Why Do Chatbots Fail?
- An Analysis of Critical Success Factors

Masterarbeit

zur Erlangung des akademischen Grades „Master of Science (M. Sc.)“ im Studiengang Wirtschaftsingenieur
der Fakultät für Elektrotechnik und Informatik, Fakultät für Maschinenbau und
der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: Grützner Vorname: Lukas

[redacted]

Prüfer: Prof. Dr. M. H. Breitner

Hannover, den 03.06.2021

Table of contents

1. Introduction

1.1 Motivation and relevance

Continuous technological progress is driving digital transformation. Technologies such as artificial intelligence (AI) have become widely adopted, but are not yet fully mature (Uren 2020, p. 1). With advances in AI, the ability to manipulate natural language text or speech by computers, i.e., natural language processing (NLP), has improved significantly (Adamopoulou and Moussiades 2020a, p. 373). Accordingly, interaction behavior of humans with computers have changed, causing new challenges for human computer interaction (HCI) research (Chaves and Gerosa 2020, p. 729), as user interaction is critical to acceptance and success of chatbots (Adam et al. 2021, p. 2). With user acceptance also on the rise (Følstad and Skjuve 2019, p. 1, 7; Sanny et al. 2020, p. 1225), chatbots have gained vast popularity in research and practice in recent years (Adamopoulou & Moussiades 2020, p. 374 - 375; Diederich et al. 2021, p. 1). Gartner predicts that by 2022 70% of all office workers will interact with chatbots on a daily basis (Gartner 2019). Further, the chatbot market is expected to grow from $17.17 billion in 2020 to $102.29 billion in 2026 (Mordor Intelligence 2021). Combined with a huge increase in scientific publications on chatbots (Zierau et al. 2020, p. 5; Adamopoulou & Moussiades 2020, p. 374 - 375), the high relevance of the field is indicated.

Chatbots are mainly Internet-based software systems interacting with humans within a simulated conversation (Feine et al. 2019, p. 138), and are implemented to automate redundant processes in a wide range of domains such as customer support (e.g., Zierau et al. 2020b), healthcare (e.g., Prakash and Das 2020), or banking (e.g., Richad et al. 2019). They can be leveraged to provide continuous availability, increase efficiency, or minimize customer support costs, for example (Richad et al. 2019, p. 1271; Adamopoulou & Moussiades 2020, p. 375 - 376). In recent years, there has been a large amount of research in various disciplines, especially in the information system (IS) and human-computer interaction (HCI) fields, that has investigated various design elements and related impacts on human behavior (Zierau et al. 2020a, p. 2; Diederich et al. 2021, p. 2). For instance, Janssen et al. (2020, p. 217, 220) provide an overview of different chatbot design elements, e.g., intelligence level or additional human support, and identified five archetypes. The importance of social characteristics in human - chatbot interaction is highlighted by Chaves and Gerosa (2020, p. 729). Conversational intelligence characteristics such as proactivity and conscientiousness can be used to influence social characteristics such as perceived personality or damage control (Chaves and Gerosa 2020, p. 769). The influence of trust and privacy concerns, as well as user perceptions of usefulness and ease of use in relation to users' intentions to use insurance chatbots, are also examined (Rodríguez Cardona et al. 2021, p. 556, 562). With perceived usefulness having the highest positive impact, chatbot functions that add practical value to a digital customer experience are most important (Rodríguez Cardona et al. 2021, p. 564).

Although many individual aspects of chatbots have been studied, and the general availability of technology and tools is increasing substantially (Galitsky 2019, p. 27), there is a significant fluctuation rate in developed chatbots. Despite this issue, to date there has been little qualitative research exploring the exact reasons why organizations take their chatbots offline permanently or stop maintaining them. In scientific literature, individual perspectives can be found, such as single case field reports from practitioners on challenges in deploying chatbots (Fiore et al. 2019, p. 9), attributes affecting failure or success based on user surveys (Følstad et al. 2018b, p. 13 - 14; Weber and Ludwig 2020, p. 324 - 325), and design principles (Winkler and Roos 2019, p. 6) as well as requirements to be considered for improving chatbots (Johannsen et al. 2018, p. 3 - 4). A holistic view considering all these technical, human behavioral, and institutional aspects and perspectives on chatbots' failures and success is not yet available. We address this research need on a global scope at the highest hierarchical level which concerns any organization across different domains, i.e., domain superior, (Bullen and Rockard 1981, p. 19 - 20), leading to our following research questions:

RQ1: What are qualitative reasons for chatbots' failures in practice?

RQ2: What are the domain superior critical success factors for chatbots?

The first research question is designed to gather knowledge about chatbots' failures in practice and to cluster it in different causes. This knowledge serves as a starting point for the second research question. Starting from the reasons for failure, CSFs for chatbots are examined.

1.2 Research procedure

Within the "gestalt mode" of HCI (Adam et al. 2021, p. 3), we apply design science research (DSR) to explore reasons for chatbots' failures in practice and identify the critical factors (CSFs) determining chatbots' success. This thesis is structured as followed in order to be able to answer the two research questions posed in the context:

First, theoretical background on chatbots and CSFs is presented. We provide a definition and brief subdivision of chatbots, state the history of chatbots, and explain basic technical concepts. Furthermore, we give a definition and overview of CSFs, describe the history and development of CSFs and present CSFs in other IS and HCI research fields. A brief overview of further success models in IS research is also given. In the third chapter we describe research design, methodology and methods used. Within our three cycle DSR process, adopted from Hevner (2007, p. 88) and supplemented by the recommendations of Vom Brocke et al. (2020a, p. 7), we develop and present our research artefact in nine steps. In the beginning, we identify our research problem by analyzing 103 real-world chatbots and adopt CSFs as our solution approach. Drawing on an extensive literature review of 154 scientific papers and 20 expert interviews, we explore qualitative reasons for chatbots' failures in practice. Next, we develop 12 domain superior CSF, outline their interrelationships based on Hilbert's (2005, p. 232) CSF framework, and evaluate them in the context of a focus group discussion (FGD). In our

discussion section, we reflect the obtained results, highlight parallels to findings from other HCI and IS research areas and give implications for research and practice. Lastly, limitations are outlined and an outlook for future research is given.

2. Theoretical background chatbots

2.1 Chatbot definition and categories

Chatbots, also known under the term conversational agents (CA) (Zierau et al. 2020c, p. 2), are mainly Internet-based software systems that interact with humans within a simulated conversation to perform tasks (Feine et al. 2019, p. 138). Chatbot are therefore computer systems which provide an interface between human users and software applications, using spoken or written natural language as the main medium of communication, some being able to perform beyond limited sets of predefined commands and answers by actual understanding content (Galitsky 2019, p. 13).

Chatbots can be categorized according to different elements. Starting from the knowledge domain, chatbots can be broken down and classified using taxonomies (Janssen et al. 2020; Zierau et al. 2020a). Chatbots can be divided into open domain knowledge and domain specific knowledge. Domain knowledge describes the knowledge a chatbot is capable of accessing, and the amount of data it has been trained with (Adamopoulou and Moussiades 2020a, p. 377). Open domain chatbots draw on a large knowledge base and might access additional databases and search engines and can accordingly talk about a diverse range of topics and answer accordingly. Domain-specific chatbots, on the other hand, are designed for a specific field of knowledge and may therefore not be able to answer questions outside the domain (Nimavat and Champaneria 2017, p. 1019, Ramesh et al. 2017, p. 348). Using the taxonomy of Janssen et al. (2020, p. 217), chatbots can be categorized based on design elements such as intelligence quotient or front-end user interface, which allows the derivation of archetypes. Another perspective allows the classification according to Zierau et al. (2020a, p. 9), which was based on a user experience perspective. Chatbots are categorized into functional, mechanical, and human dimensions, whereby these categories are then linked to service-oriented specific user cues (Zierau et al. 2020a, p. 9 - 11).

2.2 History of chatbots

The concept of chatbots had been known for a considerable time. In 1950, the Turing Test was proposed by Turing (1950, p. 433), which deals with the topic of how far machines can think similarly to humans and whether it is possible to check by means of a test whether the test participant is a human or a machine. The first famous chatbot was ELIZA, developed in 1966 (Weizenbaum 1966, p. 36 - 37). It used simple pattern matching and a template-based response mechanism to answer user utterances in the form of questions. Accordingly, no long conversations could be held, and no context could be discerned from the conversation. ELIZA's knowledge was domain specific and therefore limited (Adamopoulou and Moussiades 2020b,

scientific basis was achieved. Our results can generate added value in research and practice, e.g., derived implications for action, and serve as a starting point for further research.

9. Conclusion

Within the " gestalt mode" in HCI research (Adam et al. 2021, p. 3), we improve synergies between technologies and human behavior. In that way, we combined knowledge about human behavior (e.g., trust and usability) from the "outside mode" and knowledge about the design and maintenance of technical systems from the "inside mode" (e.g., chatbot design and chatbot progress) in the context of DSR. Our chosen three cycle DSR process (Hevner et al. 2007, p. 88; Vom Brocke et al. 2020a, p. 7) involved nine steps. Starting from an analysis of 103 chatbots that we re-examined after 15 months, we identified the problem of a high chatbot fluctuation. Based on a literature review, we found that chatbots' failures in practice have been addressed rarely in scientific literature. Within our study, we explored six experienced reasons for chatbots' failures in practice based on 20 expert interviews. To reduce failure risks, our selected solution approach is the identification of domain superior CSFs for chatbots. Taking the reasons for chatbots' failures as a starting point, we identified 12 domain superior CSFs based on our exhaustive literature review of 154 papers as well as 20 expert interviews and an FGD with 5 additional experts for evaluation. Within our discussion section, the obtained results are reflected, parallels to other HCI and IS research areas like business intelligence systems (e.g., Hawking and Selitto 2010, p. 6 - 7) or agile project management (e.g., Tsoy and Staples 2020, p. 984, 986) are highlighted and implications for research and practice are given. Next, we stated limitations of our results, such as a time limited validity as chatbots and related technology continuously evolves, or limited cultural diversification of interviewed experts, and provided an outlook for further research. Our CSFs analysis also illustrates discrepancies between a different emphasis on chatbot-related topics in practice and in scientific research, which can be used to identify further research gaps and research needs.

The list of reasons for chatbots' failures and the 12 domain superior CSFs represent our research results, based on the current status of the research and technology under consideration. We raise awareness on a high chatbot fluctuation rate and contribute by identify, address, and consequently reduce risks in the deployment and maintenance of chatbots for both researchers and practitioners.