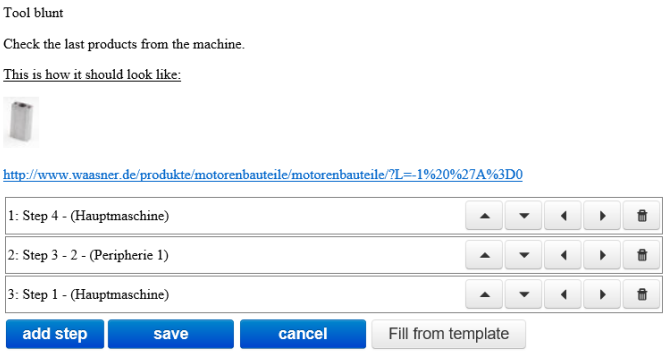

Figure 4. Screenshot create step list.

The list can be expanded by clicking the button “add step” which opens the UI shown in Figure 5.

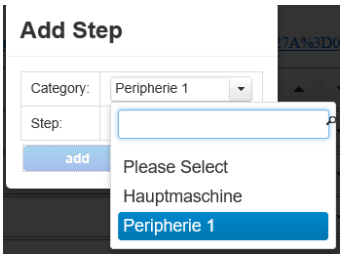

Figure 5. Screenshot add step to list.

The steps are grouped by category, so the user must select a category first and then a step from that category. Every important input field contains a search box and predefined values to simplify the usage.

Admin users have an overview of past stops that can be used to create page mappings for past situation to improve the algorithm.

IV. Conclusion and Future Work

A. Assistance System

With the created assistance system, the operators can save and share their knowledge with other users. This knowledge is stored in a wiki-based structure. Every wiki page can be freely mapped to any situation, and every mapping can be removed.

To improve the (perceived) usefulness of the software, the users from the project partner were integrated into the development, as suggested by [8] and [9]. They are integrated through multiple meetings and a test phase where all users will be able to give additional feedback.

The roll out began in June 2020 for the first machine and is scheduled to be expanded to other machines by August 2020.

B. Test machine

While we are waiting for data from our project partner, we created a simple test machine with similarities to the real machines. Our machine picks up either a magnetic or plastic chip, then moves the transport arm in different directions and releases the chip at a predefined location. The machine generates 64 bits representing different problems and sensors in the machine but has no common error code that identifies a problem cause. Then we will simulate different runs, some of them will fail with different causes. Many of these causes can be fixed with the same procedure. The new software is used to create wiki pages for these causes and map them to the runs. Different causes will then be simulated again, and the software should propose the correct wiki page as mapped before

C. Future Work

In the next months, the system will be running 24/7 in full production on two machines.

The data generated by the project partner and from our test machine can then be used to evaluate:

a) Reasoning Algorithm: The proposals of the new system can be compared to proposals only generated with the main error number. The software should at least be as useful as using only the main error code. This will be measured by comparing the needed time to fix the errors and the rating operators can give in the system. The amount of remappings or new page creation is also a good measure. If many situations have multiple wiki pages mapped to them this would indicate that cases cannot be distinguished past the main error code. The results can additionally be compared to other reasoning algorithms like distance-based algorithms which can be used with the same knowledgebase.

b) User Experience: The operators will be questioned with standardized user experience questionnaires, e.g., the User Experience Questionnaire (short version) [11], on how satisfied they are with the selected design. With this information, the software will be changed to improve the acceptance of the new system or to decrease the time needed to fix failures. This feedback could also be used to create some guidelines for assistance systems.

c) Downtime: An advanced assistance system should decrease the downtime caused by fixing machine failures. To verify this, a baseline must be created from the current production. This can be done with the machine data available for the last two years. That baseline can then be compared to new data generated after the test phase 1.

d) Match Steps with Data: Some of the steps entered by the operators can be mapped with changes in the machine data. This mapping could enable the system to ensure that a step is carried out in the correct way by checking the machine data. It could also allow the system to generate a basic wiki page by recording the changes in data while an operator fixes a failure and searching the corresponding steps mapped to these changes. The basic page can then be edited by the operator to insert the missing steps.

 ISBN: 978-1-61208-829-7

REFERENCES

[1] S. Delisle and B. Moulin, "User interfaces and help systems: From helplessness to intelligent assistance," *Artificial Intelligence Review*, vol. 18, no. 2, pp. 117–157, Jun. 2002, ISSN: 1573-7462. DOI: 10.1023/A:1015179704819. [Online]. Available: https://doi.org/10.1023/A:1015179704819.

[2] A. Maedche, S. Morana, S. Schacht, D. Werth, and J. Krumeich, "Advanced user assistance systems," *Business & Information Systems Engineering*, vol. 58, Aug. 2016. DOI: 10.1007/s12599-016-0444-2.

[3] L. Oehm *et al.*, "Cooperative fault diagnosis by operator and assistance system for processing plants," in *Technische Unterstützungssysteme, die die Menschen wirklich wollen (Band zur zweiten transdisziplinären Konferenz 2016)*, 2016, pp. 375–384, ISBN: 978-3-86818-089-3 (Print) bzw. 978-3-86818-090-9 (Online).

[4] V. Plenk, "Improving special purpose machine user-interfaces by machine-learning algorithms," *Proceedings of CENTRIC 2016*, pp. 24–28, 2016.

[5] T. Herpich, "Commissioning and evaluation of a software system consisting of c#- and java applications for real time retrieving and saving as well as analyzing machine events from a opc server," Bachelor-Thesis, Hof University - Institute of Information Systems (iisys), Mar. 7, 2019.

[6] V. Plenk, S. Lang, and F. Wogenstein, "An approach to provide user guidance in special purpose machines and its evaluation," *International Journal On Advances in Software*, vol. 10, no. 3, pp. 167–179, 2017.

[7] T. Herpich, "Designing an interactive wiki assistance system for special purpose machines," Hof University - Institute of Information Systems (iisys), Feb. 25, 2020.

[8] M. Schelkle and C. Grund, "Identifying design features to increase the acceptance of user assistance systems: Findings from a business information visualization context," in *13th International Conference on Design Science Research in Information Systems and Technology (DESRIST 2018)*, Jun. 2018.

[9] C. Aringer-Walch, S. Besserer, and B. Pokorni, "User needs for a digital assistance system in the context of industry 4.0. an explorative study in the area of assembly," in *Technische Unterstützungssysteme, die die Menschen wirklich wollen (Band zur dritten transdisziplinären Konferenz 2018)*, Dec. 2018, pp. 139–150, ISBN: 978-3-86818-245-3 (Print) bzw. 978-3-86818-246-0 (Online).

[10] M. Funk, M. Hartwig, N. Backhaus, M. Knittel, and J. Deuse, "User evaluation of assistance systems for industrial assembly," in *Technische Unterstützungssysteme, die die Menschen wirklich wollen (Band zur dritten transdisziplinären Konferenz 2018)*, Dec. 2018, pp. 213–221, ISBN: 978-3-86818-245-3 (Print) bzw. 978-3-86818-246-0 (Online).

[11] D. M. Schrepp. (2017). "Short version of the user experience questionnaire," [Online]. Available: https://www.ueq-online.org/Material/UEQS_Items.pdf (visited on 01/21/2020).

 ISBN: 978-1-61208-829-7

Impact of Advertising Intensity on Customer Churn for Web-Mail Services: Insights from a Customer Survey in Germany

Jasmin Ebert and Stephan Böhm

CAEBUS Center of Advanced E-Business Studies
RheinMain University of Applied Sciences
Wiesbaden, Germany
e-mail: jasebert@web.de,
stephan.boehm@hs-rm.de

Christian Jäger and Frank Rudolf

Deutsche Telekom AG
Darmstadt,
Germany
e-mail: christian-jaeger@telekom.de,
rudolff@telekom.de

***Abstract*—Many services on the Internet are offered free of charge to users. These include web-mail services, which allow access to e-mails via the browser without the installation of an e-mail client. Companies offer free web-mail services, for example, as a complementary service to a paid service or as an introductory or try-out offer. Advertisements are often placed on the portals of web-mail services as a revenue model or to help cover costs. Advertisements may not only contain interesting advertising messages for users, but may also be perceived as annoying depending on the content and extent of the advertising. Too much advertising can lead to a churn of users. Providers, therefore, find themselves in an area of conflict between pushing advertising to increase advertising revenues and limiting advertising to prevent customer churn. This study examines the impact of advertising intensity and the change intention of web-mail users. The study was conducted among the customers of the Telekom E-Mail Center, one of the popular web-mail offerings in Germany. A total of 2,228 customers were surveyed, and the significance of the reasons for switching was evaluated by means of discriminant analysis. After privacy concerns, the most important reason for changing web-mail providers was found to be too high advertising intensity.**

Keywords–web-mail services; customer churn; advertising intensity; discriminant analysis.

I. INTRODUCTION

Advertising is of great importance to Internet service providers, as their business models often depend on this revenue stream. Online advertising can also contain valuable information about products and services for Internet users. Currently, there are more and more sophisticated technologies supporting the process of selecting advertising messages relevant to users and deliver them with the lowest possible dispersion losses. However, the placement of advertisements may be perceived as annoying or undesirable by users [1]. Internet service providers are thus in a constant balancing act between realizing advertising revenue potentials and maintaining customer satisfaction. Existing studies on e-mail marketing and online advertising [2] focus, for example, on the opportunities for Internet service providers to increase awareness of their service portfolios, to attract attention, and to arouse buying interest among potential customers. Moreover, there are plenty of approaches to measuring customer satisfaction [3]. However, to the knowledge of the authors, there is a research gap on investigating the negative effects of online advertising on customer satisfaction as well as analyzing customer churn as a result of excessive use of advertising by Internet service providers. Additionally, while existing studies analyzed the effects of online advertising from the perspective of Internet service providers or customers (e.g. [4][5]), there is a lack of studies that try to combine both perspectives.

The main research objective of our paper is, therefore, the analysis of the effects of advertising on the web-mail portal of Deutsche Telekom on customer churn. Theoretical principles from the field of operationalizing and measuring customer satisfaction (e.g., confirmation-disconfirmation paradigm, [6]) were included in the design of the customer study. In addition to the users' perception of advertising, the study investigated influences on customer satisfaction and the intention to churn. Moreover, various advertising formats and contents (e.g., personalization of advertising content) and the parallel use of e-mail service offerings from competitors have been considered in the study. The online survey focused on the usage of the e-mail portal on the desktop browser and was conducted in May and June 2019. Customers were randomly selected among the visitors of the web-mail portal during a twelve-day survey period. Since this is still a preliminary study, the analysis is mainly exploratory, and no explicit research model, e.g., for validating cause-effect relationships, has been formulated. A discriminant analysis to identify distinguishing features and the significance of advertising intensity in the groups of users with and without the intention to churn was carried out.

Against this background, our study is structured as follows: In Section II, we first discuss the use of web-mail in Germany and describe important advertising content and formats. In the following, related research on the impact of advertising on the use of online services is examined and the research objectives of the paper are presented. Section III then covers the methodology and approach of the customer study. Important results and implications are presented in Section IV before this paper closes with the conclusions in Section VI as well as limitations and an outlook on further research in Section VII.

II. RESEARCH BACKGROUND

A. Web-mail Services and Usage in Germany

The number of users in Germany who use the Internet to send and receive e-mails has risen sharply in recent years

 ISBN: 978-1-61208-829-7

from 38 percent in 2002 to 86 percent in 2019 [7]. Thirty-five percent of the customers surveyed also use an additional e-mail service in parallel. According to this, there is no monopoly among German e-mail providers, which means that customer loyalty is becoming increasingly important.

This industry study was conducted among customers of the web-mail offering of Deutsche Telekom. Deutsche Telekom emerged from the former state-owned national telecommunications network operator in Germany and is now a leading European telecommunications company with headquarters in Germany. With around 184 million mobile customers, 27.5 million fixed-network lines, and 21 million broadband lines, the Deutsche Telekom Group is one of the world's leading integrated telecommunications companies [8]. The company launched the first mail service for the German mass market under the T-Online brand in the summer 1995. As one of the most used e-mail providers in Germany, more than 2.5 billion e-mails were received via T-Online e-mail addresses in 2019 daily [9].

B. Advertising Content and Formats

In 2017, the Internet replaced classic television as the world's most popular advertising medium and this growth has continued unabated ever since: [10], for example, forecasts that the share of Internet advertising in total global spending will rise from 39 percent in 2017 to 49 percent by 2021. Global spending on online advertising is expected to increase from USD 273 billion in 2018 to USD 427 billion by 2022 [11]. In terms of online advertising formats, the biggest growth is expected to be in display advertising (banners, online videos), which is primarily due to high-quality content, better screens and Internet connections, but also to the creative and personalized approach to target groups [4] thanks to "programmatic buying" [10]. Programmatic buying or programmatic advertising refers to "... the automated purchase and sale of advertising space" [12]. In the USA, where digital advertising generates the highest revenues worldwide, programmatic advertising is predicted to grow from 73 percent in 2017 to 78 percent by 2023 [13]. Also, since 2017, global spending on mobile advertising has been 109 billion USD, exceeding desktop advertising spending of 104 billion USD [10]. In Germany, the picture so far is still reversed: In 2018, gross advertising expenditure totaled nearly EUR 12.6 billion, with desktop advertising accounting for EUR 0.97 billion and mobile advertising for EUR 0.36 billion [14]. Nevertheless, the global development of mobile advertising media is also becoming increasingly visible in Germany: If we look at the growth of mobile advertising in 2019 (from January to May) compared to 2018, an increase of 26 percent is apparent [14]. At the same time, gross expenditure on desktop advertising in 2019 rose by only two percent compared to 2018. As a result, the focus of online advertising concepts should be increasingly directed towards mobile in the future, as the use of mobile devices and thus mobile access to Internet offerings will increase significantly [15]. Table I lists the common advertising formats within web-mail services.

TABLE I. OVERVIEW OF IMPORTANT WEB-ADVERTISING FORMATS

Format	*Description*
Skyscraper	Vertical ad, static, animated or rotating
Text Link	Single-line ad text incl. hyperlink
Inbox Ad	Ad which is integrated into the message list
Transmission Confirmation	Pop-up appearing after an e-mail has been sent

III. RELATED WORK AND RESEARCH OBJECTIVES

A. Related Work on Customer Churn at Online Services

The telecommunications industry, in particular, is interested in predicting customer churn [16] because the telecommunications sector is a rapidly growing and highly competitive market [17] and it has a direct impact on the competitiveness of a service provider [18]. Consequently, telecommunications companies often use customer churn as an important KPI to make forecasts [19]. The high intensity of competition makes it more difficult for telecommunications companies to bind customers to their services in the long term, as it is easy for customers to switch between providers [17].

B. Related Work on User Impact of Advertising

In the past, the two research areas e-mail marketing and online advertising have already been discussed in detail [1][2][5]. The focus was, for example, on the opportunities offered to Internet service providers to increase the awareness of their service portfolio through external communication in order to attract attention and arouse the buying interest of potential customers [1][2]. However, this approach neglects when the opposite could be achieved and customers churn due to excessive advertising intensity. In addition, the personalization of advertising has been intensively considered. According to [5], for example, the intrusiveness of advertising increases when it is personalized by the name of the recipient. Furthermore, there are also data protection concerns in connection with personalized advertising. In addition, penetrating Internet advertising has a negative impact on customers' purchase intentions, even if it includes discounts. In contrast, personalized online advertising can still be successful, depending on the industry of the provider. As far as the telecommunications industry is concerned, however, no correlation between personalized advertising and revenue growth could be established in the past [5].

The situation is similar to the research topic of customer satisfaction. With regard to the offline presence of companies, it has already been shown that the design of salesrooms has an influence on customer satisfaction [4]. Even just seeing the landing page of an e-commerce store can trigger emotions in users that influence their behavior [20]. Although various models and procedures for measuring customer satisfaction already exist in the literature [3], they are essentially only concerned with the evaluation of a product or service, customer service, or the company in general. This fact justifies the aim of the following study to gain insights into the influence of online advertising on customer satisfaction.

In summary, this results in two central research gaps which are closed by this scientific work: On the one hand, within the research fields of e-mail marketing and online advertising, the customer perspective has so far receded into the background, especially with regard to the extent to which the use of advertising has negative consequences. On the other hand, there is insufficient research knowledge about customer satisfaction with regard to the advertising financing necessary from the provider perspective.

 ISBN: 978-1-61208-829-7

C. Research Objectives of the Customer Survey

The aim of the analysis was to find out how strongly online advertising within web-mail services influences customer satisfaction and can even harm the providers in the form of customer churn in the long term. For this purpose, the perception and impact of different online advertising formats within the Telekom E-Mail Center were measured. Specifically, the current use of advertising in the browser application of the Telekom e-mail service was investigated. The research questions were:

- How can advertising within the e-mail portal be used sensibly without causing negative consequences on customer satisfaction, e.g., customer churn?
- How are different advertising formats evaluated from the customer's perspective?
- Does the acceptance of personalized and non-personalized advertising differ?
- What is the maximum advertising intensity that can be expected of users?
- How is the current advertising volume within the e-mail portal perceived in comparison to competing web mail services?

IV. Methodology and Study Approach

In order to find out to what extent the respondents would change their main provider, i.e., not only the Telekom E-Mail Center, due to excessive advertising intensity, a discriminant analysis was applied. The aim of this method was to find out whether online advertising has a "discriminatory significance" [21] with regard to the customer churn rate; in other words, to what extent advertising acts as a disruptive factor so that customers would consider changing providers as a result. According to the method, customers were initially divided into two groups, which are distinguished by a no-minus characteristic [21]. The "discrimination criterion" represents the customers' willingness to switch. This results in two groups: the churners and the non-churners.

The so-called churners are characterized by the fact that they consider changing their main provider within the next six months. In contrast, the non-churners estimated their willingness to change the provider within the next six months as unlikely. In a broader sense, the group of churners can be considered more dissatisfied than the group of non-churners. Of course, there are also customers who are not willing to change despite their dissatisfaction. Reasons for this can be the convenience, barriers to change, or loyalty [22]. The primary initial question of the discriminant analysis on the topic of willingness to switch was followed by ordinally scaled questions in order to determine which motives are most likely to be behind a possible change of provider by the users. Accordingly, the group of churners was asked to indicate those motives for a probable change. In contrast, the group of non-churners was asked from which motive they would change if this were hypothetically the case, contrary to their previous answer on the probability of switching within the next six months. Possible reasons for the change are given as possible answers to which the respondents could individually agree using the ordinal scale. The following reasons for switching were available [23][24]: Recommendation of a friend or acquaintance, too much advertising, too few functional and configuration options, poor usability, data protection concerns, slow update speed or data transfer, as well as the desire for change/new things to try out.

The questionnaire was played out randomly at a frequency of 1/400 per login to the browser-based Telekom E-Mail Center. In order to consider all usage rituals, all weekdays, including weekends, were considered with regard to the duration of the survey. Specifically, the survey was put online for around 12 days between Friday, May 31, and Tuesday, June 11, 2019. The test persons came from the existing customer pool of web-mail service of Deutsche Telekom, which means that the population was made up of real customers. According to a previous study, there was also an overlap between the users of the Telekom E-Mail Center and alternative providers: 44 percent of them also use a competitor's web-mail offering. In general, e-mail services are a highly competitive industry [25]. Therefore, a comparison of competitors was implemented in the questionnaire.

V. Study Findings and Implications

In total, 2,228 users took part in the survey. Table II shows some selected characteristics of the sample. The average customer who completed the online questionnaire is male, between 61 and 65 years old, who lives in a two-person household and has a monthly net income between 2,500 and 2,999 euros. The sample size shows a possible bias as the results are not representative for the total population.

TABLE II. Overview of Sample Characteristics

Sample Characteristics	*Percentage*
Male	74.0%
Female	21.0%
71 years or older	27.0%
Two-person household	52.0%
Monthly net income between 2,500-2,999 euro	9.0%

Table III shows that a majority of the respondents reject more ads in return for more functionality within the web-mail offering. Moreover, the analysis showed that 84 percent generally reject an increase in online advertising, regardless of whether it is in line with their interests (personalized content) or not. This anti-attitude of users towards the intensification of advertising makes the starting position more difficult, both for providers of web-mail services and for advertisers, since both parties are dependent on advertising or indirectly on advertising revenue or click rates. Even a functional enhancement, such as increased storage space or spam protection in return for a higher advertising volume would only be accepted by twelve percent of the respondents. As far as this practical example is concerned, it is becoming apparent that many of the customers surveyed also use the products of alternative web-mail providers.

As mentioned above, there is strong competition between web-mail offerings in Germany. Against this background, it can be assumed that the perceived advertising intensity also has an influence on customer loyalty compared to competitors. Nevertheless, there is a large discrepancy between the subjective customer opinions regarding the perception of advertising intensity. This inconsistency in the customer perspective is expressed in the fact that there is a disagreement between the researched telecommunications company Deutsche Telekom

 ISBN: 978-1-61208-829-7

TABLE III. ADVERTISING PREFERENCES OF THE STUDY PARTICIPANTS

Advertising Preferences	*Percentage*
No acceptance for more ads in return for more functionality	87.0%
Generally, not more ads, even if personalized	84.0%
Static instead of animated advertising	83.0%
Acceptance of a higher ad volume in return for unlimited storage space	73.0%
Acceptance of a higher advertising volume in return for more spam protection	71.0%
No personalized ads for data protection reasons	58.0%

and competing providers in the assessment of the extent of advertising. According to the survey, 41 percent of the respondents who actually use the services of other web-mail providers were convinced that there are differences in the intensity of advertising between web-mail providers.

Paradoxically, at the same time, 32 percent felt that the advertising volume at the Deutsche Telekom portal in contrast to other web-mail services they use is more or less the same. This result changed insignificantly if one looks specifically at the relationship between Deutsche Telekom and its largest competitor with regard to differences in the perception of advertising volume. In addition, it is not possible to generalize as to whether personalized or non-personalized advertising, in general, scored better in terms of usefulness. However, customers in Germany are particularly sensitive when their privacy is invaded in order to personalize advertising. Fifty-eight percent of the users expressed data protection concerns about personalized advertising content. In terms of advertising format, especially with regard to the examined Telekom E-Mail Center, simple text links, which redirect the user to the advertiser via hyperlink if they are interested, performed best. Only 23 percent found them very annoying. At 37 percent, respondents felt slightly more disturbed by advertisements in the form of skyscrapers. The greatest disruption was attributed to advertising banners integrated into the transmission confirmation (55 percent), closely followed by advertisements that are located within the e-mail list of a mailbox and appear in the same design as regular e-mails (52 percent).

Nonetheless, critical customer opinions about online advertising were not necessarily reflected in their actual behavior, as shown in Table IV. Forty-two percent of the respondents would just reduce usage, while 35 percent would simply ignore annoying advertising without resorting to further measures, such as complaints or changing providers. Only 20 percent of the respondents considered changing the provider in case of too much advertising.

TABLE IV. IMPACT OF WEB-MAIL ADVERTISING ON WEB-MAIL USERS

Impacts	*Percentage*
Reduction of usage in case of too many ads	42.0%
Accept the higher ad volume without a reaction	35.0%
Change of provider in case of too many ads	20.0%

Table V shows that only seven percent of customers intend to change their main web-mail provider within the next six months. For 39 percent of the customers, too much advertising was the reason, while 23 percent had performance problems with the portal currently in use.

TABLE V. INTENTION TO SWITCH AND CHURN REASONS

Intention to Switch and Churn Reasons	*Percentage*
Probability of changing the main provider within the next six months	7.0%
Too many ads as a reason for changing the main provider	39.0%
Slow update speed/data transfer as a reason for the change	23.0%

In a further analysis step, the reasons for a change of provider were examined in the group of respondents who had indicated that they would change provider within the next six months (churners). Table VI shows the ranking of the most frequent motives for switching for the web-mail provider mainly used as a result of the discriminant analysis conducted. The motive of too much advertising (57 percent) was in second place after the leading data protection concerns (74 percent).

TABLE VI. RESULTS OF THE DISCRIMINANT ANALYSIS

Rank	*Discriminants*	*Discriminant Value*
1	Privacy concerns	74.3%
2	Too much advertising	57.3%
3	Poor usability	54.6%
4	Recommendation of a friend/acquaintance	12.7%
5	A desire for change/new things to try	9.7%
6	Too few functional and configuration options	5.5%
7	Update speed/data transfer too slow	2.0%

In contrast, when asked explicitly about their risk of switching to a competitor due to the current advertising use regarding the Telekom E-Mail Center, 36 percent of customers said that switching to the competition was very unlikely, and only two percent considered it very probable. In addition, a correlation was found between customers' willingness to switch and their age. Accordingly, to a certain extent: The older researched customers are, the more likely they are to consider switching to another provider due to the advertising intensity of the considered supplier. This means that 28 percent of the interviewed users are most willing to switch providers at the age of 71 or older. In comparison, the average customer churn measured at Deutsche Telekom was only seven percent.

VI. CONCLUSIONS

The present study examined the extent to which online advertising in web-mail services affects customer satisfaction. At the core of the analysis was not only the question of how online advertising can be used optimally without triggering negative effects on customer satisfaction or even customer churn. Furthermore, empirical research was also conducted on how different advertising formats and content were evaluated from the consumer's point of view and what level of advertising intensity the users accepted. To answer the research questions, an online survey was conducted among the customers of the web-mail service of Deutsche Telekom, in which a total of 2,228 respondents took part.

In conclusion, it can be summarized that online advertising is certainly criticized by users. The extent to which customer satisfaction ultimately has an effect on negative changes in user behavior, such as customer migration to competitors, depends on the respective provider. Consequently, in addition to advertising, other factors influencing customer satisfaction should also be considered. This concerns above all the trust of the customers in the provider with regard to the handling of sensitive, personal data. Thus, a negative influence of high

 ISBN: 978-1-61208-829-7

advertising volume on the willingness of customers to switch to web-mail services could be proven (57 percent), but played a secondary role in addition to other factors, such as data protection concerns (74 percent) or poor usability (55 percent).

VII. Limitations and Outlook

In order to be able to derive long-term benefits from the results of a customer satisfaction survey, repeated measurements at regular intervals will be required in the future [4]. The more intangible the service is, the more often satisfaction measurements or complaint statistics should be analyzed [6]. In this context, the sample could generally be enlarged to prevent possible bias and to represent the total population. Moreover, a sample involving other web-mail services (e.g. from several countries) could increase the generalization of the results. Furthermore, according to the presented results, a significant percentage of customers believe that the volume of advertising on the Deutsche Telekom portal is about the same as that of other web-mail services they use. Therefore, it would be interesting to find out what causes a homogeneous increase in advertising volume for most services. It might be possible that the churn rate does not change for the same advertising volume of competitors.

The connection between the topics of the influence of online advertising on customer satisfaction and the pressure to monetize for advertising-financed Internet service providers is outside the focus of the research but is, therefore, no less important. The researched telecommunications provider Deutsche Telekom will remain indirectly dependent on the active advertising consumption of its users in the future if a free version continues to be offered, which is ultimately financed by advertising revenue. As a result, Deutsche Telekom might not only continue to depend on advertising partnerships but also on achieving satisfactory click-through rates for its web-mail customers on advertisements displayed in the researched front-end of the Telekom E-Mail Center. Conversely, Deutsche Telekom would lose revenue if the number of bookings of online advertising spaces within the Telekom E-Mail Center was to decline as a result of falling click rates.

In order to build on the insights gained, it is advisable to additionally determine how strongly the placed advertising is generally perceived by the customers. In addition, research into the willingness to pay for web-mail services of customers would be conceivable. On the basis of the current state of research [26], it could, therefore, be investigated to what extent customer satisfaction affects their willingness to pay for web-mail services.

References

[1] R. Bell and A. Buchner, "Positive effects of disruptive advertising on consumer preferences," *Journal of Interactive Marketing*, vol. 41, pp. 1–13, 2018.

[2] M. Hudak, E. Kianickova, and R. Madlenak, "The importance of e-mail marketing in e-commerce," *Procedia Engineering*, vol. 192, pp. 342–347, 2017.

[3] R. Milner and A. Furnham, "Measuring customer feedback, response and satisfaction," *Psychology*, vol. 8, no. 3, pp. 350–362, 2017.

[4] D. Ahlert, P. Kenning, and C. Brock, *Handelsmarketing: Grundlagen der marktorientierten Führung von Handelsbetrieben (Trade marketing: Basics of the market-oriented management of trading companies)*, 2. Aufl. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018.

[5] J. van Doorn and J. C. Hoekstra, "Customization of online advertising: The role of intrusiveness," *Marketing Letters*, vol. 24, no. 4, pp. 339–351, 2013.

[6] H. Meffert, M. Bruhn, and K. Hadwich, *Dienstleistungsmarketing: Grundlagen – Konzepte – Methoden (Service Marketing: Basics – Concepts – Methods)*, 9., vollständig überarbeitete und erweiterte Auflage. Wiesbaden: Springer Fachmedien Wiesbaden, 2018.

[7] Statista, *E-Mail – Anteil der Nutzer in Deutschland 2019 (E-mail - Proportion of users in Germany 2019)*, 2020. [Online]. Available: https://de.statista.com/statistik/daten/studie/204272/umfrage/nutzung-des-internets-fuer-versenden-empfangen-von-e-mails-in-deutschland/ [retrieved: 07/15/2020].

[8] Deutsche Telekom AG, *Führender europäischer Telekommunikations-Anbieter (Leading European telecommunications provider)*, 2020. [Online]. Available: https://www.telekom.com/de/konzern/konzernprofil [retrieved: 07/20/2020].

[9] Infosat, *Telekom startet neue Email-Domain @magenta.de (Telekom starts new email domain @magenta.de)*, 2019. [Online]. Available: https://www.infosat.de/entertainment/telekom-startet-neue-email-domain-magentade [retrieved: 07/20/2020].

[10] J. Barnard, *Advertising expenditure forecasts march 2019: Executive summary*, 2019. [Online]. Available: https://www.zenithmedia.com/product/advertising-expenditure-forecasts-march-2019/ [retrieved: 07/20/2020].

[11] Statista, *Programmatic advertising*, 2018. [Online]. Available: https://de.statista.com/statistik/studie/id/54223/dokument/programmatic-advertising/ [retrieved: 07/15/2020].

[12] Statista, *Ausgaben für Online-Werbung weltweit in den Jahren 2013 bis 2017 sowie eine Prognose bis 2022 (in Milliarden US-Dollar) (Spending on online advertising worldwide from 2013 to 2017 and a forecast until 2022 (in billions of US dollars)*, 2018. [Online]. Available: https://de.statista.com/statistik/daten/studie/185637/umfrage/prognose-der-entwicklung-der-ausgaben-fuer-online-werbung-weltweit/ [retrieved: 07/20/2020].

[13] Statista, *Digitale Werbung (Digital advertising)*, 2019. [Online]. Available: https://de.statista.com/outlook/216/137/digitale-werbung/deutschland [retrieved: 07/15/2020].

[14] Nielsen, *Bereinigter Werbetrend, Datenstand 17.06.2019: Werbetrend: Top Trends im Mai 2019. (Adjusted advertising trend, data status 17.06.2019: Advertising trend: Top trends in May 2019.)* 2019. [Online]. Available: https://www.nielsen.com/de/de/insights/reports/2019/top-ten-trends.html [retrieved: 07/20/2020].

[15] Cisco, *Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022*, 2019. [Online]. Available: https://davidellis.ca/wp-content/uploads/2019/12/cisco-vni-mobile-data-traffic-feb-2019.pdf [retrieved: 07/20/2020].

[16] A. Ahmed and D. M. Linen, "A review and analysis of churn prediction methods for customer retention in telecom industries," in *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 2017, pp. 1–7.

[17] O. Adwan, H. Faris, K. Jaradat, O. Harfoushi, and N. Ghatasheh, "Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis," *Life Science Journal*, vol. 11, no. 3, pp. 75–81, 2014, ISSN: 1097-8135.

[18] A. Sundararajan and K. Gursoy, *Telecom customer churn prediction*, 2020. DOI: 10.7282/t3-76xm-de75. [Online]. Available: https://rucore.libraries.rutgers.edu/rutgers-lib/62514/PDF/1/play/ [retrieved: 07/20/2020].

[19] A. Amin *et al.*, "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing*, vol. 237, pp. 242–254, 2017.

 ISBN: 978-1-61208-829-7

[20] A. Piazza, C. Lutz, D. Schuckay, C. Zagel, and F. Bodendorf, "Emotionalizing e-commerce pages: Empirical evaluation of design strategies for increasing the affective customer response," in *Advances in Intelligent Systems and Computing*, Springer, Ed., Nuremberg, Coburg: Springer, Cham, 2018, pp. 252–263.

[21] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Commun.*, vol. 30, pp. 169–190, 2017.

[22] M. Bruhn, *Qualitätsmanagement für Dienstleistungen: Handbuch für ein erfolgreiches Qualitätsmanagement. Grundlagen – Konzepte – Methoden (Quality management for services: Manual for successful quality management. Basics – concepts – methods)*, 10. Berlin, Heidelberg, 2016.

[23] G. Böttcher, *Schlechter Service ist häufiger Grund für Anbieterwechsel (Poor service is a common reason for change of provider)*, 2013. [Online]. Available: https://www.springerprofessional.de/kundenservice/kundenmanagement/schlechter-service-ist-haeufiger-grund-fuer-anbieterwechsel/6603850 [retrieved: 07/20/2020].

[24] J. Mandak and J. Hanclova, "Use of logistic regression for understanding and prediction of customer churn in telecommunications," *Statistika*, vol. 99, no. 2, pp. 129–141, 2019.

[25] Publicare Marketing Communications, *Der Deutschen liebste E-Mail-Dienste 2019 (The German's favorite e-mail services 2019)*, 2018. [Online]. Available: https://publicare.de/blog/publicare-e-mail-studie-2019/ [retrieved: 07/15/2020].

[26] N. Koschate-Fischer, "Preisbezogene Auswirkungen von Kundenzufriedenheit (Price-related effects of customer satisfaction)," in *Kundenzufriedenheit*, C. Homburg, Ed., Wiesbaden: Springer Gabler, 2016, pp. 93–121, ISBN: 978-3-658-08688-6.

 ISBN: 978-1-61208-829-7

Intent Identification and Analysis for User-centered Chatbot Design: A Case Study on the Example of Recruiting Chatbots in Germany

Stephan Böhm,
Judith Eißer,
and Sebastian Meurer

CAEBUS Center of Advanced E-Business Studies
RheinMain University of Applied Sciences
Wiesbaden, Germany
e-mail: {stephan.boehm, judith.eisser,
sebastian.meurer}@hs-rm.de

Olena Linnyk, Jens Kohl,
Harald Locke, Levitan Novakovskij,
and Ingolf Teetz

Milch & Zucker AG,
Gießen, Germany
e-mail: {olena.linnyk, jens.kohl,
harald.locke, levitan.novakovskij,
ingolf.teetz}@milchundzucker.de

***Abstract*—Chatbots are text-based dialogue systems that automate communication processes. Instead of communicating with a person, the user communicates with a computer system. Due to the use of Artificial Intelligence (AI) methods, such systems have become increasingly powerful in recent years and allow for more realistic dialogue processes. In particular, methods from the field of machine learning have contributed to an improved understanding of natural language. Nevertheless, such systems are not yet able to acquire the knowledge required to answer user queries independently. Dialogue structures and elements need to be defined as the conversational design of the chatbot. Herein, an user intent describes an information need or a goal that the user aims to achieve by entering text. For a user-centered chatbot design, a relevant set of intents must be identified and structured. In addition, training questions are required in order train the AI models for matching user input with the defined set of user intents. This article describes the procedure for developing chatbots using the example of an application in recruiting. The focus is on the appropriate identification and analysis of user intents. In our case study, the procedure for user-centered intent identification is described as well as approaches for the analysis and consolidation of intents. Furthermore, it is shown how corresponding measures affect the quality of intention identification.**

Keywords–Chatbots; Conversational Design; Prototyping; User Intent Analysis; User-centered Design; Machine Learning.

I. INTRODUCTION

The mode of communication has changed. Where in the past, information was normally only provided by companies in one direction and in a unidirectional one-to-many approach, interactive one-to-one dialogues are possible at large scales today [1]. Stakeholders can converse with companies and vice versa. Chatbots are a way to automate this dialogue process and are implemented to address this need [2]. Based on pattern matching and natural language processing methods or artificial intelligence, chatbots are automated dialogue systems for conversational scenarios [3]. They are utilized to mimic unstructured natural language dialogues normally prevailing in human-human conversations; either based on hand-built rules or on corpus-based AI functionalities, where data is mined from existent human-human conversations [4]. The potential of chatbots is vast and its diffusion continues to progress: According to a global chatbot market report by Research and Markets, the chatbot market size will be worth 9.4 billion US dollars by 2024 at an estimated compound annual growth rate of almost 30 percent [5]. Established in the 1960s, technological advancements constantly improved the technology so that today, chatbots hold the potential to support various business processes [6]–[8]. Especially in repetitive scenarios like answering Frequently Asked Questions (FAQ), AI-based technology, such as chatbots, are implemented to increase efficiency by improving quality while reducing costs [9].

This paper is about the implementation of chatbot solutions in recruiting, a special field of human resources that deals with finding and hiring new personnel for employers like companies and other institutions. In recruiting, chatbots can be deployed to transfer information to potential candidates and talents before, throughout and after the application process. They can be utilized to answer general questions regarding a certain position or the application process for example [6]. Through automation and the deployment of artificial intelligence functionalities, the processes of applicant sourcing and screening can be supported and the aspect of human bias in recruiting can be reduced [9]. In the current "war for talents", state-of-the-art technology enabling or at least facilitating the process of recruiting the most suitable talents at the most suitable points of contact for them is essential for organizational success and the formation of a competitive advantage [10]. There are several relevant and interesting use cases along the recruiting process, which can be supported by chatbot functionalities; the focus areas of this study will be shed light on when regarding recruiting chatbots in more detail in Section II-C.

The use of chatbots in recruiting is still relatively new. There are already many example applications (e.g., [11]), but these are often early pilot and test applications and in many cases not yet in permanent productive use. Nevertheless, there are more and more developers of chatbot solutions [12] and many of them use AI to promote such new applications. For decision makers in the HR sector with less technical experience, the impression sometimes arises that chatbot solutions are largely autonomous learning systems that only need to be

 ISBN: 978-1-61208-829-7

implemented in companies and then acquire the knowledge to answer user questions themselves. However, this is a major misunderstanding. The use of AI in many chatbot frameworks is still largely limited to Natural Language Understanding (NLU) and the classification of user questions to predefined user intentions. Usually, however, the user intentions have to be created in the system and linked to certain actions for output. Developers of chatbots must therefore not only implement such solutions technically, but also define and structure dialogue contents in a conversational design [13]. The selection of the user intentions to be considered plays a special role, as it defines the application domain within which a chatbot can answer user requests in a meaningful way. For the identification and further analysis of such user intentions, however, the literature contains hardly any practical descriptions of the procedure [14].

This study regards the necessity as well as the actual formation process of a suitable intent set for a corpus-based recruiting FAQ chatbot while challenging the newly trained version against the former version of the dialogue technology prototype. After this introduction, an overview of the theoretical background is given in Section 2. Related work and studies are discussed before defining the research objectives of the study at hand in Section 3. The study's outline is presented in Section 4 with the methodology and the case study approach. Section 5 deals with the case study findings and its theoretical as well as practical implications. The last Section 6 presents final conclusions, limitations and an outlook on further research.

II. Research Background

In order to understand the problem of identification and analysis of user content, a brief discussion of some background information on conversational design will be given in the following. Afterwards, the technical implementation of AI-based chatbot solutions and the importance of training data will be discussed. The section concludes with an introductory description of FAQ chatbots in general and their application in connection with applicant tracking systems used in recruiting.

A. Conversational Design

Chatbots belong to the conversational interfaces [15]. Conversational interfaces are a special kind of interactive user interface, which enables a dialogue in natural language between humans and computers and can process user input as text or speech, oftentimes based on AI functionalities [13][16]. Popular conversational interfaces are voice assistants that react to spoken user input and chatbots, which are discussed here. In chatbot solutions, conversation typically takes place through typed text input and a front-end that can be, for example, embedded in a website or messaging solution [17]. Conversational design as a special discipline of interactive design deals with all tasks of designing conversational interfaces (e.g., stakeholder and goal definition, conversational flow design [16], actual development and testing) with the goal to provide a good user experience [18].

Like other objects of interactive design, chatbots have different design elements. The design of the front-end user interface is less in focus, since text input leaves little room for variation. Fist of all, interfaces for text input can be varied by the colours or by font characteristics. Furthermore, decisions on the chatbot's personality in the form of a specific persona [15][18] (e.g., use of avatars) or the use of graphic elements for the chatbot output such as buttons, images (moving or static) as well as emoticons and emojis can be considered as design aspects [19]. The tonality of the language is another exemplary design aspect [6]. The core of the chatbot's conversational design, however, is more concerned with determining the dialogue content and its logical structure. However, the respective design options for chatbot development are determined by the particular chatbot frameworks and platforms used. The elements for the conversational design of chatbots, as well as the terms used to describe them, vary between these frameworks and platforms.

In this case study, the framework Rasa [20] was used for chatbot development. Important basic elements are utterances, intents, entities, actions, and stories (e.g., [21]):

- *Utterances* are all expressions of users that are entered as user input into the chatbot user interface.
- *Intents* refer to goals that a user intends to achieve with the dialogue or information needs that a user wants to satisfy through communication with the chatbot.
- *Entities* modify or specify an intent and are extracted from the intent by the chatbot solution for further processing. This can be, for example, time and date information, places, names, quantities, etc.
- *Actions* define the output of the chatbot as a reaction to a certain intent and can contain not only text, but also links, buttons, graphical elements or videos. For natural language communication, however, text output is the main focus.
- *Stories* are used to link the different elements with each other, e.g., to specify a defined action for a certain intent.

According to [13] and [16], it can be distinguished between one-shot questions and those allowing for subsequent follow-up inquiries: In the simplest case or with one-shot queries from users, the chatbot generates a specific answer to a specific question (e.g., user: *"What university degree do I need for the job?"* → chatbot: *"You need a master's degree in electrical engineering."*). However, more complex dialogues are only possible if the context of successive questions is taken into account and, for example, more advanced follow-up queries are possible (e.g., user: *"Do I need a university degree for the job?"* → chatbot: *"Yes, a master degree."* → user: *"In which subject?"* → chatbot: *"In electrical engineering."*). This paper is a work-in-progress and initially deals with a chatbot prototype that focuses on successive one-shot queries and thus abstracts from the complexity of a contextual dialogue. In the following, the focus therefore lays on the identification and analysis of intents. However, for a more natural dialogue, aspects of the context must be taken into account in the future if the chatbot is developed further.

B. AI-Based Chatbot Implementation and Training Measures

Over the years, several different attempts proved valuable to create an AI which is able to respond to human queries and can thus be used as a foundation for advanced chatbots. It is possible to use sequence to sequence models [22][23]. For that purpose, the encoder processes the incoming query and

 ISBN: 978-1-61208-829-7

generates a vector representation of the query. Queries contain a certain intent. As mentioned before, an intent expresses the user's intention he pursues with a made query in the sense of completing a certain task, for example to find a specific information [24]. Intents can thus be defined as predefined classes incoming inquiries can be categorised into and represent the types of queries the chatbot is capable of handling [7][25]. The decoder uses the established query representation to generate an answer. As a benefit, there is no need for a distinct set of answers, but answers are completely generated based on the user's input. In general, the models used for this do not need a task-specific setup; a domain specific corpus is required which contains generic queries and answers. But such corpora are quite scarce and rarely freely accessible.

Another possibility is to generate a vector representation of the incoming query and compare that representation to the ones of already known queries trying to find the best match [26]. If a reasonable match is found, it will be assumed that the new query is about the same intent as the known one or if no reasonable match is found, it can be assumed to have encountered an unknown query. It aims at clustering incoming queries and assign a general answer to each cluster. Although there are no unique answers created as for the sequence to sequence modeling, it is possible to easily expand the scope of such a system by adding new answers to the algorithm. The main problem is that generating sentence representation [27] is still challenging handling negations, contradictions and reciprocations.

Additionally, such a task can be seen as a classification problem. This completely limits the scope of the AI to the *a priori* set of answers. But the reduction in flexibility at least is accompanied with the AI model being specifically designed for the task [28]. In addition, the AI should be able to detect phrases, which are out of domain [29].

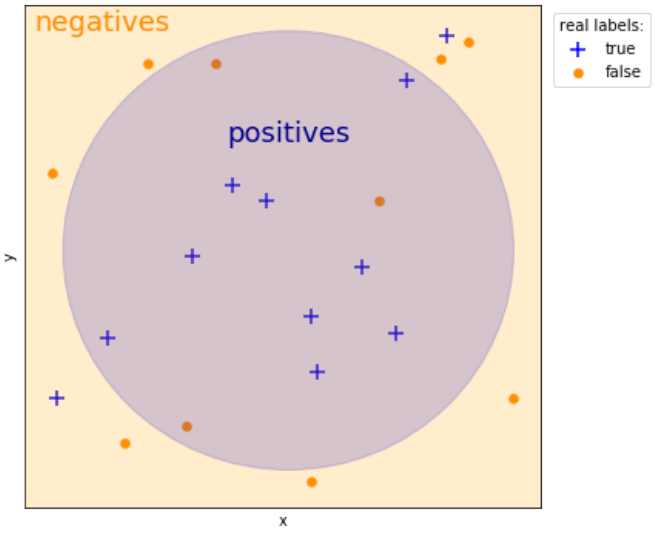

Figure 1: Example for the predictions of an algorithm

Figure 1 shows a possible prediction of an algorithm. All data points within the circle are predicted as true by the algorithm, the data points outside of the circle are predicted as false. A true positive is a data point that has the label "true" and the algorithm has also predicted this label for the data point. So a false positive actually has the label "false" but was predicted as true. The system is applied to the false labels. A true negative is a data point with the actual label false, which was also predicted by the algorithm as false. Finally, a false negative is a data point which was predicted as being true but with the real label "false". In the example above (Figure 1) there are eleven data points with the label "true" and ten with the label "false". An algorithm tried to predict the labels and predicted all data points within the circle as being true. This results in nine true positives and three false positives. An important measure for the accuracy in intent classification is the F1-score F_1 (e.g., to be seen in [30]). It is the harmonic mean of precision p and recall r. The precision p denotes the share of true positives from all positives. So in the example above (Figure 1), there are twelve data points determined as being true (the positives), but only nine of them are actually true (the true positives). Therefore, the precision $p = \frac{9}{12} = 0.75$. The recall denotes the share of true positives from all true labels. There are nine true positives in the example, but overall there are eleven data points with the label true. So, the recall results in $r = \frac{9}{11} \approx 0.82$. The F1-score is the finally calculated by $F_1 = 2 \cdot \frac{p \cdot r}{p+r} \approx 2 \cdot \frac{0.75 \cdot 0.82}{0.75+0.82} \approx 0.78$. For a non-binary classification problem, there is an F1-score for every label and the overall F1-score is usually calculated by averaging.

C. FAQ-Chatbots in Recruiting

This study applies an AI algorithm to the case of recruiting chatbots. As introduced, the recruiting process is especially suitable for efficiency enhancement by automation technology implementation [9]. Chatbots as automated dialogue systems can be deployed in various steps of the recruiting process to unburden the recruiters and leave them with the more strategic parts of the work while increasing efficiency as well as to reduce costs. They comply with the newly established requirements of potential candidates, who demand digital touch points in the form of mobile accessible websites and instant messaging [31]–[33]. According to a recent study in North America and Europe by Spiceworks [34], among organizations that currently deploy chatbots, 23 percent of administrative departments are equipped with such dialogue systems and seven percent already utilize this technology specifically within their human resource departments. Areas of application for chatbots along the recruiting process, some of which requiring components of artificial intelligence, are creating and posting job profiles, assisting job searches and the specific application process of potential candidates, handling incoming queries by applicants concerning general questions, support of recruiters during candidate pre-selection as well as during the hiring process [7]. Through automation, the efficiency of conducting these steps is improved [9]. Furthermore, the employer attractiveness is enhanced through chatbot implementation: According to a study by Phenom People based on more than 20 million chatbot interactions across over 100 chatbot deployments of the company, the number of job seekers turning into candidates applying for the job almost doubles (increase from 12 percent to 23 percent when implementing a chatbot on the career website) and the amount of candidates completing an application increases by 40 percent [35]. However, chatbots need to be integrated into the recruiters' Applicant Tracking

 ISBN: 978-1-61208-829-7

Systems (ATS), which handle application data and recruiting workflows in order to realize these potentials and to enhance the recruiting process [7]. Furthermore, chatbots cannot be seen as solution for any kind of application area and are no solution for all problems potentially occuring in recruiting.

The creation and integration of AI-based chatbots into ATS systems is being regarded in the governmentally funded research project CATS (Chatbots in Applicant Tracking Systems), which is a conjoint initiative of RheinMain University of Applied Sciences in Wiesbaden, and the talent management company milch&zucker AG in Gießen, Germany. This study project aims at the creation of a conceptual framework for a flexibly configurable chatbot toolbox, which is implementable prior, during and after the application process. An assortment of appropriate use cases as well as suitable intents is essential for relevant chatbot development. For specific use case selection of this study, interviews and surveys have been conducted with (1) technical, (2) scientific, and (3) industry experts concerning recruiting. The participants agreed upon FAQ scenarios (process guidance, application- and workflow-related questions, and guidance through the onboarding process) to be the most relevant and realistic in terms of support by chatbot implementation. This result is consistent with industry observations, which found questions related to the application status, job search and the company itself to be most common for chatbot inquiries [35]. Hence, these FAQ-related scenarios have been implemented into this study's case and an item set for an FAQ recruiting chatbot complying with these content requirements has been created.

III. Related Work and Research Objectives

A. Literature Review

Several studies already investigated the effects of AI in general (e.g., [9][10][36]) and chatbots in particular (e.g., [37]–[39]) on the recruiting process. However, as opposed to many studies incorporating either (1) perfunctory intent creation descriptions neglecting a comprehensive discussion of imperative underlying strategic considerations (e.g., [40]–[43]), (2) proposals of evaluation, i.e., rating and training methods for diverse chatbot prototypes without disclosure of the intent creation process (e.g., [44]–[47]), or (3) general investigation of intent matching and classification methods only (e.g., [48]–[51]), the interplay of intent creation and intent analysis within conversational design is not well covered by scientific research. Only two studies were found that deal with both the creation and the evaluation of intents for the use cases of (1) a hotel assistant chatbot [52] and (2) a Latvian customer support chatbot [53].

The most common error encountered within chatbot deployment according to the aforementioned study by Spiceworks [34] is the misunderstanding of incoming queries. This can refer either to (1) the intent matching capabilities of the chatbot framework, or (2) to the underlying intent set itself. Hence, developing and refining the most suitable list of intents alongside matching training and test data is fundamental to successful chatbot deployment. Encompassing evaluation is another crucial part of dialogue system design [15][46]. Human assessment is necessary within the evaluation of chatbots, either to (1) measure absolute task success, or (2) investigate user satisfaction on a more fine grained scale [4]. According to Walker et al., the users' perception of task completion success can predict user satisfaction better than actual task completion success [54]. Thus, an evaluation of the chatbot prototype from the users' perspective is conducted in this study with four users via rating of its response quality prior and after training (see Section IV-B). This kind of analysis is defined as session level user satisfaction evaluation [46].

B. Research Gap and Objective

There is an apparent lack of encompassing research dealing with both the establishment and the iterative adjustment process based on the evaluation of suitable chatbot intent sets. As seen throughout the literature review, only very few studies are known to the authors that disclose an in-depth approach to intent creation and evaluation through pre- and post-training tests of the different versions at the same time. This study gives detailed insights to the process of intent set creation and enhancement and furthermore proposes a structured approach for a recruiting FAQ chatbot. The central research questions are:

1) What is a relevant intent set for an FAQ recruiting chatbot?
2) Which effects can be seen when training the chatbot with enhanced data (intents and formulation variations) for improvement?

In the following, the approach to answer these research questions will be explained within the methodological section of the study.

IV. Methodology and Case Study Approach

As shown in the literature review, the identification of user intentions is an essential starting point of user-centric conversational design for chatbots. In the following, a case study in recruiting is used to describe how a basic set of user intents can be generated from various information sources. Starting from this basic set, the intents are analysed, cleaned up and provided with variations of user queries in a multi-stage process involving users. In addition, AI models are trained and evaluated with the sets of intents and corresponding variations of user questions. Finally, the resulting AI-based chatbots versions are subjected to user tests in order to evaluate the achieved quality of intent recognition.

A. User-centered Intent Identification

As described in the introduction, the starting point of our case study presented in this paper is first of all the composition and structuring of a comprehensive list of intents in the context of recruiting. Therefore, this section will explain the methodological approach in the sense of a user-centered attempt to identify user intents towards the chatbot (user centered intent identification) as well as suitable alternative formulations (to be used later as training and test questions). The following two sections then describe the approaches used to analyze and consolidate the developed intent sets, as well as the effects of modifications and model training on selected metrics and satisfaction values when using a corresponding chatbot prototype.

The overall methodological approach consists of five steps:

1) *Intent Sourcing*: Accumulation of potential intents from (1) website FAQs, (2) mail inquiries, (3) an expert review, and (4) user tests (see Figure 2).

 ISBN: 978-1-61208-829-7

2) *Intent Funneling*: Reduction of the initial item set via consolidation, reviewing and merging processes.
3) *Intent Variation*: Variation of the finalized item set through word substitution and splitting in into training and testing phrases.
4) *Intent Optimization*: Optimization of the item set through training, testing and intent matching coefficient improvements.
5) *Intent Validation*: The finalized item set is validated via a structured user test.

From the four sources described within the intent sourcing process, almost 500 initial intent propositions were drawn, which were reviewed, merged and eliminated in case of duplicates so that 82 final items emerged (see funneling process in Figure 2).

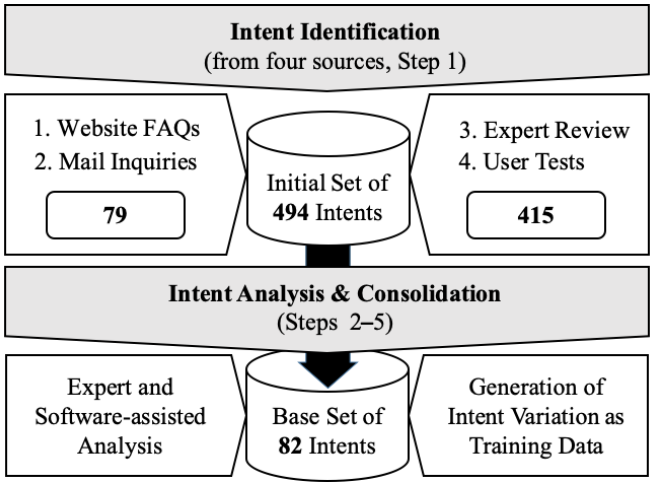

Figure 2: Overview of Intent Identification and Analysis

B. Analysis and Consolidation of Intents

After creation of the data set (base set), several instances of a natural language understanding artificial intelligence were trained to classify the intent of an input phrase. As mentioned before, the framework Rasa was used in general for the AI. Five different pipelines for the processing of the input messages were created and the DIET classifier was used in all instances. Sparse features were created in all cases by count vectorization of n-grams. The first one consisted of a white space tokenizer and only created sparse features for the tokens by the means of a Regex featurizer and count vectorization of words and n-grams. The second one additionally used the spacy components for creating tokens and dense features. The third one used the HFTransformerNLP with the "Bert"-Model applying the bert-base model-weights for uncased words implementing the associated tokenizer and featurizer. In the remaing two instances, a white space tokenizer was used again and a neural network incorporating a biLSTM was used to create dense word embeddings from the char sequences of the input words. The corpus used to train both of these networks was chosen specifically for the task of job search consisting of over 400,000 job ads and 12,000 anonymized support emails from a company's human resources management. One of these networks was trained by the approach by Ling *et al.* [55], training the previously mentioned embedding network as a part of a natural language processing task. In the other one, the network was trained to mimic the vectors created by a glove embedding as suggested by Pinter *et al.* [56]. These two networks based on character sequences rather than look-up tables were chosen in order to prevent the occurrence of out-of-vocabulary (oov) words. For comparison of all these setups, a five-fold cross-validation was performed on all models. The instance using the

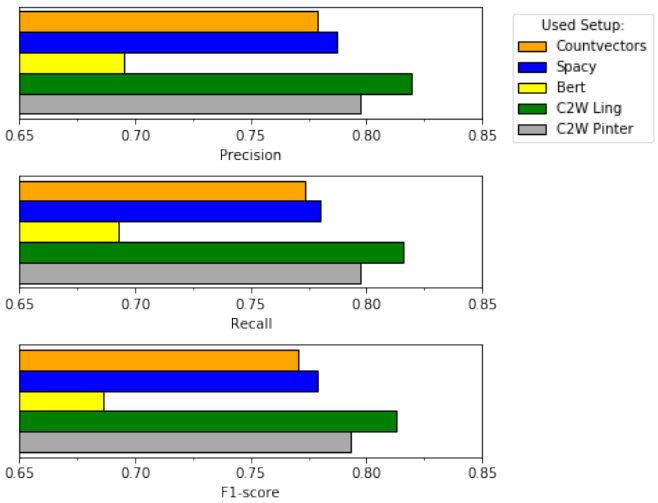

Figure 3: Comparison of setups for Rasa

character to word embedding network as suggested by Ling *et al.* outperformed the other instances reaching an F1-score of 0.81 in average (see figure 3). The second best setup was the one incorporating the word embedding model of Pinter *et al.* with an average F1-score of 0.80. The spacy setup followed in third place barely beating the pure count vectorization of words approach by 0.01 comparing their respective F1-scores of 0.78 and 0.77. The most plausible explanation for the character-based neural networks outperforming spacy is that the corpus for them was chosen specifically for the task, while a general corpus based on news articles was used for spacy. Surprisingly, although consisting of a very sophisticated architecture, the Bert model performed worst for this task reaching an F1-score of only 0.69 in average. Fine-tuning the Bert-model might drastically improve the performance, but as Bert is by far the most resource-consuming model in this study and also performing worst, it was further excluded from investigation.

To understand the sources of the errors, the confusion matrices were investigated and compared to the cosine similarities between the phrases of all intents. To calculate the cosine similarities, a sparse vector was assigned to every phrase with every entry consisting of the text frequency inverted document frequency value for every word in the corpus. This allowed the detection of several nearly or fully identical phrases within different intents, which explained at least some of the errors in the intent classification task. The data set was reworked, eight of the intents were removed and the corresponding phrases were shifted to other intents, ten were reworked and two new ones were created. The set of answers was reworked, too in order to fit to the new list of intents. Again, a five-fold cross-validation was performed to estimate the performance. This time, only the previously best-performing pipeline was used applying the character sequence to word embedding model of

 ISBN: 978-1-61208-829-7

Ling *et al.*

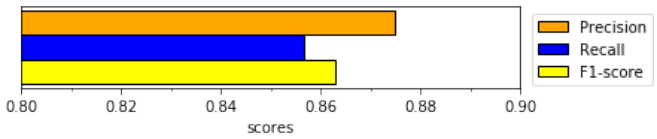

Figure 4: Scores for the setup using the character-based word model by Ling *et al.* after reworking the corpus

The F1-score was 0.86 in average for the new dataset (see Figure 4). A comparison of the scores of the two data sets was neglected, as there is a different number of classes and data points used. A reduction in classes should in general be accompanied with an increase in accuracy.

The intra-rater reliability $\kappa = \frac{p_0 - p_c}{1 - p_c}$, where p_0 is the accuracy of the chatbot in choosing the right intent and p_c is the probability to select the right intent by chance, is a measure showing how reliably a query is classified to the right intent. This reliability metric was calculated for both chatbot instances, the one trained on 88 and the one on 83 intents, to be 0.81 and 0.85 respectively (see Figure 5).

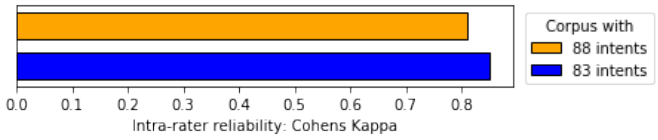

Figure 5: Intra-rater reliability for the two corpora with a different number of intents

It is important to note that the two values cannot be directly compared but can provide qualitative measure of performance. Although this metric does also not allow a direct comparison, a value above 0.8 usually shows that the predictions made by an algorithm are substantially reliable and not caused by chance. The higher score of 0.85 for the refined version might suggest also an higher reliability whereas viewing this as a general improvement has to be done with great care.

C. Measuring the Impact of Improved Intent Sets

To still compare the two variants of the chatbot, it was tried to capture the user experience when confronted with the AIs. In order to do so, two new instances of the chatbot using the well performing char sequence to word vector embedding were trained on their respective whole data set. One data set being the original one and the other data set being the new one with an reduced number of intents and reformulated answers. An independent test set consisting of 1,400 phrases was created and both versions of the chatbot predicted the answers to these phrases. Finally, the number of four students raters, R1 to R4, had to rate these answers as ”good”, ”mediocre” or ”bad”. They were asked to rate an answer as ”good” if the answer fitted the question. A ”mediocre” answer meant, that the chatbot gave an answer which at least corresponded to the right topic but did not exactly answer the question. A ”bad” answer was one that did not match the intent at all. This threefold evaluation scheme is loosely based on [44], who rated the appropriateness of the dialogue system based on the three categories (1) appropriate, (2) interpretable (evasive answer), and (3) inappropriate.

The students rated the two chatbots quite differently: One of the raters strongly favored the chatbot training on the refined corpus, giving ”good” ratings more often while reducing the number of ”mediocre” and ”bad” ratings for its answers. Two of the student raters only favored the refined version of the AI by a slight margin, with the tendencies towards ”good” ratings being less distinct as compared to the first student. The remaining student even gave fewer ”good” ratings for the answers of the refined version of the chatbot. Also, this student rated fewer answers as ”bad” mainly resulting in an increase in ”mediocre” answers. Overall, the answers of the refined chatbot were rated ”good” and ”mediocre” more often (72.0% vs. 71.1% and 7.5% vs. 6.8%) and ”bad” less often (20.5% vs. 22.1%) (see Figure 6). Due to the low number of test persons, especially viewing the standard deviation of the ratings, these results are not significant enough to claim a general trend.

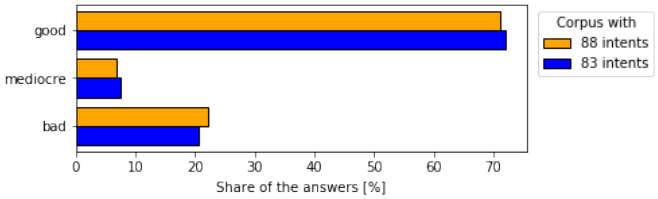

Figure 6: Ratings of the answers in average over all queries and students

One major problem of such a small number of test persons is that different mindsets are not averaged out and strongly dictate the outcome of the testing. Hence, all answers from both chatbot setups (corpora) were picked where all students gave the same rating. These should be very ”good” or very ”bad” answers, as the idea of what is ”mediocre” is more a question of the mindset than the extremes which are ”very good” or ”very bad”. Unsurprisingly, no answer was rated ”mediocre” by all students, but 67.4% of the answers for the chatbot trained on the original corpus with 88 intents and 69.6% of the answers of the refined chatbot got the same rating (see Figure 7).

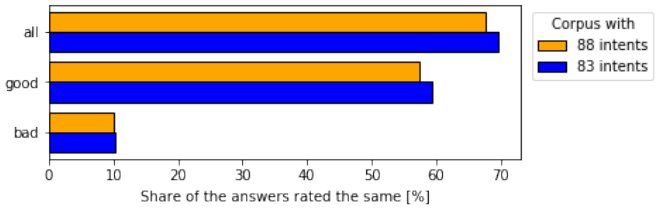

Figure 7: Percentages of all queries where the students gave the same rating for the answer of the corresponding chatbot

For the original chatbot 10.1% and for the refined one 10.2% of the answers were rated ”bad”. One might suggest that the the general setup of the chatbot combined with the limited training data is just not capable of understanding queries that are unknown. So, either the training corpus has

to be extended or the word embedding needs to better capture semantic similarities. Further, 57.4% of the answers for the first version and 59.4% of the answers for the refined version were rated as ”good” by all test persons. This slight improvement at least suggests some positive effect of the intent refinement.

Focusing on consistent ratings, it can be seen in Figure 8 that out of the 6,500 evaluated cases in total, 3,464 ratings remained unchanged with either a consistant ”good”, ”mediocre” or ”bad” rating after the training of the chatbot. In 532 cases, all reviewers consistently gave a ”good” rating prior and after the training while the total amount of unchanged good ratings was 3,024. As mentioned before, there was no case of uniform ”mediocre” rating throughout the evaluation study amongst all four reviewers while the total amount of consistent ”mediocre” ratings was 60 out of the 5,600. 380 cases were rated badly in total while only 25 of them were reviewed as ”bad” by all four reviewers. Looking at the positive (improvement, edged in green) and negative changes (deterioration, framed in red), it becomes apparent that overall, more cases improved (1,101) than worsened (1,035) throughout the training.

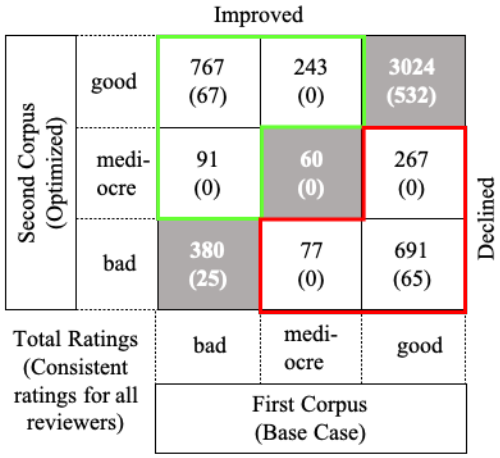

Figure 8: Overview of the user rating distribution

In Figure 9, the rating changes through the training of the chatbot are broken further down. While several of the reviewers' ratings seem to be similar, there are some noticeable differences between R1 and R3, especially regarding the verbatim ratings and the decline from the first towards the second corpus of the chatbot.

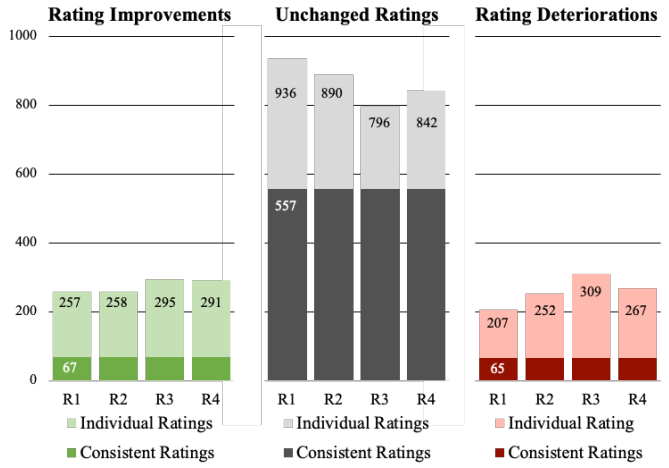

Figure 9: Rating comparison of improvement, verbatim state and deterioration

For the unchanged rating amounts, there is a spread of 140 differently rated cases and regarding the deteriorations, a gap of 98 stands out. However, with an exception of R3, the improvements or unchanged ratings outweigh the potential deterioration of the rating structure.

V. Conclusions and Managerial Implications

In summary, the chatbot composition and especially its conversational design is not finished yet. The training corpus still seems to be too narrow, suggested by the number of answers rated as ”bad” by all test persons. On the other hand, it seems that a useful setup for the embedding was found and that the refinement of intents had some effect. In the end, a lot of minor improvements will give rise to an overall powerful chatbot system.

The use of chatbots in recruiting will play a prominent role in the next few years in the handling of service dialogues within the organisation and towards the candidate. Especially in companies and organizations with a high number of applicants, the support of candidates in the recruiting process takes a considerable amount of time with frequently recurring requests for the same information. Here lies a significant savings potential on workforce (man-days) through the use of chatbot offers without worsening the quality of support. Furthermore, chatbots will play an increasingly important role in the dialogue between the hiring mamanger and the personnel/recruiting department. Chatbots will help with general FAQs but also with questions about the requirements of positions (skill management) or about the classification according to collective bargaining agreements, and they will also help with the formulation of advertisements. As they mature, they will also be able to help in the selection of applicants by autonomously conducting structured interviews.

The most important component to get there is to provide the best possible recognition of intents in the respective specific domain. As a basis, the conversational design in the form of a relevant and suitable intent set is indispensable. General user acceptance will then depend largely on apt responses and thus relevant content as well as a low number of incorrect answers within the dialogues.

 ISBN: 978-1-61208-829-7

VI. Limitations and Further Research

This case study described how an initial intent set for a FAQ chatbot can be developed for a specific application in recruiting and enhanced via a structured consolidation process (see Figure 2). The conversational design of this chatbot was initially limited to single-shot queries. Follow-up queries or a more complex context has not yet been considered in the dialogue modeling. In the following research work, such follow-up queries and context must also be included in dialogue modelling. It should also be noted that a user-centered improvement of the chatbot prototype can only be achieved if it interacts more extensively with real users and the resulting questions are used to extend and improve intent recognition. Nevertheless, the limits of the current chatbot development have already become clear due to the simple design. Only if the relevant intents can be captured, the chatbot will be able to provide a real benefit. The case study also showed that interdisciplinary cooperation between experts from different fields is necessary to successfully develop a chatbot. Conversational designers need to understand the basics of interactive design for conversational interfaces as well as the basics of AI solutions. Further research should also focus on how teams should be put together and which specific qualifications and skills are required for the individual roles and phases of the chatbot development process. Furthermore, a larger set of participants need to be exposed to the chatbot as a next step in order to yield generalizable information.

A research gap is also evident in the area of how the technical quality of an AI model and its improvement is related to the effect on the users. Developers and operators of chatbot solutions need, for example, technically derived information on how much training data is required or how a set of intents can be suitably improved. This is imperative as user tests are often complex and expensive and can have performance deficits at various levels of conversational design and AI components. Better research of such correlations is the basis for more sound recommendations for the design of chat offers in practice.

Acknowledgment

The study was carried out as part of the research project CATS (Chatbots in Applicant Tracking Systems) of the Rhein-Main University of Applied Sciences. This project (HA project no. 642/18-65) is funded in the framework of Hessen Modell-Projekte, financed with funds of LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, Förderlinie 3: KMU-Verbundvorhaben (State Offensive for the Development of Scientific and Economic Excellence).

References

[1] J. Rowley, "Just another channel? Marketing communication in e-business," *Marketing Intelligence & Planning*, vol. 17, no. 4, pp. 332–351, 2006.

[2] S. Böhm and J. Eißer, "Hedonic motivation of chatbot usage: Wizard-of-oz study based on face analysis and user self-assessment," in *Proceedings of CENTRIC 2017*, 2017, pp. 59–66.

[3] A. Mittal, A. Agrawal, A. Chouksey, R. Shriwas, and S. Agrawal, "A comparative study of chatbots and humans," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 6, no. 6, pp. 1055–1057, 2016.

[4] D. Jurafsky and J. H. Martin, "Speech and language processing (draft)," *Chapter 24: Dialog Systems and Chatbots (Draft of September 11, 2018). Retrieved March*, vol. 19, p. 2019, 2018.

[5] Research and Markets, *Chatbot market by component (solutions and services), usage (websites and contact centers), technology, deployment model, application (customer support and personal assistant), organization size, vertical, and region - global forecast to 2024*, 2019. [Online]. Available: https://www.researchandmarkets.com/reports/4858082/chatbot-market-by-component-solutions-and? [retrieved: 07/13/2020].

[6] L. Schildknecht, J. Eißer, and S. Böhm, "Motivators and barriers of chatbot usage in recruiting: An empirical study on the job candidates perspective in germany," *Journal of E-Technology Volume*, vol. 9, no. 4, pp. 109–123, 2018.

[7] S. Meurer, J. Eißer, and S. Böhm, "Chatbots in applicant tracking systems: Preliminary findings on application scenarios and a functional prototype," in *Proceedings of the IWEMB 2019: Third International Workshop on Entrepreneurship in Electronic and Mobile Business*, in press.

[8] G. V. Research, *Chatbot market size, share & trends analysis report by end user, by application/business model, by type, by product landscape, by vertical, by region, and segment forecasts, 2018 - 2025*, 2017. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/chatbot-market [retrieved: 07/13/2020].

[9] B. Hmoud *et al.*, "Will artificial intelligence take over human resources recruitment and selection," *Network Intelligence Studies*, vol. 7, no. 13, pp. 21–30, 2019.

[10] S. İşgüzar and C. Ayden, "New decision mechanism in the recruitment process: Artificial intelligence," in *A New Perspective in Social Sciences*, London: Frontpage Publications, pp. 196–205.

[11] S. Reviews, *The Top 10 Best Recruiting and HR Chatbots - June 2020*, 2020. [Online]. Available: https://www.selectsoftwarereviews.com/buyer-guide/hr-chat-bots [retrieved: 07/12/2020].

[12] A. Multiple, *Top 60 Chatbot Companies of 2020: In-depth Guide*, 2020. [Online]. Available: https://research.aimultiple.com/chatbot-companies/ [retrieved: 07/12/2020].

[13] M. F. McTear, "The rise of the conversational interface: A new kid on the block?" In *International Workshop on Future and Emerging Trends in Language Technology*, Springer, 2016, pp. 38–49.

[14] C. Pricilla, D. P. Lestari, and D. Dharma, "Designing interaction for chatbot-based conversational commerce with user-centered design," in *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, 2018, pp. 244–249.

[15] M. McTear, "Conversation modelling for chatbots: Current approaches and future directions," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*, pp. 175–185, 2018.

[16] S. Janarthanam, *Hands-on chatbots and conversational UI development: build chatbots and voice user interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*. Birmingham: Packt Publishing, 2017.

[17] J. Feine, S. Morana, and U. Gnewuch, "Measuring service encounter satisfaction with customer service chatbots using sentiment analysis," in *14th International Conference on Wirtschaftsinformatik*, 2019, pp. 1115–1129.

[18] R. Batish, *Voicebot and Chatbot Design: Flexible Conversational Interfaces with Amazon Alexa, Google Home, and Facebook Messenger*. Birmingham: Packt Publishing, 2018.

[19] A. Fadhil, "Domain specific design patterns: Designing for conversational user interfaces," *arXiv preprint arXiv:1802.09055*, 2018. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1802/1802.09055.pdf [retrieved: 07/12/2020].

 ISBN: 978-1-61208-829-7

[20] Rasa, *Rasa: Open source conversational AI - Rasa*, 2020. [Online]. Available: https://rasa.com/ [retrieved: 07/12/2020].

[21] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open source language understanding and dialogue management," *arXiv preprint arXiv:1712.05181*, 2017. [Online]. Available: https : / / arxiv . org / abs / 1712 . 05181 [retrieved: 07/12/2020].

[22] O. Vinyals and Q. Le, *A neural conversational model*, 2015. arXiv: 1506.05869. [Online]. Available: https://arxiv.org/abs/1506.05869 [retrieved: 07/12/2020].

[23] A. Sojasingarayar, *Seq2seq ai chatbot with attention mechanism*, 2020. arXiv: 2006 . 02767. [Online]. Available: https : //arxiv.org/abs/2006.02767 [retrieved: 07/12/2020].

[24] U. Şimşek and D. Fensel, "Intent generation for goal-oriented dialogue systems based on schema. org annotations," in *1st International Workshop on Chatbots Co-Located with the 12th International Conference on Web and Social Media (ICWSM2018)*, 2018, pp. 1–7.

[25] S. Srivastava and T. Prabhakar, "Intent sets: Architectural choices for building practical chatbots," in *Proceedings of the 2020 12th International Conference on Computer and Automation Engineering*, 2020, pp. 194–199.

[26] N. Lair, C. Delgrange, D. Mugisha, J.-M. Dussoux, P.-Y. Oudeyer, and P. F. Dominey, "User-in-the-loop adaptive intent detection for instructable digital assistant," *Proceedings of the 25th International Conference on Intelligent User Interfaces*, Mar. 2020.

[27] N. Reimers and I. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks*, 2019. arXiv: 1908 . 10084. [Online]. Available: https://arxiv.org/abs/1908.10084 [retrieved: 07/12/2020].

[28] A. Perevalov, D. Kurushin, R. Faizrakhmanov, and F. Khabibrakhmanova, "Question embeddings based on shannon entropy: Solving intent classification task in goal-oriented dialogue system," in *Proceedings of International Conference on Applied Innovation in IT*, 2019, pp. 73–78.

[29] S. Larson *et al.*, "An evaluation dataset for intent classification and out-of-scope prediction," in *roceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 1311–1316.

[30] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Interspeech 2016*, 2016, pp. 685–689. DOI: 10.21437/Interspeech.2016-1352. [Online]. Available: http : / / dx . doi . org / 10 . 21437 / Interspeech.2016-1352.

[31] C. Lieske, "Digitalisierung im Bereich Human Resources (Digitization in the field of human resources)," in *Digitalisierung in Industrie-, Handels-und Dienstleistungsunternehmen (Digitization in industrial, commercial and service enterprises)*, Wiesbaden: Springer Gabler, 2020, pp. 149–160.

[32] D. Bollessen, *Der fortschreitende Fachkräftemangel infolge des demographischen Wandels: Denkbare Konzepte und Erfolgsstrategien zur langfristigen Mitarbeiterbindung (The continuing shortage of skilled workers as a result of demographic change: Conceivable concepts and successful strategies for long-term employee retention)*. Diplomica Verlag, 2014.

[33] R. Hartmann, "Rekrutierung im Mittelstand: Trends und Herausforderungen im Personalmanagement oder von Trüffelschweinen und Wollmilchsäuen (Recruitment in medium-sized businesses: trends and challenges in human resources management or of truffle pigs and Jack of all trades)," in *Rekrutierung in einer zukunftsorientierten Arbeitswelt (Recruiting in a future oriented working world)*, M. Hartmann, Ed., Wiesbaden: Springer Gabler, 2015, pp. 215–234.

[34] Spiceworks, *Data Snapshot: AI Chatbots and Intelligent Assistants in the Workplace*, 2018. [Online]. Available: https : / / community . spiceworks . com / blog / 2964 - data - snapshot - ai - chatbots - and - intelligent - assistants - in - the - workplace [retrieved: 07/13/2020].

[35] Phenom People, *Chatbots for Recruiting - Benchmarks 2020*, 2020. [Online]. Available: https : / / www . phenom . com / resource/chatbots-for-recruiting-2020-benchmarks [retrieved: 07/13/2020].

[36] Q. Jia, Y. Guo, R. Li, Y. Li, and Y. Chen, "A conceptual artificial intelligence application framework in human resource management," in *Proceedings of the International Conference on Electronic Business*, 2018, pp. 106–114.

[37] Q. V. Liao *et al.*, "All work and no play?" In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[38] N. Nawaz and A. M. Gomes, "Artificial intelligence chatbots are new recruiters," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 10, no. 9, pp. 1–5, 2019.

[39] G. Suciu, A. Pasat, C. Bălăceanu, C. Nădrag, and A. Drosu, "Design of an internship recruitment platform employing nlp based technologies," in *ECAI 2018-International Conference, June 2018*, 2018, pp. 1–6.

[40] M. Fleming *et al.*, "Streamlining student course requests using chatbots," in *29th Australasian Association for Engineering Education Conference 2018 (AAEE 2018)*, Engineers Australia, 2018, pp. 207–211.

[41] G. M. Mostaco, I. R. C. De Souza, L. B. Campos, and C. E. Cugnasca, "Agronomobot: A smart answering chatbot applied to agricultural sensor networks," in *14th international conference on precision agriculture*, vol. 24, 2018, pp. 1–13.

[42] M. Carisi, A. Albarelli, and F. L. Luccio, "Design and implementation of an airport chatbot," in *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*, 2019, pp. 49–54.

[43] T. L. Vu, K. Z. Tun, C. Eng-Siong, and R. E. Banchs, "Online faq chatbot for customer support," in *Proceedings of the 2019 10th International Workshop on Spoken Dialogue Systems Technology*, 2019, pp. 1–6.

[44] Z. Yu, Z. Xu, A. W. Black, and A. Rudnicky, "Chatbot evaluation and database expansion via crowdsourcing," in *Proceedings of the chatbot workshop of LREC*, 2016, pp. 15–19.

[45] Å. Kamphaug, O.-C. Granmo, M. Goodwin, and V. I. Zadorozhny, "Towards open domain chatbots—a gru architecture for data driven conversations," in *International Conference on Internet Science*, 2017, pp. 213–222.

[46] W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit, and T. Theeramunkong, "A survey on evaluation methods for chatbots," in *Proceedings of the 2019 7th International Conference on Information and Education Technology*, 2019, pp. 111–119.

[47] E. Ruane, R. Young, and A. Ventresque, "Training a chatbot with microsoft LUIS: Effect of intent imbalance on prediction accuracy," in *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, 2020, pp. 63–64.

[48] B. Behera, "Chappie-a semi-automatic intelligent chatbot," *Write-Up*, pp. 1–5, 2016.

[49] M. Y. H. Setyawan, R. M. Awangga, and S. R. Efendi, "Comparison of multinomial naive bayes algorithm and logistic regression for intent classification in chatbot," in *2018 International Conference on Applied Engineering (ICAE)*, 2018, pp. 1–5.

[50] S. Alias, M. S. Sainin, T. S. Fun, and N. Daut, "Intent pattern discovery for academic chatbot-a comparison between n-gram model and frequent pattern-growth method," in *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 2019, pp. 1–5.

[51] K. Balodis and D. Deksne, "Fasttext-based intent detection for inflected languages," *Information*, vol. 10, no. 5, p. 161, 2019.

 ISBN: 978-1-61208-829-7

[52] L. N. Michaud, "Observations of a new chatbot: Drawing conclusions from early interactions with users," *IT Professional*, vol. 20, no. 5, pp. 40–47, 2018.

[53] K. Muischnek and K. Müürisep, "Collection of resources and evaluation of customer support chatbot," in *Human Language Technologies–The Baltic Perspective: Proceedings of the Eighth International Conference Baltic HLT 2018*, 2018, pp. 30–37.

[54] M. Walker, C. Kamm, and D. Litman, "Towards developing general models of usability with paradise," *Natural Language Engineering*, vol. 6, no. 3 & 4, pp. 363–377, 2000.

[55] W. Ling *et al.*, "Finding function in form: Compositional character models for open vocabulary word representation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1520–1530.

[56] Y. Pinter, R. Guthrie, and J. Eisenstein, "Mimicking word embeddings using subword RNNs," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 102–112.

 ISBN: 978-1-61208-829-7

User-Centered Methods Applied to 4D/BIM Collaborative Scheduling

Hugo Carvalho Mota
PErSEUs EA7312
Université de Lorraine
Metz, France
Email: hugo.carvalho-mota@univ-lorraine.fr

Benoît Roussel
PErSEUs EA7312
Université de Lorraine
Metz, France
Email: Benoit.roussel@univ-lorraine.fr

***Abstract*—4D simulation (linking 3D models with time schedules) allows stakeholders to better understand the construction process of a building. However, 4D tools' recurrent usability issues contribute to its limited adoption. We used user-centered methods to better understand the nature of those issues and identify current practices and users' needs. Based on the gathered data, we framed and then conducted a creativity session with architecture, engineering, construction professionals and researchers, as well as construction software editors. Through the creativity process, users produced 46 ideas that led us to define and develop new functionalities for a new 4D prototype.**

Keywords- User-Centered Design; User needs; 4D BIM; Collaboration; Creativity.

I. INTRODUCTION

The design of a construction building is a complex activity that requires great coordination and a lot of information exchange from different disciplines (architects, engineers, contractors, project managers, etc). Given the fragmented nature of the construction market and information flows, close collaboration between the different stakeholders is essential for the construction process to meet the time, financial and aesthetic requirements of the owner. To that end, the planning and scheduling processes must be carried out in a rigorous and collaborative way. These two processes are complementary, although often confused [1]. Thus, in the construction field, planning consists in defining the objectives, modalities and resources of the project while scheduling makes it possible to allocate resources to tasks as well as to define their sequencing and the time necessary for their completion. [2]. Nowadays, this activity is still carried out manually and requires a large amount of individual and collective work to produce and exchange information [3]. However, a significant amount of information is lost during these exchanges, which results in low quality documents strongly impacting the project's planning and scheduling process [4]. Collaborative scheduling process mainly occurs during the "Design stage" and more precisely in the pre-construction phase. It is during this pre-construction phase that the teams work together on the definition of the construction process. This collaborative activity mainly occurs during coordination meetings, during which plans and diagrams from different disciplines are pooled to resolve design errors. These documents, in 2D or 3D, are digital or paper representations of the physical characteristics of the building and are therefore mediators in the process of collaboration and decision-making [5].

To meet this growing need for collaboration and information sharing, recent decades have seen the emergence of Building Information Modeling (BIM). The Associated General Contractors (AGC) of America defines BIM "as a data-rich, object-oriented, intelligent and parametric digital representation of the facility, from which views and data appropriate to various users; needs can be extracted and analyzed to generate information that can be used to make decisions and to improve the process of delivering the facility" [2]. BIM is, therefore, both a technology and a new work process requiring a reorganization of workflows. Documents and information centralization allow each actor to have access to all of the information throughout the entire project lifecycle. With the help of BIM tools, each stakeholder produces a 3D digital artifact representing a "business-oriented" vision of the building. These are then regularly updated and merged to form a centralized artifact. When applied to scheduling processes, BIM turns a 3D model into 4D. 4D artifacts are created by linking a 3D model with a planning, resulting in a model that visually simulates the construction process over time. Trough visualization of construction process, 4D offers many benefits such as work process optimization, rework reduction, increased errors detection, construction time reduction, and better communication among stakeholders [4]. Thus, 4D can be crucial to improve collaboration, decision making and reduce misinterpretations among teams.

While BIM's adoption and use in coordination meetings are growing, 4D's remain low. Despite its numerous advantages and potential uses, it is still mainly used as a visualization tool [6]. Researches have identified softwares' visualization issues as a barrier to 4D adoption at the individual level [7].

This paper is organized as follows: Section 2 gives an overview of related works on 4D's flaws and usability issues. Section 3 presents the creativity method, tools used and participants. Section 4 presents and analyzes the results of the creativity session. Section 5 concludes the paper and give an outlook of future work.

 ISBN: 978-1-61208-829-7

II. RELATED WORK

Some studies have sought to identify the usability issues of 4D softwares that limit their adoption. Guevremont and Hammad [6] and Castronovo, Lee, Nikolic and Messner [7] provide an overview of 4D issues. Through semi-structured interviews of AEC professionals, they found, among other issues, that there is a lack of visualization and interaction standards, such as common color coding, information filtering (i.e obtaining a precise information about a mechanical object), zooming to visualize precise spatial information or modifying the graphical level of detail. Both studies propose guidelines to enhance 3D models and schedule representation of 4D models.

Others researchers have been focusing on observing and analyzing the interaction between stakeholders and artefacts during coordination meetings from an ethnographic point of view [8].

In their study on the use of 3D and 2D digital, or paper, artifacts during coordination meetings, Mehrbod, Tory and Staub-French [8] identified the navigation between artifacts as the major problem. When solving a problem, actors spend a large part of their time navigating between 2D and 3D artifacts searching for more detailed 2D views, trying to obtain measurements and trade-specific information or even annotating documents. These navigation difficulties, as well as the lack of visualization and interaction standards, greatly slow down the collaboration and decision-making process during coordination meetings.

In summary, few studies have attempted to determine the AEC professionals needs about the use of 4D for scheduling purpose through the use of user-centered methods on all stages of the design process. Among them, none have described how users' needs have been used to generate and provide professionals with adapted solutions. In this paper, we present a collective creativity method used to generate functionalities that meet users' requirements.

III. SCOPE OF PROJECT AND METHODS

This paper describes part of a research project, 4DCollab [9], which involves the use of a collective creativity method to produce new 4D features. The project objective is to develop, through a user-centered methodology, a synchronous and co-located collaborative tool to help AEC professionals with site planning in the pre-construction phase. Using BIM and 4D technologies, the tool should improve communication, decision-making and information sharing between the various stakeholders. To tackle the identified issues, user needs were determined through a user-centered approach by first conducting semi-structured interviews with professionals in order to collect data on their individual and collective practice with 4D tools. Then, a multimodal analysis of speech gestures, exchanges and interactions around 2D, 3D and 4D documents during coordination meetings was performed. The data collected highlighted the main obstacles to the adoption of these technologies, as well as their main advantages. These data were then used to define the framework of a creativity session.

A. Means of the creativity process

In accordance with Parjanen [10], we define collective creativity as an approach of creative activity that emerges from the collaboration and contribution of many individuals so that new ideas are produced collectively by individuals connected by the common concern. Once the subject is defined (in our case, it is centered on user needs), the course of a collective creativity session is designed around an alternation of divergence-convergence. Thus, the creative process is structured in 4 main phases: an analysis phase, a divergent phase, a convergent phase and post workshop, a synthesis phase (Figure 1).

The objective of the divergent phase is to move away from the subject by diverging through exploratory reasoning. Its purpose is to produce new, unexpected, even crazy ideas. They can be lifted in this way, as their potential links with the subject have not yet been highlighted. The role of the convergence phase is then to bring the subject (user needs) and these ideas together in order to be able to respond to the problem at hand. Once this rapprochement is achieved, the term "crazy" ideas may then disappear and give way to "interesting" ideas to solve user needs. Our collective creativity action (during 2 continuous half-days of work) aimed to explore the theme of "new functionalities to share knowledge with others" and to formalize as many idea sheets as possible.

B. Participants

The working group was made up of 13 people (3 women and 10 men) from different professions: 2 architects, 2 computer scientists, 1 building construction professional, 1 researcher in architecture, 1 researcher in psychology, 2 mecanical engineers, 2 programmers in BIM, 2 software editors.

IV. RESULTS

The creative process was structured in 4 main phases : an analysis phase, a divergent phase, a convergent phase and a synthesis phase. The analysis phase was carried out using a purge tool (in our case, we used the mindmapping tool [11]. The purpose of the purge is to define the scope of the group's understanding of the initial subject. The purge resulted in a representation of ideas and concepts in the form of a Mind map which allowed the emergence of generic work themes. 14 thematic areas comprising a total of 68 items emerged.

The divergent phase opened up the initial topic by drawing from other areas concepts, notions and ideas that could later feed into the initial topic. Three tools were used: "Hot Potato" [12], Brainstorming [13], and then Analogy [12]. The initial questions were "What evokes for you the words : Compare, Appreciate, Confront, Bring together,...?" and "How to facilitate an instructive discovery in a city abroad ?". The convergent phases focus on returning to the initial subject by integrating the elements found in the divergent phase. It is during these phases, provoked at different moments of the creativity session, that the creativity group collectively brings out 38 embryos of ideas. An "idea embryo" is the first step of an idea explained by a member of the creativity group to the

 ISBN: 978-1-61208-829-7

other members in the form of a drawing. "Idea embryos" are most often generated during the phases of divergence and convergence.

The selection (by voting) and classification of ideas allowed the identification of 12 embryos of ideas that most caught the attention of the members of the working group. The drafting of 8 of the 12 idea sheets was carried out in groups of 2 or 3 people (Figure 2). A total of 46 idea sheets were produced, 8 of which were produced directly by the group during the creativity session.

For the 4DCollab project, idea generation (one of the results of collective creativity) is not an end in itself. A synthesis of the results was presented visually in a CK (Concept-Knowledge) Tree [14][15] (Figure 3). The classification "by families of Idea sheets" was carried out (after the session by the facilitator) with the formalization of C tree (concepts). This is intended to provide a vision of the links between the sheets produced as well as an overview of the fields explored (and not explored) by the group's production.

V. CONCLUSION AND FUTURE WORK

In this paper, we have discused 4D's issues that limits its use by AEC profesionnals and how user-centered methods can be used to resolve those issues. We have presented and defined a user-centered method, based on the principles of the collective creativity, that we used to generate new 4D functionnalities adapted to users' needs. From a quantitative point of view, 14 thematic areas (including 70 items) emerged during the analysis phase and were presented in a Mindmap. During the creative production phase, 38 ideas were generated. They open up new ways of solutions, complementary to existing "main stream" solutions. The collective creativity session was centered on user needs and its conception was organized around successions of divergence/convergence. The group of professionals participating in the session generated a total of 38 ideas. 12 of them were evaluated by them as the most interesting to meet the user needs of "new functionalities to share knowledge with others". Following this collective creativity session, some of these new functionalities were also evaluated as relevant from a business point of view. These were developed and implemented on the first version prototype, whose usabilility is being iteratively evaluated with the user testing method.

ACKNOWLEDGMENT

The authors acknowledge financial support from the Fonds Nationale de la Recherche (FNR), Luxembourg, and the Agence Nationale de la Recherche (ANR), France, to 4DCollab project, grant reference: 11237662 (LU) / ANR-16-CE10-0006-01(FR).

REFERENCES

[1] A. Baldwin and D. Bordoli, Handbook for construction planning and scheduling. John Wiley & Sons, 2014.

[2] S. A. Mubarak, Construction project scheduling and control. John Wiley & Sons, 2015.

[3] H. Liu, Z. Lei, H. X. Li, and M. Al-Hussein, "An automatic scheduling approach: building information modeling-based onsite scheduling for panelized construction", Construction Research Congress 2014: Construction in a Global Network, pp. 1666-1675, 2014.

[4] P. H. da Silva, J. Crippa, and S. Scheer. "BIM 4D no planejamento de obras: detalhamento, benefícios e dificuldades." PARC Pesquisa Em Arquitetura E Construção 10 (2019): e019010-e019010.M.

[5] M. Tory, S. Staub-French, B. A. Po, and F. Wu. "Physical and digital artifact-mediated coordination in building design," Computer Supported Cooperative Work (CSCW), vol. 17, no. 4, pp. 311-351, 2008.

[6] M. Guevremont and A. Hammad, "Criticality visualization using 4D simulation for major capital projects," Winter Simulation Conference (WSC), pp. 2360-2371), 2017.

[7] F. Castronovo, S. Lee, D. Nikolic, and J. I. Messner, "Visualization in 4D construction management software: a review of standards and guidelines," Computing in Civil and Building Engineering, pp. 315-322, 2014.

[8] S. Mehrbod, S. Staub-French, and M. Tory, "Interactions with BIM tools in design coordination meetings," Proceedings, Annual Conference—Canadian Society for Civil Engineering, Vol. 2, 2013.

[9] 4DCollab project web site. [Online]. Available from: https://www.4dcollab-project.eu/ 2020/10/16

[10] S. Parjanen, "Creating possibilities for collective creativity", Approches cognitives et ergonomiques, Ed. Acta Universitatis, 2012, ISBN 978-952-265-234-8.

[11] T. Buzan, B. Buzan, & J. Harrison, *The mind map book: Unlock your creativity, boost your memory, change your life*, Pearson BBC Active, New York, 2010.

[12] H. JAOUI, *La créativité mode d'emploi : connaissances du problème, applications pratiques*, ESF Edition, 2ième édition, 1996.

[13] A. F. Osborn, Applied Imagination: principles and procedures of creative thinking, Ed. Charles Scribner's sons, New York, 1953.

[14] A. Hatchuel and B. Weil, "C-K design theory: an advanced formulation", Research in Engineering Design, vol 19 no 4, pp. 181-192, 2008.

[15] T. Gillier and G. Piat, "Exploring over the Presumed Identity of Emerging Technology", Creativity and Innovation Management, Wiley, vol. 20 no 4, pp. 238-252, 2011, hal-00641765.

 ISBN: 978-1-61208-829-7

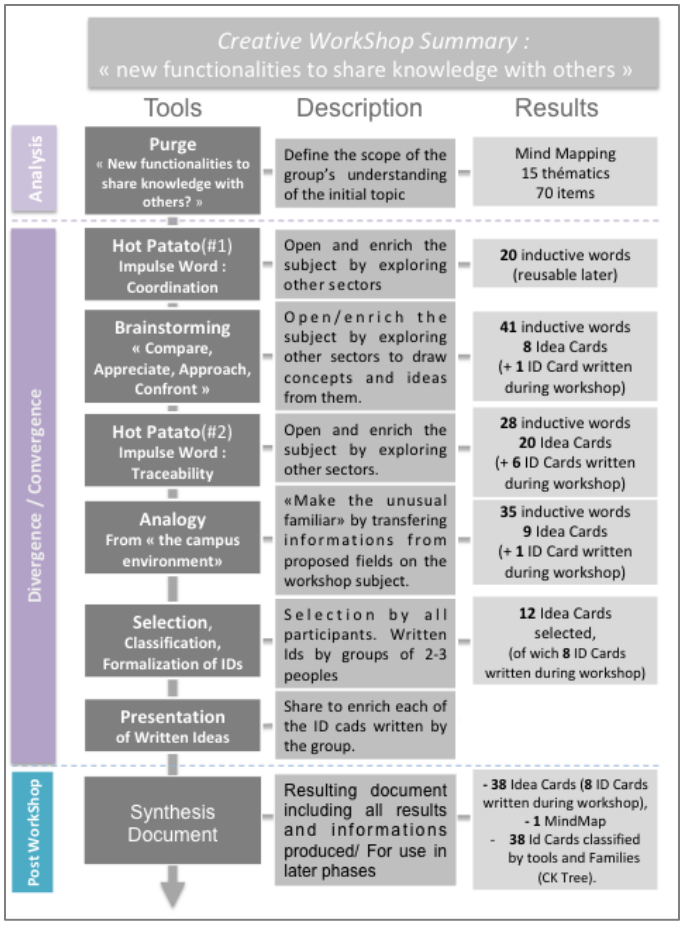

Figure 1. Visual summary of the creative process used for the workshop

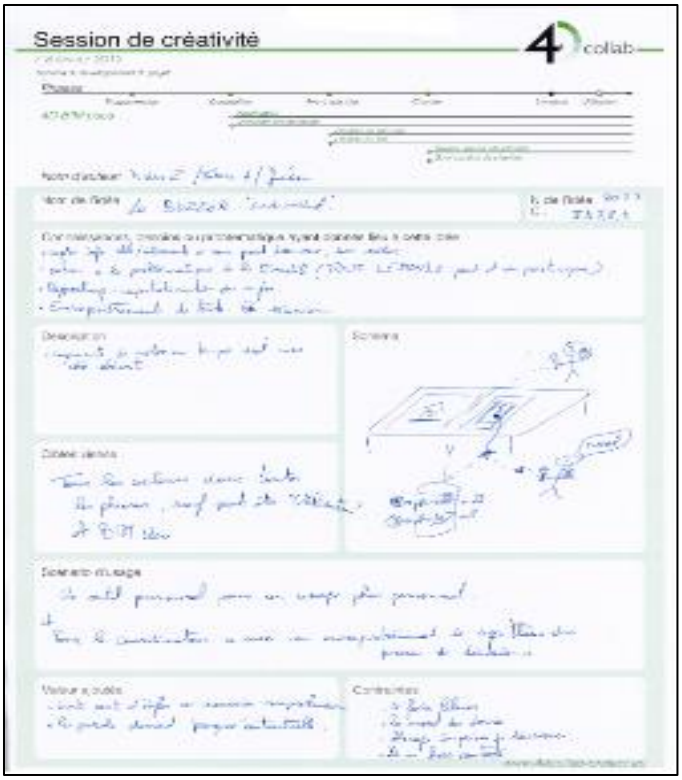

Figure 2. Example of Idea sheet (PAT 7-8-9) written by a group during the workshop (see d) in Fig.2)

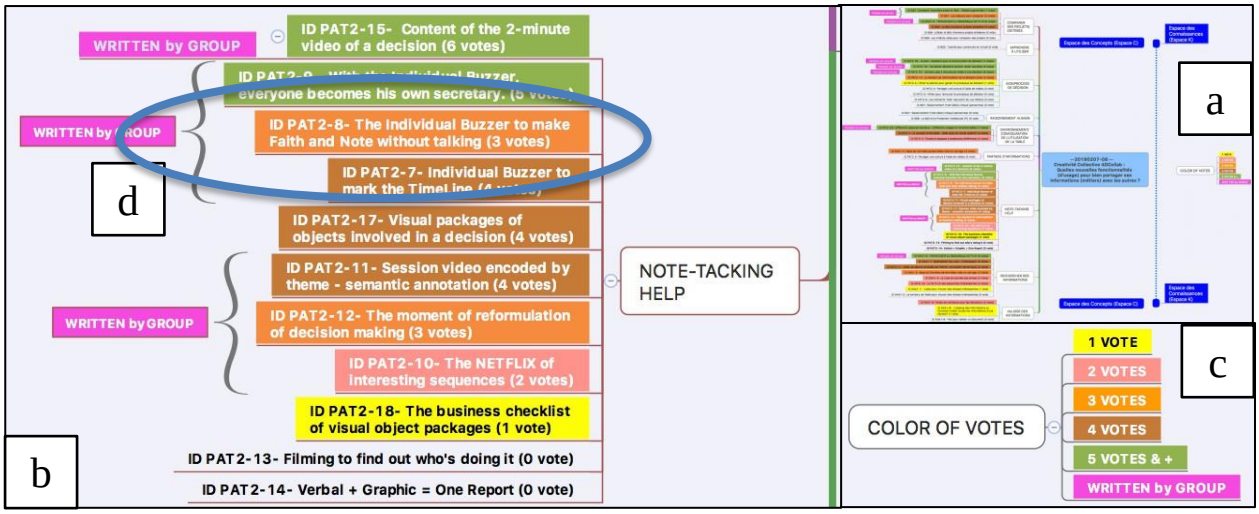

Figure 3. CK Tree : Classification "by families of Idea sheets" :
a) Global view of CK tree, b) Detail of a family, c) Caption, d) Example of Idea sheet written (detail in Figure 2)

 ISBN: 978-1-61208-829-7

Wizard-of-Oz Testing as an Instrument for Chatbot Development

An experimental Pre-study for Setting up a Recruiting Chatbot Prototype

Stephan Böhm, Judith Eißer, and Sebastian Meurer

CAEBUS Center of Advanced E-Business Studies
RheinMain University of Applied Sciences
Wiesbaden, Germany
e-mail: {stephan.boehm, judith.eisser, sebastian.meurer}@hs-rm.de

***Abstract*—Chatbots can be utilized to automate various business processes to add value for companies and users – for example, in the form of efficiency enhancement. Throughout the process of chatbot development, the integration of user feedback within a user-centered conversational design process is essential. In our study, we investigated chatbots in recruiting, a field within human resource management that is characterized by a high proportion of repetitive and standardized tasks. This pre-study applies a Wizard-of-Oz approach in which a basic dialog concept is tested in a very early phase of the project, simulating the chatbot functionality by a human operator. In this way, valuable user feedback on the general suitability of the dialog design can be gathered without coding chatbot functionalities. In total, eight users participated in our 60-minute experiment to conceptionally validate our idea and test the simulated Frequently Asked Questions (FAQ) chatbot. The research brought important insights into the basic concept and allowed us to collect new user intents not considered in the design. As a result, the tested concept proved to be suitable and of value for the users. Despite relatively long response times, only one participant suspected that they were not interacting with a chatbot but a human operator. The feedback on the user satisfaction with the completeness of the predefined answers and competence setup of the simulated chatbot was indifferent and rather moderate. However, most of the participants considered the tested scenario as relevant and stated a high user value for implementing the proposed chatbot in a recruiting process. Moreover, the Wizard-of-Oz approach generated appropriate input for improving the chatbot concept (e.g., intents, entities, criteria for satisfaction and acceptance enhancement) and valuable practical insights for developing a recruiting FAQ chatbot aligned to user needs.**

Keywords–chatbot; Wizard-of-Oz testing; prototyping; chatbot development; recruiting.

I. INTRODUCTION

Chatbots as a way to automate repetitive stakeholder (i.e., customers, prospects) inquiries in the form of conversational dialogues are more and more implemented into internal and external business communication processes [1][2]. In order to unfold their potential of enhancing the efficiency of such processes, it is imperative to create a suitable conversational design concept considering the envisioned users' requirements and expectations concerning this automation technology [3][4]. The integration of early user feedback is crucial in the development process [5]–[7], which makes it a common practice in technology development processes [3]. One way to yield stakeholder feedback and thus necessary input for the creation and advancement of a chatbot in an early stage without possessing a functional chatbot system is to conduct a Wizard-of-Oz (WOz) experiment [3][6]. In a WOz test, the executors lead the test subjects to believe that they are interacting with a fully developed technological system, whereas it is the test operators themselves acting as such, in this case serving as chatbot disguising their human form [8]. In our pre-study, on the practical example of a FAQ chatbot for recruiting [9][10], a WOz experiment was conducted within a broader chatbot user testing scenario in order to:

- evaluate the intent database of the developed recruiting FAQ chatbot prototype in terms of relevancy and answer suitability,
- collect feedback on the conversational design and specifically the (1) preliminary content, (2) the perceived user satisfaction, (3) the user's level of acceptance, and (4) utilization limitations, and
- yield not yet considered but relevant content in the form of novel chatbot intents, as well as potential training data for the chatbot.

This work in progress will first shed light on the theoretical background of chatbot prototyping followed by a discussion of Wizard-of-Oz testing in general, as well as the current state of WOz testing applied for chatbots in Section 2. The third section deals with the study approach in terms of the overall goal and the strategic, as well as technical set up of the WOz testing environment and framework. In the fourth section, we present preliminary findings of our pre-test and implications for practice before presenting the study's limitations and conclusions in Section 5.

II. THEORETICAL BACKGROUND

Iterative, user-centric design of chatbots is essential for good performance and to ensure the relevancy of the technology to the intended process of deployment. This section deals with the current state of chatbot development and the corresponding role of prototyping. Furthermore, it gives insights into the procedure of WOz experimenting and its application within chatbot development and research.

A. Chatbot Prototyping and Development

Chatbots are a kind of conversational interface [4] and belong to the field of Human-Computer Interaction (HCI)

 ISBN: 978-1-61208-829-7

research [11]. The need to involve users in the development process becomes apparent as it is the human users who need to see the overall relevancy of the technological system and be able to utilize it appropriately in order for it to add value. As per common practice (e.g., [5][12]), user testings are integrated into the system design process as an essential development step. Overall, there are many requirements to consider when developing a chatbot (see [13] for a multi-perspective overview), such as an adequate and useful reaction to input, behavioral appropriateness, and friendliness. Unlike graphical user interfaces, chatbot development is more difficult to separate interaction with the system from system functionality. Also, for chatbots, clickable dialog flows can be created and visualized for testing (e.g., [14]). However, such prototypes do not react directly to text input and, therefore, strongly abstract from the later usage scenario. Thus, technically, development requires already some sort of a development platform, high levels of programming skills and development experience [15] in order to build a functional chatbot prototype to be tested in a real-world scenario. Contentwise, the intent and response database is essential and determines the quality of the chatbot in the form of response appropriateness [15]. Hence, the creation of a suitable, encompassing intent list with an accompanying set of matching, relevant responses as an adequate reaction is crucial within chatbot development (e.g., [16]). Conversational interfaces can be seen as a progression from visual layout and interaction design [11]. They serve as an interface allowing for a dialogue with human users based on natural language entered by text input. As such, they leave little room for front-end user interface design as text input is rather static and not very variable [7]. Hence, it is the content itself [11] and the way of communication (e.g., chatbot personality [4]) that is in focus in chatbot conversational designing.

B. Wizard-of-Oz Experiments for Technological Innovations

The term Wizard-of-Oz originates from a story in a children's book by [17], in which one of the protagonists hides behind a curtain to control a scene from a remote, through which he can pretend to be a powerful wizard. A Wizard-of-Oz experiment, as coined by [18], is thus a simulation where the researchers interact with the users themselves in a concealed way while posing as a fully functioning technology whereas, in reality, the technological system is in a prototypical, incomplete state [19][20]. WOz studies are conducted to let the participants believe that they interact with a computer system processing natural language dialogues whereas in reality, they are not: a human, called wizard in this kind of experiment, mediates the conversation in order to circumvent the constraints of current technology and thus pretending to showcase an operating, sophisticated kind of technology [8]. The method is not new [6] but still represents a practical, resource-saving way of early user testing within the development process since no full-fledged prototype needs to be built for yielding first feedback. However, due to the integration of a competent, skillful wizard, a WOz scenario does not depict a fully realistic representation of the examined technology and is somewhat idealized so that it cannot be treated as a holistic testing approach -– it rather gives first ideas to build upon [3]. Especially in early stages of prototyping with incomplete functionalities, WOz experiments are advantageous as they resemble realistic, human-like conversational behavior and capable dialogue management as opposed to existing, potentially erroneous systems [21].

WOz studies are integrated into various fields of research and add value to technology development projects of all kinds. Complex technology, such as systems integrating Artificial Intelligence (AI) functionalities are especially well suited for this approach. The following section examines WOz testing in the specific domain of chatbot development.

C. Wizard-of-Oz Experiments for Chatbot Development

Wizard-of-Oz setups are applicable to various systems and architectures for testing before actual implementation [3]. The advantages of WOz experiments, such as the early user feedback on the system to have it comply closely to all relevant user requirements and the savings in (especially technical) resources, can also be exploited within chatbot development. As conversational systems conversing with human users in natural language, chatbots oftentimes encompass AI functionalities and are thus especially suited for WOz tests during the development process: The AI components can be mimicked without the necessity of sophisticated AI framework implementation. Within chatbot development, there are various aspects to consider in terms of technical, content, and design requirements, as presented in Section II-A. Alongside these prerequisites, there are certain restrictions concerning the creation of conversational systems: Chatbots are bound to predefined databases and thus predetermined input, which makes WOz-based prototype tests relevant to cover unexpected and thus non-considered content [22]. This is an ideal setup to assess first user perceptions of the preliminary conversational design while also allowing for new content compilation, which is in line with [8], who states that WOz studies can be utilized to gather data. In this study, the WOz experiment yields relevant intents and accompanying training, as well as test data for the chatbot prototype at hand to be implemented in the chatbot prototype as a next step.

The interface itself is predetermined as well in the form of a certain social media channel or messaging application as most common access point for chatbots [10]. Hence, the WOz framework needs to be integrable into this environment and must fit in a way that it cannot be distinguished from the expected fully developed chatbot. The WOz approach has commonly been applied to chatbot research (e.g., [20][21][23]–[25]). In the focused field area of chatbots for human resources, a few first studies exist as well (e.g., [26][27]). However, no study is known to the authors providing more detailed insights on the WOz framework and its implementation, as well as the findings generated for the user-centered improvement of a chatbot concept. This study seeks to close this gap.

III. Methodological Approach

The study at hand focuses on WOz testing for the simulation of an advanced chatbot. The chatbot is applied to the use case of answering FAQs on different topics and process steps within an electronic (i.e., web-based) recruiting process. In this section, the methodology of the study, including its goals, the chatbot concept, and the WOz framework, are presented.

A. Goals of the Wizard-of-Oz Study

There are three overarching goals of this ongoing study:

 ISBN: 978-1-61208-829-7

1) Intent matching and answer suitability assessment: As introduced, chatbot concepts can be simulated in a WOz testing environment. To get a real user feedback, the reactions of the wizard must reflect not only the functions but also the limitations of the intended chatbot. The wizard, therefore, does not answer freely but must follow predefined rules and settings. In our case, we did use the underlying content in terms of an initial intent set and corresponding predefined answer phrases developed during a previous project work on which the prototype is based on. Via the experiment, we tried to evaluate for which user inquiries the wizard could match an existing user intent to answer the user request, in which cases the wizard had to modify the answers, or no predefined intent was found at all, and thus a response had to be formulated based on the wizard's expertise. All in all, the completeness and suitability of our initial intent set should be assessed.

2) Conversational design evaluation: In addition, after setting up the first version of our recruiting FAQ chatbot, its (1) content, and (2) the experience with the chatbot in the specific application area of recruiting FAQ – assessed via the user's satisfaction and perceived usability, as well as the performance of the demonstrated solution – are studied by gathering user feedback via a qualitative (thinking aloud) as well as a quantitative (user survey) approach.

3) Intent generation: Besides testing of the topics already implemented in our recruiting FAQ chatbot concept, further information needs, and corresponding user intents need to be identified and integrated into the intent set. For acceptance reasons, this set must be extended to a point so that the chatbot provides relevant answers for the most prevalent questions. Apart from intent generation, potential training data can be derived from the WOz testing by the integrated collection and assignment of user input phrases to intents. However, to gain relevant amounts of data, this would require a larger scale of testing than in this pre-study. Furthermore, such use of WOz experiments might get more important and productive in later phases when the chatbot solution is implemented and needs to be trained. Training is necessary as the natural language understanding and intent matching components of advanced chatbots are based on (pre-trained) machine learning algorithms and thus rely on domain-specific training data to evolve and improve [28].

The WOz approach is utilized to test and validate the recruiting FAQ chatbot prototype from the corresponding perspectives as presented above. Based on the findings, the chatbot will be iteratively adapted, enhanced, and further developed.

B. Chatbot Composition and Configuration

This pre-study is part of the research project CATS (Chatbots in Applicant Tracking Systems, for further information see acknowledgment section) that focuses on the identification of value-adding chatbot use cases and implementation of chatbot functionalities in applicant tracking systems. The general relevance of the specific use case of a FAQ chatbot to support applicants and answer questions in the recruiting process was already the subject of previous research [29].

In order to satisfy the needs of the target group, intents were collected from different sources: (1) potential candidates on the verge of applying to a job were asked to walk through an application process in an applicant tracking system and to formulate questions on problems and challenges, (2) questions and answers in existing FAQs on websites on career websites and job portals were screened and consolidated, and (3) information inquiries from other channels (e.g., e-mail requests to employers with job offers) were collected. Moreover, recruiting experts were involved in reviewing and improving the resulting set of intents, suitable answers, and an initial set of example user questions (to make the intents easier to understand and as initial training data). In total, 113 intents have been identified to be included in the FAQ recruiting chatbot concept. This intent set with the accompanying answers has been utilized as the wizard's database throughout the experiment.

C. Study Design

The study at hand was designed to comply with the goals as defined in Section III-A: Intent matching in the form of answer fitness assessment, conversational design evaluation, and intent generation. The experiment resp. study design consists of four sub-sequential parts, as presented in Figure 1, and is described in the following paragraphs.

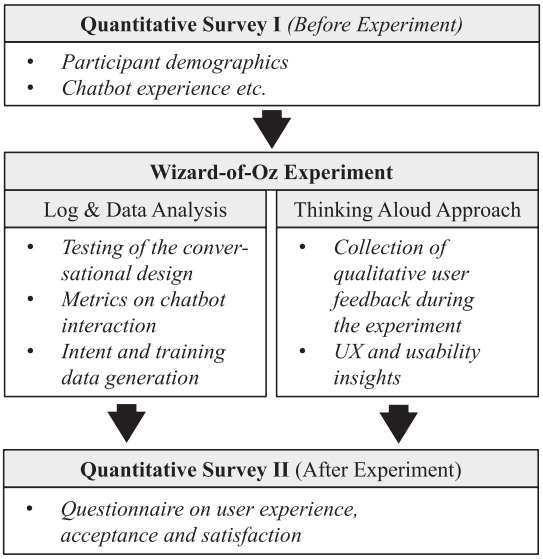

Figure 1: Flowchart on the Wizard-of-Oz Study Approach

1) Quantitative Survey I: Prior to the chatbot experiment, some socio-demographic characteristics of the participants (e.g., study program and qualification as filter-questions to confirm a required fit with the made-up job ads prepared for the study), as well as their experience with (recruiting) chatbots were surveyed.

2) Log and Data Analysis: At this stage, the WOz experiment started and the participants (students) were asked to apply for one of three pre-chosen open positions presented by job advertisements. The job ads were selected based on the qualification and skill profiles defined for the acquisition of study participants. As per the digital job advertisement landing page, the participants were free to consult the chatbot for any upcoming question or insecurity during their information and application phase up to the final application step of document and information submission. Even though the participants could have applied with their own documents due to their qualifications, application documents were provided for data protection reasons. All interactions were logged for the later data analysis.

 ISBN: 978-1-61208-829-7

3) Thinking Aloud Approach: The participants were asked to conduct a chatbot-supported application process in a thinking aloud approach, thus stating their thoughts, irritations, opinions, and actions while performing the task. The thinking aloud approach helps to gain relevant user experience (UX) insights and is a standard tool within user experience research [30]. Qualitative results are highly valuable for assumption and opinion validation and exploration of usability aspects [31].

4) Quantitative Survey II: After the participants have completed the WOz experiment and successfully submitted their application to the system, they were asked to answer a quantitative survey focussing on their satisfaction with the chatbot support and the corresponding user experience.

A moderator accompanied the participants through this process (on-site or remote) while one of the researchers posed as the wizard in the WOz framework; the details will be discussed in the following description of the WOz experiment setup.

D. Setup of the Wizard-of-Oz Experiment

Depending on the technological system, the WOz experiment concept needs to be integrated in a way that the wizard can operate covertly, which can be problematic for some setups [6]. However, the users must be led to believe that they are interacting with the technology itself for the WOz experiment to become successful and measuring the intended aspects. In the following, the WOz testing strategy and setup will be explained from conceptual, as well as from the technical perspective.

1) Wizard-of-Oz Experiment Concept: Maulsby et al. [31], who conducted a study on WOz testing with an automation agent, stress the importance of a strict behavioral plan for the wizard (they even recommend implementing an algorithm). This is important to maintain consistent behavior and, thus, experimental reliability [31]. In the four parts of the experiment as presented in Section III-C, several components had to be conceptualized: Required (1) roles, (2) documents, and (3) sequences.

In general, the following roles were assigned:

- *Participant:* The recruited chatbot users belonging to the target group of potential candidates, who converse with the chatbot during their application process.
- *Wizard:* A researcher belonging to the research project, who operates the WOz framework by sending preformulated messages or creating ad-hoc responses as seemingly AI-based automated answers from a separate room/on remote based on the experimental study framework.
- *Moderator:* Another researcher also belonging to the research project, who accompanies the participant through the experiment giving an introduction, instructions, and guidance through the process.

The participants were provided a set of application documents (CV, internship certificate, master's certificate) to allow for a realistic application scenario while maintaining privacy and data protection requirements. The moderator guided the participants through the whole process and was also responsible for writing down the participants' answers to the introductory and the conclusive quantitative questionnaires himself for a consistent moderator-participant experience. The first quantitative questionnaire was conducted after the moderator's introduction in the form of a brief explanation of the experiment and the according procedure and prior to chatbot utilization for first participant classification concerning their demographics. It consisted of five questions regarding their professional situation, their study program as well as their experience with online applications, chatbots, and recruiting chatbots in specific.

In the main part of the WOz experiment, the participants accessed a job search portal with three predefined job ads; they had to choose between. Upon making a choice, they were able to make any kind of inquiry to the alleged chatbot prototype, positioned in an embedded chat window in the lower right-hand side of the job ad landing page. They had to gather all information they presumed necessary for taking up an application and then actually apply via a specially configured testing application platform provided by the cooperating industry partner of the authors. During this process, the participants were once again told to make use of the chatbot whenever it felt necessary in situations of upcoming questions. The utilization phase ended after information and document upload upon submission of the application. Eight checkpoints had been established for further encouragement of chatbot utilization in the form of active requests to formulate every possible question coming to mind, but this method proved unsuccessful in the initial experiments and was perceived as rather interrupting concerning the overall procedure. For this reason, the checkpoints were removed from the study design, and feedback collected this way was not further considered in the study.

Throughout the phase of chatbot use and application in the system, qualitative user feedback was yielded via a thinking aloud approach. The participants were encouraged to articulate all upcoming thoughts, perceptions, and feelings towards the chatbot and their interaction with it. The quantitative survey after the WOz experiment, contained ten UX items (concerning the interaction via the interface not focusing on the design), questions concerning the participants' satisfaction with the answer quality (completeness, competency, and speed), the perceived added value from the chatbot support in general, as well as for each application step of the application process. The quantitative survey concluded with questions on the perceived (dis-)advantages concerning (1) any previous recruiting chatbot usage and (2) the FAQ chatbot prototype presented in the WOz experiment.

2) Technical Infrastructure for the Wizard-of-Oz Study: According to [3], the only components necessary for WOz testing are the interface software and databases. Correspondingly, the WOz framework was technically set up via Rocket.Chat [32], a free open source chat platform allowing for back-end and front-end chat interfaces for the wizard and the experiment participant. Moreover, this chat server system provided functions for storing data, i.e., logging the messages with additional information for later analysis (e.g., time-stamps). Rocket.Chat as chat server was chosen as it comes as an installation option with the Ubuntu server operating system and due to its simple handling and configuration without the need for advanced programming expertise. The Live Chat feature of the platform was utilized as a communication channel for the participants. The chat server was installed on a dedicated Ubuntu server.

 ISBN: 978-1-61208-829-7

Apache webserver was installed and configured for setting up websites required for the study. By using a JavaScript code snippet as advised by Rocket.Chat, a chat window, was integrated into an HTML document, which was then accessible by the participants via a web browser. The HTML document was also used to embed an IFRAME with a job search platform presenting the job ads. The job ads were linked to made-up career websites with access to a test installation of the applicant tracking system BeeSite [33] (operated on the servers of the cooperation partner Milch & Zucker AG) where the participants entered their data and completed the application process.

The participant's front-end configuration is presented in Figure 2. While operating in the career portal as shown on the left-hand side, the chatbot was accessible throughout the whole process as an overlay in the lower right corner (depicted on the right-hand side). This way, all upcoming problems in the form of questions or irritations could be directed to the chatbot from the participants. Messages sent by the respondents in the chat window were sent to and stored in Rocket.Chat. The chats can be accessed via a certain interface and saved as JSON files. Via JavaScript, the JSON data, were then converted into CSV data for further analysis and handling via Microsoft Excel. The researcher acting as wizard utilized the Rocket.Chat administration interface to receive and process incoming inquiries while posing as a chatbot.

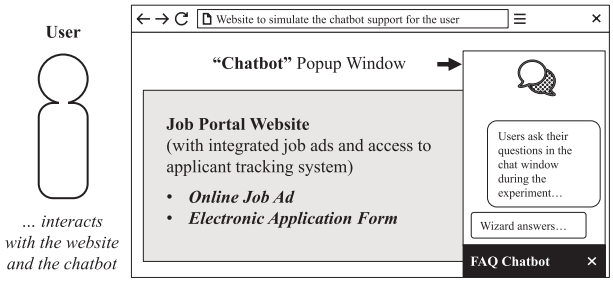

Figure 2: Wizard-of-Oz User Front-end Configuration

As shown in Figure 3, a special cockpit was designed within a web application for the wizard to either (1) choose from the predefined answer related to a specific intent considered in the predefined intent set, (2) take a predefined answer and modify it according to the unexpected input, or (3) enter answers in real-time to create novel, individual content for distribution to the participant.

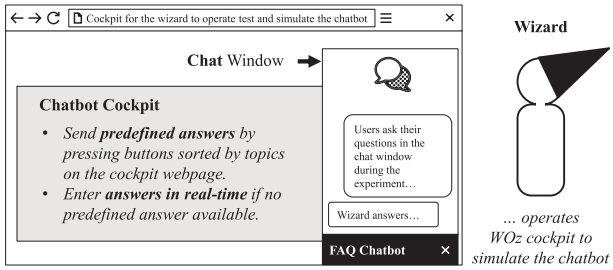

Figure 3: Wizard-of-Oz Wizard Front-end Configuration

Figure 4 shows the framework as procedure embedded into the overall study design, including the different roles, documents, and processes. With the servers hosting the Rocket.Chat chat environment and the career portal as central parts, the users accessed the framework from front-end perspective (left-hand side) while the wizard operated in secret from the back-end perspective imitating the expected FAQ recruiting chatbot (right-hand side).

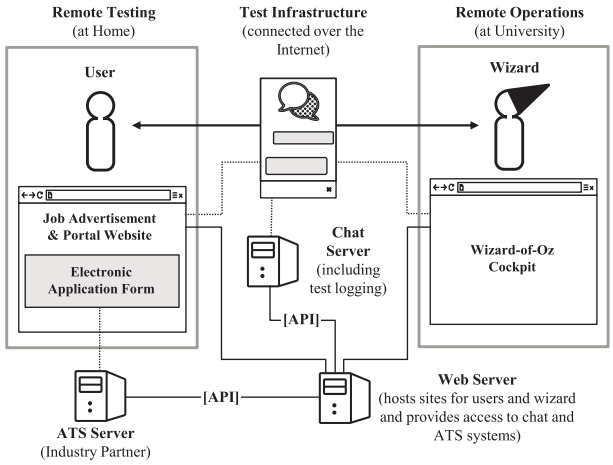

Figure 4: Wizard-of-Oz Framework of the Study

During the study, the setup had to be revised because of the outbreak of COVID-19. As a result, after an initial test with a first participant, the whole technical infrastructure had to be moved from a physical server of the laboratory in intranet (protected from access from the public internet by restrictive firewalls) of the university to (cloud) hosting providers in the public internet to allow all stakeholders in the form of the participants, the moderator, and the wizard to access the necessary interfaces remotely (without VPN access). For the moderator in the WOz experiment, who accompanied the experiment in a room together with the participants in presence mode, an adequate alternative had to be found to be able to perform this part of the experiment remotely. As a solution, Lookback [34] was identified. With this product, various user tests can be easily moderated and remotely performed. By means of Lookback, it was possible in the present experiment to guide the test participants to the already established test site (career portal incl. chat) and to accompany them during use. By integrating a video solution, the test participant and the moderator could stay in contact during the experiment.

IV. Preliminary Findings and Implications

A. Metrics on Chatbot Interaction

In total, eight users actively participated in the WOz experiments. One user did not use the chatbot as he did not required or considered support within the application process and was thus excluded from the analysis on the chatbot interaction in this and the next chapter. Another participant had to be excluded from this section analyzing the metrics, as changes in the setup of the WOz environment were required, as described in the previous section.

The remaining six participants interacted with the wizard in 79 chatbot sessions (cf. Table I). A session describes here a coherent sequence of interactions between chatbot (i.e., the wizard) and the user associated with a single user intent. The

ratio of chatbot interactions per session (column c) varied between 1.00 and 1.43, with a mean value of 1.21. This evaluation shows, first of all that the activation for use and intensity of use varied greatly among the participants in the study. Moreover, it becomes evident that some respondents expected to get a prompt answer (comparable to a search request on a website) where others got more involved in an interactive dialog with the chatbot to get the intended information.

The wizard's response behavior in the experiment is also shown in Table I in columns (d) to (f). As described earlier, the wizard had three different response options to user queries in the experiment. More than half of all answers by the wizard (55; 63 percent) were given by predefined answers via the button option in the wizard cockpit (d), only in four cases (5 percent) the predefined answers were modified by the wizard (e). For about one-third of the user requests (28), there was no matching intent, and so the answer had to be formulated by the wizard (f). In a productive mode with a chatbot implemented based on the given concept, the questions with no matching intent could not have been answered. Two participants took advantage of the opportunity to be forwarded by the chatbot to a human contact person to answer a question, but only once each (human hand-over (g)).

TABLE I: Wizard-of-Oz Experiment Metrics (Absolute Values).

(#)	(a) Chatbot sessions	(b) Chatbot interactions	(c) Interactions per session	(d) Wizard answer via button	(e) Wizard answer edited	(f) Wizard answer free	(g) Human hand-over request
(1)	16	21	1.31	10	1	10	1
(2)	7	7	1.00	4	0	3	n.a.
(3)	12	17	1.42	6	0	8	1
(4)	13	14	1.08	8	2	2	n.a.
(5)	23	33	1.43	20	1	4	n.a.
(6)	8	8	1.00	7	0	1	n.a.
Sums	79	100	–	55	4	28	2
Means	13.2	16.7	1.21	9.2	0.7	4.7	0.3

As a next usage metric, the average response times required by the Wizard were recorded in this stage of the experiment (see Table II). Not surprisingly, the average response times of the wizard were lowest for the predefined answer buttons (column a), at an average of 20 seconds. Here, the wizard had to capture an incoming user request, then search the chatbot cockpit with keywords to find a matching intent/answer pair in the list and send the corresponding answer to the chat. The wizard in the experiment took noticeably longer (34 seconds on average) for the answer option (column b), where a slight adjustment of the predefined answers available in the intent/action list was made. Only marginally shorter, average response times were achieved for the option of free answers written by the wizard. Here, an average of 32 seconds passed between receiving the user inquiry and posting the answer in the chat (column c). Across the various response options and the participants in the experiment, the wizard took an average of 25 seconds to answer an user inquiry.

Overall, it can also be recognized that the response times were significantly longer than it could be expected from an automated answering system. However, the participants in the study were briefed in such a way that it is a test system with yet limited performance.

TABLE II: Wizard-of-Oz Experiment Mean Answer Times (In Seconds)

(#)	(a) Wizard answer via button	(b) Wizard answer edited	(c) Wizard answer free	(d) Overall
(1)	19	27	22	21
(2)	16	n.a.	23	19
(3)	26	n.a.	45	37
(4)	20	44	27	25
(5)	18	32	36	21
(6)	23	n.a.	39	25
Means	20	34	32	25

B. Quantitative User Experience Survey

After the experiment, the participants were asked for quantitative feedback in the form of a short user survey, e.g., with regard to satisfaction ratings on selected topics in the field of user experience.

Table III shows a summary of the result of this survey. The table shows the survey results for the seven participants that interacted with the simulated chatbot. It can be seen that the satisfaction with regard to the answer completeness is rather indifferent and moderate. Three of the seven participants were rather satisfied, another three partly satisfied, and even one not satisfied at all. The quality of the answers and thus the perceived competence of the chatbot is another important evaluation criteria in the WOz experiment. As shown in Table III, two of the seven participants stated that they considered the answers they received as rather competent. The remaining five participants were moderately satisfied only. Not surprisingly, the satisfaction rating with the chatbot performance, i.e., the speed of the chatbot answers, turns out to be recognizably poor, which is of course also due to the character of the WOz project: Since the chatbot is only simulated by a human, the person needs time to record and process the questions and to write the answers. This finding does not seem to have influenced the perception of the general added value of chatbots. This is probably due to the fact that the test persons were aware of the test situation and performance limitations. Six of the seven participants consider the tested use case as relevant and the general added value in applicant support of such a FAQ chatbot as (very) high.

TABLE III: User Experience Evaluation of the Chatbot Prototype

(Absolute Values; N = 7)	completely satisfied	rather satisfied	moderately satisfied	rather not satisfied	not at all satisfied
Answer Completeness	0	3	3	1	0
Competence	0	2	5	0	0
Speed and Performance	0	1	1	2	3
	very high	rather high	moderate	rather low	very low
General Added Value	2	4	0	1	0

Beyond the general added value, the test persons were also asked to evaluate the added value of the presented FAQ chatbot

 ISBN: 978-1-61208-829-7

in the individual process phases of the application. As shown in Figure 5, the three areas to which the participants attribute the greatest added value are answering questions about the job advertisement, questions about registration and general questions on the application process, and the further procedure after the application.

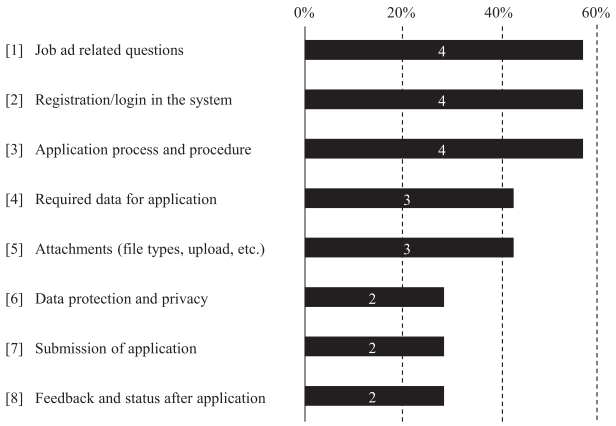

Figure 5: User Assessment on the Added Value of the FAQ Recruiting Chatbot Prototype (Multiple Answers Possible).

The last part of the quantitative survey contained questions on the user experience and usability of the chatbot as perceived by the participants in the WOz experiment. As shown in Figure 6, all participants perceived the FAQ chatbot simulated in the WOz scenario as easy to use. Furthermore, all but one participant agreed that potential applicants quickly learn to use such a chatbot. Four of the seven participants could still imagine using such a FAQ chatbot on a regular basis, following the example of the one used in the experiment during the application process. The other answers regarding the perceived technical complexity, quick learnability, or data security show that chatbot systems like the one presented can be quickly adopted and easily mastered. However, feedback on the integration of the chatbot and inconsistencies in the answers indicate a potential for improvement.

C. Qualitative User Feedback

In the following, some important observations and findings from the thinking aloud approach will be summarized. The most apparent problem was the long latency times caused by the human wizard simulating the chatbot. However, only one of the participants suspected that the delay might result from a human acting as a counterpart in this experiment. Due to the delay, some participants started to adapt their asking behavior by reformulating inquiries or reducing the number of questions to prevent further waiting frustration. The response times in the WOz experiment are, therefore, clearly too long. In future experiments, response times must be reduced. This could be done by optimizing the wizard cockpit, integrating a recommender system to pick answers (instead of searching for answers in the cockpit), or the integration of language-to-text interface to avoid typing in text for free answers. It should be noted, however, that this finding is more a problem of simulation than of the actual chatbot concept.

A more substantive observation concerns the complexity of the questions. The solution intended for implementation of the chatbot does not support the identification of multiple intents in a single user prompt. For a realistic scenario, the chatbot did answer questions one at a time. Ignoring question portions led to misunderstandings and confusion among individual participants. In a later implementation, solutions must be found to identify such problems and provide users with appropriate feedback to simplify questions. Another frequent remark was the perceived superficiality of several chatbot answers, which overall did not satisfy the users but rather frustrated them and was perceived as inept in some instances. This indicates a need for improving the intent set used in the experiment, as well as the corresponding answers. For chatbot implementation, response quality and relevance to a user might be improved by integrating the usage context. For example, responses could be personalized by processing information about the applicant already entered in the applicant management system. It has also become clear that a chatbot system must be able to distinguish between requests that can be answered with standardized information or with advice from a human contact person. One participant suggested human hand-overs for important questions and leaving the chatbot for rather simple inquiries. As discussed in the previous section, however, it is not to be expected and occurs rather rarely that the users of a chatbot themselves request such an offered option for a hand-over to second-level support.

Although there was little to criticize by the users regarding the usability of the chatbot, and there is little scope for design, individual possibilities for improvement were identified. A typing indicator was not implemented in the WOz front-end and was reported missing by the participants. Such an indicator can show that the request is being processed on the other side and shall be included in the future version of the WOz setup, as well as in the real chatbot if significant processing time would occur. There were also several helpful remarks regarding the positioning of the chatbot: Some participants felt

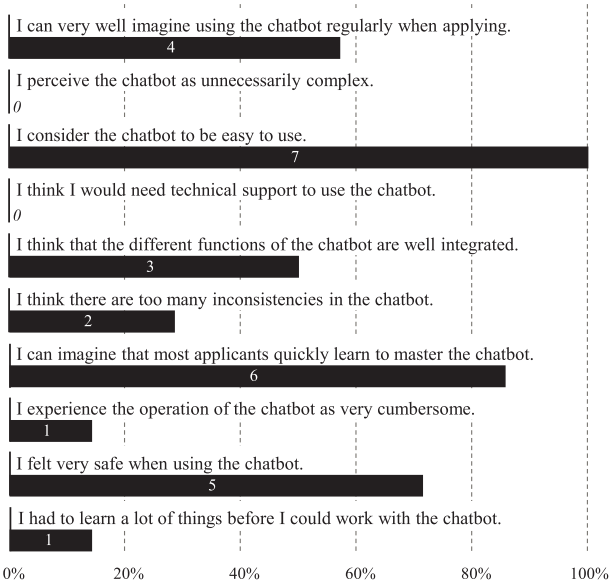

Figure 6: User Experience Rating of the FAQ Recruiting Chatbot Prototype (Sum of Response Options ”Totally Agree” and ”Rather Agree”).

it was partly hidden behind the banner informing about the website's cookies, and one participant did not recognize the chatbot button when retracted into the small starting button at the beginning of the experiment.

In total, it can be said that from candidate-sided user perspective, there are several aspects in the WOz framework that need enhancing and further development prior to continuation of the experiment, as well as consideration in the envisioned chatbot prototype. Consistent with Maulsby et al. [31], the authors learned a lot about required improvements of the chatbot concept by posing as wizards during the experiment. That way, not only the feedback of the participants can be integrated, but also the researchers' perspective can be considered through their role as back-end chatbot operators. About the content and scope of the chatbot, we learned that several topics and questions, for example, concerning the career portal, have not yet been considered and need to be included for a more comprehensive intent set of the envisioned chatbot.

V. Conclusion and Limitation

This study has demonstrated that WOz experiments can be used in the early phases of system development to validate concepts for chatbots. Appropriate infrastructures are to be implemented with a manageable amount of resources based on existing open-source web and chat server solutions. In such WOz setups, participants can be credibly convinced to interact with a real chatbot. However, the time needed to select appropriate answers is problematic if the restrictions of the chosen chatbot design are to be maintained in the experiment, and the wizard should not simply answer freehand. The experiment has also shown that users do not automatically accept support offered by a chatbot and do not necessarily enter into a more comprehensive dialogue with such a system.

The findings of the study indicate that the implementation of FAQ chatbots in application processes is seen by the participants as easy to master and valuable. However, it is important during implementation that the chatbot is actively promoted and indicated on the respective website. When interacting with a user, the chatbot must not only provide suitable answers for questions but also need to point out necessary simplifications in case of complex inquiries or even take the initiative to offer a hand-over to second level support by human experts.

WOz experiments can also provide important insights into the required content and scope of the chatbot concept. While most user questions could be handled by predefined answers from a given intent set that reflected the current status of the chatbot concept, more than 30 percent of the user inquiries in the experiment had to be answered freehand by the wizard showing a need to extend the intent set to the topics not covered yet. This is supported by the findings of the quantitative survey on the answer completeness that was not fully satisfying and thus needs to be improved. The perceived superficiality of the chatbot answers is another quality-related problem of the chatbot concept identified in the experiment that indicates further improvements of the intent/answers sets.

Certain limitations need to be taken into considerations: Our findings are based on a WOz study with a very small sample of eight participants only. However, in early phases of development and in studies focusing more on general feasibility and usability than generalizable results, small groups of test persons are quite common [35]. More critical, however, are the statements in our study about the added value or usefulness of the presented solution, which must definitely be verified by surveys with more participants. Another inaccuracy with regard to the implementation of the chatbot concept is if the intent matching performance of the wizard can be achieved by today's chatbot platforms available in the market. While the coverage of user inquiries and the responses of the chatbot were realistically limited by the intent set, intent matching may still vary considerably in a later implementation, which may influence user satisfaction as well. In general, for WOz experiments, maintaining consistent wizard behavior and the incapability to simulate errors or suboptimal system performance are limiting aspects of studies of this kind [6].

In future research, the authors can profit from the insights of this pre-study by optimizing the chatbot infrastructure or utilizing a hybrid approach, as suggested by [28]. Such a hybrid approach could be implemented, for example, by integrating a functional chatbot prototype into the WOz framework and limit the scope of human intervention to areas where the chatbot does not respond appropriately to inquiries. Other future studies might look into the field of speech-based dialogue systems in the form of voice assistants, predicted to be the even more efficiency enhancing and generally next logical step after establishment of text-based chatbot solutions [10].

Acknowledgment

The study was carried out as part of the research project CATS (Chatbots in Applicant Tracking Systems) of the Rhein-Main University of Applied Sciences. This project (HA project no. 642/18-65) is funded in the framework of Hessen Modell-Projekte, financed with funds of LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, Förderlinie 3: KMU-Verbundvorhaben (State Offensive for the Development of Scientific and Economic Excellence).

References

[1] L. Schildknecht, J. Eißer, and S. Böhm, "Motivators and barriers of chatbot usage in recruiting: An empirical study on the job candidates' perspective in germany," *Journal of E-Technology*, vol. 9, no. 4, pp. 109–123, Nov. 2018.

[2] U. Gnewuch, J. Feine, S. Morana, and A. Maedche, "Soziotechnische gestaltung von chatbots (socio-technical design of chatbots)," in *Cognitive Computing*, Springer Fachmedien Wiesbaden, 2020, pp. 169–189.

[3] D. Jurafsky and J. H. Martin, *Dialog systems and chatbots*, 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, 2018.

[4] M. F. McTear, "The rise of the conversational interface: A new kid on the block?" In, J. F. Quesada, Francisco-Jesús, M. Mateos, and T. L. Soto, Eds., Berlin: Springer International Publishing, 2017, pp. 38–49.

[5] J. Nielsen, "The usability engineering life cycle," *Computer*, vol. 25, no. 3, pp. 12–22, Mar. 1992.

[6] S. Schlögl, G. Doherty, and S. Luz, "Wizard of oz experimentation for language technology applications: Challenges and tools," *Interacting with Computers*, vol. 27, no. 6, pp. 592–619, May 2014.

[7] S. Böhm *et al.*, "Intent identification and analysis for user-centered chatbot design: A case study on the example of recruiting chatbots in germany," in *The Thirteenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, CENTRIC 2020*, (in press), 2020.

 ISBN: 978-1-61208-829-7

[8] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of oz studies," in *Proceedings of the 1st international conference on Intelligent user interfaces - IUI '93*, ACM Press, 1993, pp. 193–200.

[9] B. Hmoud and V. Laszlo, "Will artificial intelligence take over human resources recruitment and selection," *Network Intelligence Studies*, vol. 7, no. 13, pp. 21–30, 2019.

[10] L. Dudler, "Wenn bots übernehmen – chatbots im recruiting (when bots take over – chatbots in recruiting)," in *Digitalisierung im Recruiting*, T. Verhoeven, Ed., Wiesbaden: Springer Gabler, 2020, pp. 101–111.

[11] A. Følstad and P. Bae Brandtzæg, "Chatbots and the new world of HCI," *Interactions*, vol. 24, no. 4, pp. 38–42, Jun. 2017.

[12] T. K. Landauer, *The Trouble with Computers*. Cambridge, Massachusetts: The MIT Press, 1996.

[13] N. Tavanapour and E. A. Bittner, "Automated facilitation for idea platforms: Design and evaluation of a chatbot prototype," *Thirty ninth International Conference on Information Systems*, pp. 1–9, 2018, San Francisco.

[14] Botsociety, *Design chatbots and voice experiences*, 2020. [Online]. Available: https://botsociety.io/ [retrieved: 07/16/2020].

[15] S. A. Abdul-Kader and D. J. Woods, "Survey on chatbot design techniques in speech conversation systems," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 7, pp. 72–80, 2015.

[16] S. Ghose and J. Joyti Barua, "Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor," in *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, IEEE, May 2013, pp. 1–5.

[17] F. L. Baum, *The Wonderful Wizard of Oz*. Chicago: George M. Hill, 1900.

[18] J. F. Kelley, "Wizard of oz (woz): A yellow brick journey," *Journal of Usability Studies*, vol. 13, no. 3, pp. 119–124, 2018.

[19] R. Eynon and C. Davies, "Supporting older adults in using technology for lifelong learning," *Proceedings of the 8th International Conferenceon Networked Learning*, pp. 66–73, 2012.

[20] J. Eißer and S. Böhm, "Hedonic motivation of chatbot usage: Wizard-of-oz study based on face analysis and user self-assessment," in *The Tenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, CENTRIC 2017*, 2017, pp. 59–66.

[21] L. El Asri *et al.*, "Frames: A corpus for adding memory to goal-oriented dialogue systems," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Association for Computational Linguistics, 2017, pp. 207–219.

[22] F. Guerin, "Learning like a baby: A survey of artificial intelligence approaches," *The Knowledge Engineering Review*, vol. 26, no. 2, pp. 209–236, May 2011.

[23] W. R. Kearns *et al.*, "A wizard-of-oz interface and persona-based methodology for collecting health counseling dialog," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems: 1-9*, ACM, Apr. 2020.

[24] L. Riek, "Wizard of oz studies in HRI: A systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 119–136, Aug. 2012.

[25] S. Quarteroni and S. Manandhar, "A chatbot-based interactive question answering system," in *Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, 2007, pp. 83–90.

[26] M. X. Zhou, C. Wang, G. Mark, H. Yang, and K. Xu, "Building real-world chatbot interviewers: Lessons from a wizard-of-oz field study," in *Joint Proceedings of the ACM IUI 2019 Workshops*, 2019, pp. 1–6.

[27] R. Kocielnik, D. Avrahami, J. Marlow, D. Lu, and G. Hsieh, "Designing for workplace reflection," in *Proceedings of the 2018 Designing Interactive Systems Conference*, ACM Press, 2018, pp. 881–894.

[28] J.-W. Ahn *et al.*, "Wizard's apprentice: Testing of an advanced conversational intelligent tutor," in *Tutoring and Intelligent Tutoring Systems*. Nova Science Publishing, 2018, ch. 12, pp. 321–340.

[29] S. Meurer, S. Böhm, and J. Eißer, "Chatbots in applicant tracking systems: Preliminary findings on application scenarios and a functional prototype," in *In Böhm, S., and Suntrayuth, S. (Eds.): Proceedings of the Third International Workshop on Entrepreneurship in Electronic and Mobile Business*, (in press), 2019, pp. 209–232.

[30] W. Albert and T. Tullis, *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Waltham, Massachusetts: Morgan Kaufmann, 2013.

[31] D. Maulsby, S. Greenberg, and R. Mander, "Prototyping an intelligent agent through wizard of oz," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, 1993, pp. 277–284.

[32] Rocket.Chat, *The ultimate communication hub*, 2020. [Online]. Available: https://rocket.chat/ [retrieved: 07/16/2020].

[33] M. Z̃. AG, *Beesite recruiting edition – job posting applicant management talent pools*, 2020. [Online]. Available: https://www.milchundzucker.com/products/beesite-recruiting-edition-job-posting-applicant-management-talent-pools/ [retrieved: 07/16/2020].

[34] Lookback, *Talk to your users: See how they're using your app or website.* 2020. [Online]. Available: https://lookback.io/ [retrieved: 07/16/2020].

[35] J. Nielsen, *How many test users in a usability study?* 2012. [Online]. Available: https://www.nngroup.com/articles/how-many-test-users/ [retrieved: 07/16/2020].

 ISBN: 978-1-61208-829-7