Researching Human-AI Collaboration through the Design of Language-Based Query Assistance

by

Marvin Pafla

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2020

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Interactions with artificial intelligence (AI) are uniquely difficult to design because of the complexity of its output and the uncertainty of its capabilities for designers. Additionally, AI can be error-prone and needs human oversight. To try to overcome this issue, we followed a research through design (RtD) process to develop a high-fidelity interface prototype and to investigate human-AI collaboration in a realistic scenario. Specifically, we developed a support answer assistant that collaborates with customer service representatives to answer questions from customers. Our contributions highlight interaction designs that allow people to guide AI towards more satisfactory output, or to improve the utility of unsatisfactory AI output by making it possible to review and edit AI output efficiently. Early studies showed that participants used these designs to leverage the AI to provide a first answer draft or to automatically complete what they wanted to write. We describe our design and evaluation process and discuss how the insights of this research can help improve human-AI collaboration. Finally, we discuss how explainable artificial intelligence (XAI) can help identify incorrect AI output.

Acknowledgements

First, I would like to thank my two supervisors Stacey Scott and Mark Hancock for their help and support. Their knowledge, expertise, feedback and dedication shaped my thoughts and guided my work. I am deeply thankful for being able to come to Waterloo to do my Master's. I also want to thank my other two committee members, Jim Wallace and Siby Samuel.

Thanks to Deltcho Valtchanov and Julian Whiting. I had a blast working at Axonify for which I am deeply grateful. I learned so many things about AI and how to flourish in a corporate structure. I am also grateful for our discussions about politics, philosophy and justice, and my newly acquired Super Smash Bros skills.

I also like to thank the people that make the Games Institute a great place to learn, grow and just hang out. It always puts a smile on my face when I arrive at this place. Thanks to Neil Randall, Agata Antkiewicz, Marisa Benjamin, Pamela Maria Schmidt and Grace Van Dam.

Grad work can be hard and isolating. These people made it more fun: Cayley MacArthur, Alex Yun, Ali Rizvi, Joseph Tu, Rishav Agarwal, Tina Chan, Jin Lan Cen, Kenny Fung, Kateryna Morayko, Marco Moran-Ledesma, Caroline Wong, Justin Song, Oliver Schneider, Marcela Bomfirm, Robert Gauthier and Tanay Singhal. They also provided valuable feedback in our research meetings.

I am thankful for the Canadian governmental organizations (SWaGUR and Mitacs) that supported me financially.

Thanks to my family for their unquestioned love, availability and support. They have given me everything to flourish, grow and succeed in this world, something I just started to realize and appreciate. For their sacrifice, I am deeply grateful. Thanks Manuela, Heinz, Laureen, Oma Annemarie, and Emilija.

Dedication

For Ronny, Teena, and Maise.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI artificial intelligence iii, iv, x, xi, 1–36, 38–43, 49–63

DL deep learning 1, 3, 7, 8, 10, 12, 17, 58, 63

HCI human-computer interaction 1–3, 10–12, 14, 16, 17, 21

HCT human-computer trust x, xi, 6, 33–37, 43–45, 48, 54, 56, 63

ML machine learning 10, 12–14

MS MARCO MAchine Reading COmprehension 34

NLP natural language processing 4, 8, 13

RtD research through design iii, 1, 3–7, 17–19, 21, 22, 31, 32, 52, 61, 62

UEQ user-experience questionnaire x, xi, 6, 33–37, 43, 44, 46–48, 56, 63

XAI explainable artificial intelligence iii, 3–6, 11–14, 57–59, 62, 63

Chapter 1

Introduction

In this thesis, I focus on human-artificial intelligence (AI) collaboration in a prototype developed through a research through design (RtD) process. Throughout this process I endeavour to find out how AI can be triggered, edited, discarded and understood in a customer service scenario. This work represents a contribution to the field of human-computer interaction (HCI), particularly to the study of human-AI collaboration in interactive and realistic prototypes.

I will start this chapter by laying out my motivation for this research (section 1.1), before placing it into the larger context of HCI (section 1.2). I then discuss my research questions (section 1.3), and describe the approach I took to answer these questions (section 1.4). Finally, I list the contributions I made following this approach (section 1.5), and provide an outline for the rest of this thesis (section 1.6).

1.1 Motivation

AI is now a revitalized research field after the emergence of its subfield, deep learning (DL). DL is concerned with the training of deep (i.e., multi-layered) neural networks (or neural nets) which was made possible by the availability of high compute resources. In training, these neural nets 'learn' patterns found in training data. They also have the ability to reproduce these patterns once learned. This allows the technology to be generative: AI[1] can draw, write, or create music (Child et al., 2019; Ha & Eck, 2017; Sutskever et al., 2011).

[1]In this thesis, I will use the term 'AI' to describe the current state-of-the-art AI which often happens to rely on neural nets. Thus, AI and neural nets are used somewhat interchangeably.

While creative work is often thought of as something that machines cannot do, AI assistants designed for creative work have already emerged as research prototypes (McCormack et al., 2019; Oh et al., 2018; Qiu et al., 2019). These prototypes showcase the capabilities of AI and highlight how they can change the way creators such as writers, digital designers, or educators do their work. As Oh et al. (2018) suggest, AI can free humans from tedious or repetitive tasks and give suggestions or insights to a person who ultimately guides the collaboration.

However, it is unclear what form these collaborations should take. As AI is constantly evolving, interaction designers may not understand the capabilities of AI and what tasks can be fully automated (Yang et al., 2020). Furthermore, even though some tasks *can* be automated, it does not always mean that they *should* be. Humans are complex creatures—many might want to maintain control over human-AI interaction, as Oh et al. (2018) point out, including interaction within tasks that can already be automated or will be in the near future. Another challenge in the design of human-AI collaboration arises when the AI is incapable to reliably produce output that is satisfying to people using the AI.

This challenge of AI that is not powerful enough to be autonomous can be observed in the deployment of AI to industry. In a study of chatbot-type AI solutions in customer support, Xu et al. (2017) found that AI solutions could effectively handle emotional requests on social media, but that they fell short on informational requests. Their results, among others (Company, 2020; Morgan, 2020), show that AI solutions are still insufficient for some knowledge tasks. The authors argue for the integration of chatbots with human agents; that is, allowing chatbots to assist a human agent to provide more effective customer support. In this paradigm, the AI takes the role of a virtual employee assistant whose usage by information workers is predicted to grow to 25% by 2021 (Omale, 2020).

To overcome this and other challenges, guidelines that define how humans and AI *should* interact together have been proposed for various applications (Amershi et al., 2019; Horvitz, 1999; Tecuci et al., 2007). These guidelines, for example, specify that users should be informed about the capabilities of AI, or that it should be easy to detect and correct unsatisfactory output or AI behavior. However, while these guidelines help detect *ineffective* interaction design, it is still a challenge to realize *effective* interaction design in specific applications.

In this research, I take on this challenge and try to develop an AI assistant to support a human-guided task. Specifically, I consider the task of answering customer queries within the customer support industry. As laid out above, AI is currently unable to work autonomously in customer service but might be able to assist human customer service workers. From the HCI perspective, it is then unclear how this human-AI collaboration

should be achieved. Thus, my research explores how people can oversee and interact with the AI to collaboratively answer customer queries. In the following, I place this research into the context of existing research and define this scope of this work.

1.2 Context & Scope

In chapter 2, I will describe influences from different research fields that impacted this work. Specifically, this research is situated in the field of HCI, but was strongly influenced by the field of explainable artificial intelligence (XAI) and social sciences such as cognitive psychology and philosophy. This makes it difficult to place this research into a context as it does not follow traditional research fields such as HCI or AI separately but relies on knowledge and techniques from both.

In doing so, I employed HCI design techniques (e.g., pilot study, interviews, etc.) in a RtD process (Zimmerman et al., 2007) which relied on findings of cognitive psychology (e.g., considering the mental workload of users). This process was followed to produce a set of tangible designs of human-AI collaboration and a realistic prototype that is situated in the task of answering customer service questions (Zimmerman et al., 2007). To produce these designs and the prototype a significant amount of work and prototyping was required to train an AI model relying on DL literature (e.g., neural net architecture and training). Additionally, I developed and evaluated interactive visualizations to improve the user's understanding of AI behaviour. These visualizations originated in the field of XAI (e.g., post-hoc explanations).

Within the scope of this thesis, I conducted research through design to produce a realistic prototype that I then evaluated and discussed. In doing so, I focus on the interaction with the AI and how it is perceived and conceptualized. Furthermore, I investigate the impact on this interaction by a post-hoc explanation that aims to give insight into AI behavior. This research does not examine the efficiency of the human-AI collaboration (vs. just the human) or the task performance (i.e., the quality of produced answers), nor does it involve a performative evaluation of different designs of human-AI interaction.

1.3 Research Questions

In this thesis, I explore two main research questions:

Question 1: *How can humans interact with unsatisfactory AI?*

While AI reaches human-level performance in many tasks already today, tasks that require a significant amount of reasoning, memory, or creativity cannot be solved satisfactorily (i.e., the AI cannot produce output above a task-related threshold reliably). This drawback prevents the AI from solving the task autonomously. However, the AI might be 'good enough' to add value to human-AI collaborations. In this work, I explore what form this collaboration might take. Specifically, I look at how AI generation can be triggered, accepted if satisfactory, and edited and discarded if unsatisfactory.

Question 2: *Can a post-hoc explanation improve the collaboration of a human with an unsatisfactory AI in a realistic scenario?*

A key component of a successful human-AI collaboration is to detect if the AI produced unsatisfactory output. To provide this component, the field of XAI has produced (interactive) visualizations and metrics to inspect the AI itself and its behavior. In this research, I developed a visualization that allows the human to 'see' input that the AI considered relevant when it produced the output. I explore in a user study how participants conceptualize the visualization and how the visualization impacts human-AI collaboration in a realistic scenario. Specifically, I try to understand whether and how users can inspect AI output with the help of the visualization to determine its quality.

1.4 Approach

In this research, I take the viewpoint of leveraging AI to support a human-guided task and focus on designing appropriate interactions between humans and AI assistants. Specifically, I consider the task of answering customer queries within the customer support industry, a $350 billion industry that is heavily outsourced and already has a strong push for AI solutions (Morgan, 2020; Wadhwani & Gankar, 2020). However, the nature of the task—formulating reasonable responses based on available information—makes it generalizable to many other content-creation tasks that rely on large, text-based knowledge bases. For instance, in education, instructors commonly create lectures and tests based on existing textbooks and other literature.

To investigate human-AI collaboration in this task, I iteratively designed, using the RtD approach (Zimmerman et al., 2007), an interface that allows a user to engage with a state-of-the-art natural language processing (NLP) model, the Transformer (Vaswani et al., 2017), to answer customer queries. In this process, I went through many cycles of design, development, and evaluation to conceptualize the capability of the Transformer

model and to create a user interface that can handle complex AI output. This included the training of the Transformer model with the MSMARCO dataset to generate answers to customer queries based on a reference text. I then ran a pilot study with a strong focus on the qualitative experience of participants of the interaction with the interface and the AI. Afterwards, I ran a thematic analysis on the interview data collected in this study.

1.5 Contributions

My work produced the following contributions to research:

1. I followed a RtD process to design and develop an interface that enables users to collaborate with AI to answer customer queries. In this process, I described 'ideal' human-AI collaboration and produced interaction designs to collaborate with the model (i.e., auto-complete functionality). I describe the RtD process and my design decisions in detail.

2. I present a study with HCI experts that demonstrates that participants leveraged the AI through the design to provide a first draft to their answer and to finish sentences that users intended to write. In doing so, I found that most participants perceived the AI "just like a tool" which allowed them to have supervisory control over the AI. I then discussed the benefit of invoking the AI through activities related to the fulfillment of the task (e.g., writing) and how this integration with the task can make the interaction with the AI 'invisible'.

3. I developed a post-hoc explanation to give insight into the behaviour of the AI. I presented and discussed the results of a thematic analysis and quantitative evaluation of questionnaire data, which showed that users tried to match their mental model of what the visualization is supposed to highlight with what it actually highlighted. I then discussed how this explanation failed to 'explain' the AI and made a call to stop the over-promising of XAI related to neural nets.

1.6 Outline

The remainder of this thesis is organized as follows.

- In chapter 2, I discuss related research to situate my work. This includes the discussion of findings of the fields of AI, human factors, XAI and psychology.
- In chapter 3, I describe the RtD process. I then lay out specific design questions that I try to answer through iterative design.
- In chapter 4, I describe my pilot study with eleven participants using a within-participants design. In this study, participants interacted with the designs created in the RtD process to answer customer questions. After the interaction, I conducted a semi-structured interview with participants.
- In chapter 5, I describe the thematic analysis that was conducted on the interview data collected in the pilot study. This includes a description of what participants stated they did, how they made sense of the system, and my interpretation of their conceptualizations with regards to the context and the task.
- In chapter 6, I analyze the data from two questionnaires that I administered in the study. These questionnaires are the human-computer trust (HCT) scale and the user-experience questionnaire (UEQ). I also analyze log data from the system. This included running statistical tests and the production of informative graphs.
- In chapter 7, I synthesize the findings from the thematic and quantitative analysis. I then discuss these findings in the light of the related research and speculate about human-AI collaboration beyond the developed prototype and task. Specifically, I discuss the benefit of invoking the AI through activities related to the fulfillment of the task (e.g., writing) and the advantage of high-recall systems in low-risk environments if the costs of recovering from failure is low.
- In chapter 8, I summarize this work, list and describe my major contribution, and describe the limitations of this work. I then provide ideas for future research.

Chapter 2

Related Research

In this chapter, I present the related work informing my research. First, I will discuss how artificial intelligence (AI) is going through a 'second summer' with the emergence of deep learning (DL) after periods of reduced funding in the 1970s and 1990s. Second, I will give an overview of some human-AI collaboration prototypes that were enabled by the recent advancements of this field. Third, I will lay out why the development of these prototypes comes with challenges that are unique to AI. Fourth and fifth, I present how these challenges manifest in issues of AI interpretability (i.e., understanding AI behaviour) and AI leadership (i.e., when does AI take control in the human-AI collaboration). Finally, I introduce the process of research through design (RtD) to tackle these challenges and design human-AI collaborative prototypes to inform human-AI collaboration.

2.1 A New AI Summer

The field of AI not only "attempts to understand intelligence" but also to build intelligent agents (Russell & Norvig, 2009). Defining what (artificial) intelligence precisely is, however, has been agonizing AI researchers (Russell & Norvig, 2009). The resulting frustration can be heard when Tesler states that "[artificial] intelligence is whatever machines haven't done yet" (Hofstadter, 1979). Consequently, this lays open the contradiction of AI research: "When we know how a machine does something 'intelligent', it ceases to be regarded as intelligent" (Reed, 2020). To avoid this problem and a general definition of intelligence, the field of AI and its researchers have traditionally focused on subfields close to human cognition (e.g., (uncertain) knowledge, reasoning, perceiving, acting, and learning; Russell and Norvig, 2009). Depending on these subfields, different (mathematical) AI models and

agents have been developed that are distinguishable by a multitude of factors such as their representation of 'mental' states (e.g., knowledge- or rule-based, latent numerical; Fodor and Pylyshyn, 1988), the task (e.g., classification, reasoning, clustering, language generation), or the type of learning (e.g., expert-induced, supervised, unsupervised, reinforcement; Russell and Norvig, 2009).

AI has made major advances in recent years, reaching human-level performance in tasks such as the classification of objects in images (He et al., 2015), speech recognition (Xiong et al., 2016) or language translation (Hassan et al., 2018). These advancements, mainly sparked by the field of DL, were achieved by training large neural networks with high computational resources (LeCun et al., 2015). Within the field of natural language processing (NLP), advancements by deep and large neural networks can be observed as well: the autoregressive Transformer model is the current state-of-the-art language model and possesses hundreds of millions of parameters (Vaswani et al., 2017). Using an internal attention mechanism, the Transformer scores text input according to its relevance when it predicts the next best word that follows its input (Vaswani et al., 2017). The Transformer has been used in a variety of tasks such as text summarization (Radford et al., 2018b), answering questions (Lan et al., 2019), or even creating music (Dhariwal et al., 2020). In fact, by mapping language tasks onto a sequence researchers have posited that this model can be used for *any* language task (Raffel et al., 2019). This makes the model versatile and powerful enough to enable the development of AI tools in a multitude of task (such as answering customer queries).

2.2 Human-AI Collaboration Prototypes

Given this progress in DL, researchers predict AI will become more prevalent in people's lives and work in the near future (Grace et al., 2017; Oh et al., 2018). Although the impact of AI may be socially disruptive and potentially displace millions of jobs (Grace et al., 2017), there will also be room for humans and AI to work collaboratively guiding each other in new ways (Kittur et al., 2013; Oh et al., 2018). These collaborations are likely to exceed small, repetitive, and tedious tasks, and produce more complex and creative ones such as drawing, writing, and in general, creating (Bowman et al., 2015; Oh et al., 2018). Indeed, a wide variety of human-AI prototypes have emerged that are beyond the scope of this thesis. In the following, I will list some exemplary research prototypes that are relevant to my research:

- Steiner et al. (2018) showed that a human-AI collaboration was more successful in diagnosing breast cancer than a human or an AI alone. This highlights the potential

of human-AI collaboration and gives a reason to developers to pursue them. Though not explicitly tested in the domain of customer service, this finding encouraged me to explore human-AI collaboration rather than the evaluation of the performance of the AI or the human alone.

- Oh et al. (2018) developed a prototype that allows users and AI to collaboratively draw together. The researchers found that users generally enjoyed working with the AI, but liked to take the lead in the interaction. Furthermore, they liked to receive feedback from the AI only when they asked for it. These findings reveal many questions about taking the initiative in human-AI collaborations which guided my work.

- McCormack et al. (2019) investigated how the communication of internal AI states can help humans and AI improvise music together. They found that their emoticon-based visualization increased the human musical engagement in music improvisation. This finding motivated me that visualizations can improve human-AI collaboration.

- Clark et al. (2018) explored the design of 'machine-in-the-loop' systems to assist the creative writing of stories and slogans. They found that participants reported the interaction to be fun, yet would not use the system (in this form) again as it was disruptive and not competent enough for the process of creating coherent text. Some participants "noted they would have preferred words instead of full sentences as suggestions [for the slogan completion]" (Clark et al., 2018). In my research, I was inspired by Clark et al.'s system (2018) to develop a writing assistant while significantly improving the AI and the interaction with it.

- Arnold et al. (2020) developed a similar system but focused more on the task performance (rather than the subjective experience). They reported that participants could write image captions faster when an AI tool provided suggestions for the captions. The consequence of this speed advancement was, however, that the text captions were shorter and more predictable when assisted with an AI. This highlights the potential usefulness of AI, but also shows that the quality of writing (i.e., slogans) is impacted by the AI itself as humans 'tune out' and become complacent (Parasuraman et al., 2000; cf. Kaur et al., 2020). In this research, I discuss my prototype with regards to these advantages (e.g., faster query answering) and disadvantages (e.g., human complacency) of providing AI assistance to customer service.

2.3 Why Human-AI Interaction Is Uniquely Difficult To Design

While these prototypes highlight potential advantages of and interactions with AI, open issues and questions remain (Abdul et al., 2018; Amershi et al., 2019). With the success of DL, for example, Abdul et al. (2018) notice a trend from production rules-based models towards larger, more complex machine learning (ML) models (i.e., mostly neural nets). While the human-computer interaction (HCI) community recently started to explore how to design and develop user-centered applications around these models, there is still a disconnect between the AI and the HCI community (Abdul et al., 2018; Yang et al., 2020). Specifically, Abdul et al. (2018) find that the AI and ML community have traditionally been concerned with algorithmic fairness and 'explainability' (cf., interpretability of AI, section 2.4), but not on usable and transparent applications that work for, benefit, and empower people. On the other side, many studies with a focus on user validation and empirical testing have been run in the HCI community to investigate 'smart' applications (Abdul et al., 2018). These applications include decision support in domains such as medical or business processes, recommenders, smart homes and devices, and intelligent user interfaces (Abdul et al., 2018). While Abdul et al. (2018) consider the contributions of both fields significant, they call for more HCI studies that take into account "the human side of explanation" and investigate the usability and practicality of large AI models in realistic applications drawing from the rich body of research in HCI on interaction design and software learnability. Yang et al. (2020) are more affirmative when they proclaim a "real need for HCI and AI research in collaboratively translating fairness as an optimization problem into a feature of AI [within a] socio-technical system, and into a situated, user experience of fairness".

To achieve this symbiosis of the HCI and AI research community, attempts have been made to increase the 'technical literacy' of designers and the interdisciplinarity of research teams (Yang et al., 2018; Yang et al., 2020). These attempts highlights the focus on the algorithmic complexity of AI that is often described as the main barrier for HCI researchers to prototype with AI (Yang et al., 2018; Yang et al., 2020). In response, Yang et al. (2020) point out that the *uncertainty around AI capabilities* and the *complexity of AI output* make AI applications uniquely difficult to design. On the issue of AI capabilities, the authors write that the capabilities of AI are highly uncertain especially in the early design ideation stage (Yang et al., 2020). This ties back to the question at the beginning of this chapter about what AI even is and makes it hard for designers to "understand what design possibilities AI can generally offer" (Yang et al., 2020). Nonetheless, anticipating the user-centered perspective on AI and its capabilities is crucial in user experience (UX) design (Yang

et al., 2020). On the issue of the complexity of AI output, Yang et al. (2020) note that designers need to conceptualize what the AI produces as a possible output to create effective interaction design. This can be a challenge as HCI techniques like sketching or prototyping become more complex and demanding. According to the researchers, a key challenge for designers is to anticipate the types of errors the AI can make and how interaction design can be leveraged to overcome them (Yang et al., 2020).

Together with the uncertainty of AI capabilities, the complexity of AI output is what differentiates AI applications from non-AI applications (Yang et al., 2020). Developing an AI assistant to answer customer queries, the research in this thesis faces this same challenge. Specifically, it is unclear what type of customer queries the AI will be capable of answering (cf., uncertainty of AI capabilities) and what form these answers will take (cf., complexity of AI output). While both of these characteristics of AI provide challenges to designers by themselves, they also impact other aspects of human-AI interaction. For example, if the output of AI is complex (i.e., the output space of the AI is large), how can users understand how and why the AI generated the output it generated? Or, if the capabilities of AI are uncertain, when is the AI capable 'enough' to take the initiative in human-AI collaborations? In the following two sections, I will look more into these two issues of AI interpretability and leadership.

2.4 Interpretability of AI

Before automation and AI, humans held the monopoly on agency in society (Lipton, 2016). This changed as AI became more powerful and established in decision making processes, such as diagnosis in healthcare or loan applications in finance (Lipton, 2016). With this shift, decisions made by AI (and humans) can be consequential for people which is why explanations for these decisions are demanded (Lipton, 2016). Providing these explanations, however, describes a challenge because "today's predictive models are not capable of reasoning at all" (Lipton, 2016). Lipton (2016) argues that the problem lies in the fact that most AI systems are trained as supervised learners: "[they] do not know why a given input should receive some label, only that certain inputs are correlated with that label" (Lipton, 2016). To overcome this problem and for humans to understand these models and systems, the concept of *AI interpretability* has been proposed in the field of explainable artificial intelligence (XAI) (Gunning, 2017; Lipton, 2016).

AI interpretability is, however, ill-defined (Lipton, 2016). There are multiple definitions of interpretability in the literature which suggest that interpretability is not a single concept but a conglomerate of ideas that need to be disentangled and demarcated from other terms

and concepts, such as intelligibility or explainability (Kaur et al., 2020; Lipton, 2016). Indeed, Lipton (2016) argues that AI interpretability is related to trust (of users in AI), causality, transferability (i.e., how easy can the AI be deployed in a different situation?), informativeness (of AI output), and fair and ethical decision making. The separation of these concepts of AI interpretability can be found in Abdul et al.'s (2018) citation network: trust has been investigated in intelligent user interfaces and automation within the HCI community, causality within the field of cognitive psychology (besides explanation and reasoning), and fair and ethical decision making within the ML community. As mentioned above, investigations of (user) trust in realistic scenarios with fair and ethical AI are rare (Abdul et al., 2018).

While the HCI community only recently started to provide these investigations, the ML/XAI community has been investigating the question of how to make AI models (especially neural nets) 'explainable' with more effort since DARPA announced their XAI program in 2017 (Barredo Arrieta et al., 2020; Gunning, 2017). This line of research mostly focuses on the goal of making AI models more informative and transferable by looking at the intelligibility of AI models (Abdul et al., 2018; Barredo Arrieta et al., 2020; Weld & Bansal, 2019). Without considering user trust or interaction design, researchers have focused on the simulatability, decomposability, and algorithmic transparency of different AI models, and what makes some these models inherently interpretable (i.e., "glassbox" models like GAMs, sparse linear classifiers, decision trees and association rule lists, case-based models; Abdul et al., 2018; Kim, 2015; Weld and Bansal, 2019). However, due to breakthroughs in DL, much attention has been paid to large and powerful, yet often uninterpretable, neural networks (i.e., "blackbox" models; Abdul et al., 2018; Kaur et al., 2020; Lipton, 2016; Weld and Bansal, 2019). To make these non-transparent blackboxes intelligible, research has provided text explanations, visualizations, and explanations by examples that are produced after the model's prediction/generation (i.e., post-hoc) for a specific input (i.e., local) (Bahdanau et al., 2015; Lipton, 2016; Ramanishka et al., 2016; Samek et al., 2017). According to Lipton (2016), these "explanations" provide useful information for users and ML engineers without sacrificing predictive power nor fully elucidating how the model works. Therefore, neural nets in combination with explanations from the field of XAI are interesting for my research as they enable the development of AI tools that could have not been realized with other forms of AI.

To provide explanations for neural nets, for example, attempts have been made to compute the sensitivity of predictions to input changes and to decompose a classification decision dependent on input variables (Samek et al., 2017). Specifically, a heat map of an image can be generated that highlights pixels that were relevant in predicting the class of the object that is in the picture (Samek et al., 2017). Similar to this heat map approach,

Ramanishka et al. (2016) highlight regions of an image representative of caption words the AI model found for the image. In the domain of NLP, attention weights of neural networks have been visualized (Bahdanau et al., 2015): Given a text input to the model, the visualization showed which parts of the input were relevant (i.e., what did the model "pay attention to") in producing the output. The Transformer has such a mechanism as well: Vig (2019) developed a visualization that highlights parts of the input text that were relevant to certain parts of the model (i.e., layers and heads) when predicting the output. In this research, I tried to improve this visualization and evaluate how it impacts human-AI collaboration in a realistic scenario. To the best of my knowledge, my work constitutes the first time this visualization has been evaluated in a user study.

Despite these efforts in XAI (e.g., post-hoc explanations), it is unclear how well these feedback mechanisms allow users to understand the model's decision and to draw conclusions from it. For example, Kaur et al. (2020) discovered that GAMs and Shapely additive explanations (SHAP) (i.e., a post-hoc explanation for ML models) were misused, misinterpreted and over-trusted by data scientists. Furthermore, they found that some participants did not critically evaluate the explanations but "took [them] at face value, using their existence to convince themselves that the underlying models were ready for deployment" (Kaur et al., 2020). To avoid this superficial evaluation of the explanation, the researchers warn of attempts to provide "full transparency" that can lead to information overload (Kaur et al., 2020). Finally, they found that participants with more ML experience reported lower confidence ratings in the "reasonableness of the explanations and the underlying models" (Kaur et al., 2020).

While these findings again highlight the need for more user studies that try to discover how humans conceptualize and interact with post-hoc explanations, research has indicated that there might be a fundamental discrepancy between human and XAI post-hoc explanations. For example, Kaur et al. (2020) argue that their explanations were not effective because they did not fit the mental model of participants (Kaur et al., 2020). This is consistent with findings from social sciences as XAI research "does not appear to be strongly informed by Cognitive Psychology in terms of how humans can interpret the explanations" (Abdul et al., 2018). Drawing on social science research, Miller (2017) argued that indeed the process of explanation is not the provision of a long causal chain of events that is made accessible with the help of visualizations. According to Miller, explanations of behaviour are not single entities provided to the "explainee" by the "explainer", but rather a social process in which the explainer chooses one of many explanations which is then evaluated by the explainee (Miller, 2017). Thereby, the explanation is not necessarily a cause of the behaviour but a selected cause in a (often very long) causal chain that is put in relation with some counterfactual event that *did not* occur (e.g., "Why did *this* happen rather than

that?") (Miller, 2017). While Halpern and Pearl's pioneering work (Halpern & Pearl, 2005a, 2005b) on causal models tried to formalize explanation from a computational perspective, common ML models like neural networks only model correlation (not causality), and thus are (currently) not able to provide human-level explanation (Abdul et al., 2018; Weld & Bansal, 2019).

Nonetheless, neural networks added with post-hoc "explanations" may still be the best way to solve complex tasks and gain some human insight into the model (Weld & Bansal, 2019). Meanwhile, it needs to be pointed out that most post-hoc explanations in the XAI community are static and intended for ML researchers themselves (and not users) (Abdul et al., 2018; Miller et al., 2017). In order to avoid "inmates running the asylum", Miller et al. (2017) calls upon researchers to rely on the social sciences and produce HCI studies with real users. Finally, there seems to be strong agreement in the XAI literature to provide these user studies with interactive explanations (instead of static ones, e.g., Abdul et al., 2018; Kaur et al., 2020; Miller, 2017; Weld & Bansal, 2019; Yang et al., 2020). It is unclear, however, if and how interactive post-hoc explanations can overcome information overload experienced by users while trying to gain intelligible insight into large and complex AI models (cf., Kaur et al., 2020). Thus, a key question in this research was whether a post-hoc visualization can help 'explain' the AI or its prediction, what conclusions users draw from the visualization, and how these conclusions impact the interaction with the AI.

2.5 Leadership in Human-AI Interactions

On the question of leadership, researchers have been asking how much control and initiative should be allocated to AI, and automation in general, for over forty years (Horvitz, 1999; Sheridan, 1992; Sheridan et al., 1978; Shneiderman & Maes, 1997). This line of research originates in the field of human factors and was started in 1978 when Sheridan et al. (1978) investigated how humans can control undersea teleoperators. With the availability of new technology, the researchers argued that these teleoperators do not need to be directly controlled (i.e., a human operator controls every movement) but can make autonomous decisions and take control for short periods and restricted conditions (Sheridan et al., 1978). Sheridan et al. (1978) referred to this control as 'supervisory control'. While teleoperators (or any other 'hardware') are not part of my research, the criteria of autonomy is a defining feature of and goal for both teleoperators and AI in general. Though Sheridan et al. (1978) argued in 1978 that AI "was a long way from being on-board controllers for teleoperators" they still had the wisdom to point out that the sophistication of AI will likely enable the automation of more human functions. Interestingly, they predicted that this sophistication

will lead to an increase in the need for human-computer interaction (instead of a decrease) as there will be a stronger focus of the human functions that are not automated (yet) and a need to integrate with the rest of the system (Sheridan et al., 1978).

Meanwhile, Sheridan et al. (1978) provided a model of the ten levels of automation in human-computer decision-making. Ranging from "the human does the whole job" to "the machine does the whole job", the researchers describe how automation can be used to provide, select, and implement actions. Throughout this process and depending on the level of automation, the human operator gains different levels of control and approval over the machine's actions. For example, the machine might provide and select actions which it implements after the approval of the human operator (cf., level 5 in Sheridan et al.'s levels of automation, 1978). While focusing on the decision selection and the action implementation, this model was later extended with the two precedent tasks of information acquisition and information analysis (Parasuraman et al., 2000). According to Parasuraman et al. (2000), this provides a framework for designers to determine the level of automation in each of these four dimensions separately. With AI that assists writing, information acquisition is automated by the system (as text can just be read in) while the information analysis and decision selection 'merge together' (i.e., AI input is used to predict text). Finally, implementing the action (i.e., how to present the text to the user) provides design opportunities for meaningful human-AI interaction. In this research, I pay attention to this last step and create designs that enable users to be assisted by AI in creating answers for customer queries.

To evaluate these designs, Parasuraman et al. (2000) provide specific criteria that consider human abilities and limitations. These criteria include the mental workload (i.e., is the information level appropriate for the human operator?), situation awareness (i.e., is the human operator still aware of the situation?), complacency (i.e., will the human operator lose their monitoring precision?), skill degradation (i.e., will the human operator lose skills because of the automation?), the consequences of a wrong decision (i.e., what is the cost of an error made because of automation?), and the reliability of the automation (i.e., can the same level of automation be guaranteed?).

While the framework of Parasuraman et al. (2000) has been recognized to improve the effectiveness of human-automation interaction, researchers have argued that this frame is not flexible enough to enable the modeling and delegation of tasks that can be decomposed into smaller sub-tasks (Miller & Parasuraman, 2003). Since then, research has looked into how human operators can take a more 'active' role in the delegation and control over these smaller sub-tasks (Linegang et al., 2006; Miller & Parasuraman, 2003). For example, Linegang et al. (2006) have developed a goal-driven, human-automation system to enable human operators to actively collaborate in dynamic mission planning by relying on control

theory to increase the amount of feedback from and human control over the automation. This highlights the focus of this line of research to provide a more active role to human operators with high levels of control. Providing these high levels of control to the user is also reflected in the literature of human-AI collaboration (Clark et al., 2018; Oh et al., 2018; Shneiderman, 2020). For example, Oh et al. (2018) report that humans in creative work like to keep control at every decision but appreciate feedback from the AI when requested. Clark et al. (2018) agree with this notion when they state that a high level of control enables machine-in-the-loop writing systems to adapt to a wider range of writing styles (Clark et al., 2018). Shneiderman (2020) generalizes this finding when he calls for 'human-centered' AI that strives for both high levels of human control and automation. To achieve this, Shneiderman argued for recognizing situations in which full human and computer control is either necessary or excessive. This is in line with the research on mixed-initiative user interfaces (UI) that follow the design principle of taking the initiative "at the right time" (Horvitz, 1999; Tecuci et al., 2007).

Finally, researchers have argued that reliable, human-centered systems are trustworthy to users (Parasuraman et al., 2000; Shneiderman, 2020). Though trust is a multi-dimensional concept and difficult to define (Lipton, 2016; Pavlou, 2003), research has long been investigating how trust can be built in automatic systems such as online shopping systems or service robots (Lu et al., 2019; Pavlou, 2003). For example, trust has been shown to be negatively correlated with the perceived risk of online shopping systems (and positively with the intention to use them) (Pavlou, 2003). The importance of the perceived risk in the system has also been ascertained in autonomous cars (Choi & Ji, 2015). Besides perceived risk, performance efficacy has been described as a factor to predict trust in service robots (Lu et al., 2019). There is preliminary evidence that being able to 'noncommittally' (or without consequence) test a system can help build trust in an AI system (Berge, 2018; Mesbah et al., 2019). These findings are relevant to this work as they highlight how trust in AI may be achieved through reliable, low-risk interaction. Finally, it remains an open question if a focus on reliable, low-risk interaction with AI can provide a counterweight to the above mentioned issues of AI interpretability (i.e., does the AI need to be fully explainable, or is it enough to understand what the AI is able to do on average?).

2.6 Research Through Design

To overcome the challenges of explainability and control in AI deployment, the HCI community has defined what 'ideal' human-AI interaction and collaboration could look like (Amershi et al., 2019; Horvitz, 1999; Tecuci et al., 2007). However, the challenge for how

to reach this ideal state in specific AI applications and tools remains. This is a problem, as much of AI-produced behaviour is still "alien" to people and they struggle to understand how technology may affect them (Abdul et al., 2018; Weld & Bansal, 2019). To overcome this, researchers have generally called for more HCI studies with a focus on interaction to focus on technology that benefits and empowers people (Abdul et al., 2018; Kaur et al., 2020; Shneiderman et al., 2016; Weld & Bansal, 2019; Yang et al., 2020). Specifically, Kaur et al. (2020) asked for more studies that focus on the qualitative aspect of human-AI interaction. Yang et al. (2020) criticize the lack of RtD studies . In these studies, human-AI interaction design could be developed and investigated with AI as a "design material" (Holmquist, 2017).

The reason why RtD studies are warranted is because research on the design of interaction between human and AI represents a "wicked problem": to explore how AI can be handled, understood, and corrected within a human-AI collaboration, "one has to develop an exhaustive inventory of *all* conceivable solutions ahead of time" (Rittel & Webber, 1973). In other words, how do you investigate interaction that depends on the design that you developed (Harris, 2019)? While this problem highlights the difference between design (i.e., "the specific, intentional, and non-existing") and research (i.e., "the universal and existing"), RtD processes try to overcome this difference by "making the *right* thing" (Stolterman, 2008; Zimmerman et al., 2007). This includes, in my case, the framing and articulation of the preferred state from the current state (e.g., how should designs be made to effectively guide human-AI interaction?), a series of interface designs and documentation of the design process (Zimmerman et al., 2007). Finally, Gaver (2012) argues that the "design research community should be wary of impulses towards convergence and standardisation" and be proud of "its aptitude for exploring and speculating, particularising and diversifying, and—especially—its ability to manifest the results in the form of new, conceptually rich artefacts". This is why my work puts an emphasis on the perception of the human-AI collaboration (rather than its performative evaluation) which I investigate with qualitative methods.

2.7 Summary

As prior research has shown, the design of human-artificial intelligence (AI) interaction and collaboration describes unique challenges to AI. These challenges are the uncertainty around AI capabilities (i.e., what even is AI and what can it do?) and the complexity of AI output (i.e., what different AI output and behaviour can be expected?). Amplified by recent advancements in the field of deep learning (DL) and the provision of powerful

neural networks, there is a strong need for AI to be more 'interpretable' before widespread deployment in commercial products. Furthermore, it is an open research question when and how AI can seize control in human-AI collaboration to provide benefits (and not disadvantages) to the human operator. The research through design (RtD) approach is able to provide answers to these questions.

In the following chapters, I will present the usage scenario that motivates my RtD study. I will describe the process of how design decisions were informed by findings in human-AI interaction and how they can inform future interaction. After my design process, I will present an evaluation of my final design. Finally, I will present the findings of this study and discuss it in the broader context of human-AI interaction.

Chapter 3

Designing for Human-AI Collaboration

In this chapter, I specify and describe the usage scenario for which I designed my interface. Consequently, I lay out how I developed the interface through a research through design (RtD) process in which I designed the interface in iterative cycles, each of which included feedback, reflection, and intentional redesign, as explained below.

3.1 Usage Scenario

In this research, I focus on the task of answering customer queries online in written form by customer service workers which is a demanding, labour-intensive, often outsourced, business. For example, in order to address a wide variety of customer queries, workers need to possess expert-level knowledge of products, the business, and problems customers usually face. In my study, I focused on technical domains such as setting up user accounts or interacting with cloud software and clients. Furthermore, there is a shift to faster, chat-based interaction in the customer service industry (Morgan, 2020). To overcome this challenge, the industry has sought the development of artificial intelligence (AI) agents to replace and assist human agents (Morgan, 2020; Omale, 2020). However, customer-facing AI solutions attempting to tackle this complex problem can be brittle; they might not understand the customer or might give incorrect or unsatisfactory answers. This poor performance can frustrate the customer and leave them seeking interaction with a real person. Nonetheless, such systems might still be of use to customer service workers by allowing them to rapidly

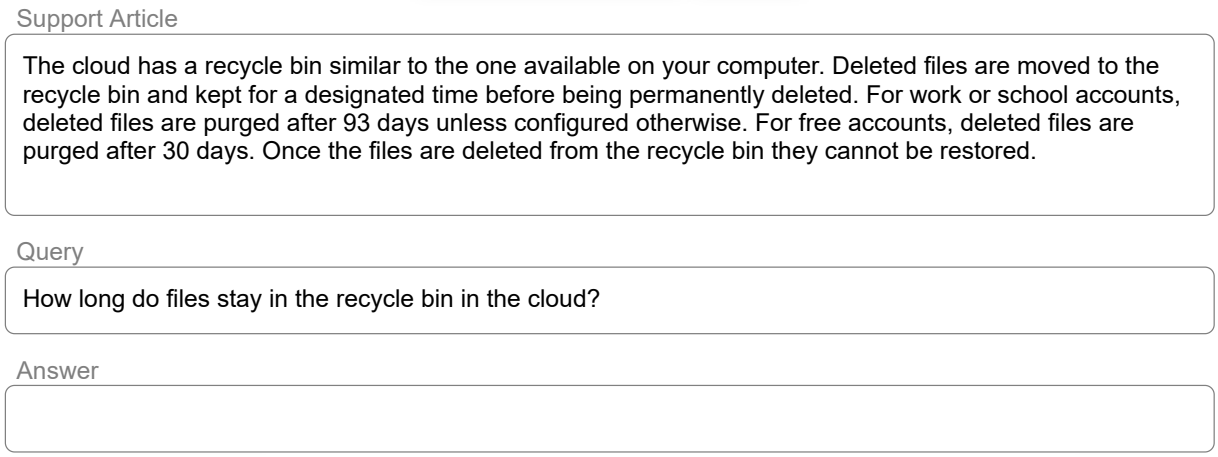

Figure 3.1: Example of the task for participants in which both the support article (i.e., information needed to answer the query) and the query is provided. The task for participants is to provide an answer to the query. There is only one support article provided for each query. That means that participants do not have to cross-reference information.

find and draft answers to customer queries by generating suggestions based on previously answered queries and highlighting relevant information from knowledge repositories.

While I only investigate customer service, many other commercial products and academic successes are centred on systems that involve queries on a large dataset. These systems require the AI to process large amounts of information to produce the right answers to the queries. This highlights the similarity of customer service systems to other systems as all of them try to automate the information pre-processing (e.g., processing information in text paragraphs) and decision selection stage (e.g., creating an answer to a pre-processed customer query) with the application of AI (Parasuraman et al., 2000). Writing summaries to multiple texts, creating quiz questions for students on a text book, or composing music from many inspirational pieces are all examples of specific realizations of such systems. To further simplify and generalize the task of answering customer queries, I assume that the answer can be found in one document/text alone so there is no cross-referencing necessary. The usage scenario can thus be described in simple terms: **Given a support article from a knowledge repository and a customer query, customer service workers (and the AI) must find an answer by referencing the text.** For example, participants are asked to answer the query "How long do files stay in the recycle bin in the cloud?" in Figure 3.1 while they can find relevant information in the support article.

To investigate this scenario, I reframed and conceptualized the problem through an iterative process of creating and critiquing design solutions in a RtD process (Zimmerman et al., 2007). As Gaver (2012) argues that the "design research community should be wary of impulses towards convergence and standardisation", I evaluated my design through qualitative methods such as short design presentations or short pilot sessions with HCI experts from my research institute. The comments made by the experts throughout the design process were collected and considered for the next design iteration to gradually improve the interface as the negative aspects of the interface get optimized (Harris, 2019). In the following section, I describe how and which three distinct design features emerged out of this design process. While the designs' creation was guided by persistent feedback and consideration in this process, I wanted to evaluate the interface added with the final designs in a study with participants. This study is described in the following chapters.

3.2 Research Through Design

In this section, I describe the design process of my machine-in-the-loop system (cf., Clark et al., 2018). While drawing from both human-computer interaction (HCI) and AI expertise, I was able to develop this system iteratively to assess the consequences of design on potential users. Throughout this process three design features crystallized through my RtD process that incorporate my principles of and ideas about an interface that improves human-AI collaboration. First, I collocated the workspace of the user and the AI to enable easy recovery from AI errors and to allow the human to correct the AI (see Figure 3.2). Second, I created an auto-complete functionality that allows the AI to directly react to input from the user in the workspace (see Figure 3.4). This feature also gives the user the chance to seamlessly guide the AI towards a desired outcome. Third, I developed an interactive visualization that enables users to trace the AI output back to its origin (see Figure 3.6). My intent was to give the user better means of inspection into the AI. Before I present these design features, however, I provide an overview of my efforts to narrow down the space of AI capabilities to generate answers to questions.

3.2.1 From Questions to Answers

There are many ways to design and train AI. For example, different architectures (e.g., the Transformer model) and forms of training (e.g., semi-supervised) impact the quality of the answers that the AI can generate. However, the design and training of AI does not only impact the performance of the AI but also how it receives input and provides output.

This has significant impact on the way humans can interact with AI. In this section, I briefly discuss two forms of possible interaction with the current state-of-the-art language model, the Transformer (Vaswani et al., 2017), that became apparent when I built small prototypes.

First, humans and the Transformer can work together to interactively finish the sequence the Transformer learned in training. Being an autoregressive model, the model tries to produce the next best words depending on the input words it received and the sequence it learned in training. For example, such a sequence could look like `text + $ + question + $ + answer` where `text` refers to any text (e.g., source articles) and `$` is a delimiter that signals different parts of the sequence to the model (Radford et al., 2018a). Depending on how much of an input the Transformer receives from this sequence it tries to complete the rest of the sequence by iteratively providing the next best word. In my RtD process I realized that this characteristic of the Transformer (i.e., its autoregressive nature) can be leveraged for different interactions. Specifically, this means that both the human and the AI can work together to interactively finish the sequence the AI learned. For example, the AI could freely generate a `text` and the human could ask a `question` about this `text` which the AI tries to `answer` in the end. For my prototype, I provided both the `text` (i.e., the source article) and `question` (i.e., customer query) to the AI. The `answer`, however, can be interactively created (see section 3.2.3).

Second, I experimented with providing alternative words for each word in the answer the AI produced. In essence, creating answers to questions is technologized as a classification task by the Transformer (i.e., picking the next best word out of a vocabulary of words). As for all classification tasks, the Transformer produces a probability value between 0 and 1 for each word in its vocabulary (i.e., all the words it 'knows'). With this probability distribution the word with the highest probability can be sampled (i.e., greedy strategy) and the user can be provided with a set of alternative words with the next highest probability. After testing my design, I decided to discard this design as the potential selection of alternatives for every word felt too overwhelming and distracted from solving the task. Nonetheless, it might provide an interesting thought to readers to provide not just words but (parts of) sentences to users as alternative generations. This can also be a way to guide the model towards a preferred output when they select one of these alternatives (see section 3.2.3).

3.2.2 From Text Boxes to Collocated Workspaces

While the inner workings of AI can be hidden away from the user, the delivery of input to the AI model as well as the handling of its output define the interaction with it. Thus, in

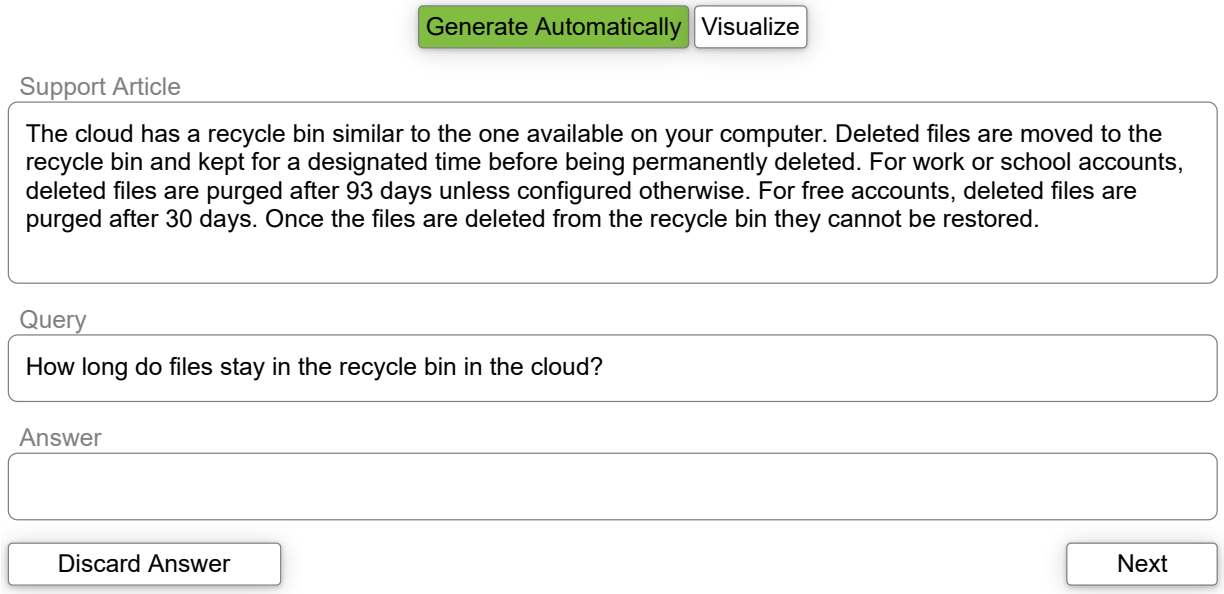

Figure 3.2: My initial design of a human-AI collaborative system to provide a response to a customer query. Three text boxes are provided in the workspace: 1) the "Support Article" displaying the source text that the AI and human use, 2) the "Query" to be answered, and 3) the "Answer" to be generated. An answer can be generated from the AI assistant (via the *Generate Automatically* button) or the user enter an answer themselves. In use, a query is provided by the customer, and a support agent generates an answer, possibly with the help of the AI. If the AI is used to generate an answer, the text can be altered by the user. The *Visualize* button can also be used to show what words the AI deemed "relevant".

my first realized design (see Figure 3.2) I decided to create text boxes to provide the user with interface elements to handle text input/output to the AI model. These text boxes provide the scaffold to my interface and allow users to create input for the AI, trigger its generation and receive output from it. In total, I provided three text boxes: one to display the provided support article, another to display the query, and another to provide space for the user to answer the query. This design iteration also had a *Generate Automatically* button to trigger the AI generation, a *Discard* button that discards the answer, and a *Next* button to proceed to the next query.

While this interface allows for a basic handling of the AI model, the interaction with the AI is still restricted to simple button functionality. This makes it difficult for the interface

to facilitate recovery when the AI makes a mistake. A fast recovery, however, is ideal in non-critical systems that can 'afford' to provide frequent assistance at the cost of precision (Kocielnik et al., 2019). To enable this recovery, I made the AI output editable in my second iteration of the design. This means that once the AI output was provided, the user could click on and directly edit the output in the text boxes in which it was displayed. A user could, for example, correct grammatical errors in the generated answer, improve the wording of it, or discard the generated answer altogether. My goal was to ease interaction by allowing fast, in-place editing of AI output. If both the user and the AI "work in the same space" the user has more oversight over the AI and can correct its output more efficiently. This can give the users a chance to harness the benefits of AI by relying on the AI to provide a first draft of the answer even though the generated answers might be flawed. In my short pilot sessions, HCI experts mentioned that it was useful to rely on the AI to provide a first draft of the answer because they could efficiently edit the answer to their liking in-place. An HCI expert mentioned that "though flawed, the AI output as an inspiration was helpful". This indicates that the AI can help with cognitive inertia in creative tasks (i.e., 'writer's block') (Clark et al., 2018; Garfield, 2008).

Notably, leveraging the AI in my interface changes the role of the user in fulfilling the task. The user's role changes from 'being a worker' that simply writes answers to customer queries to one that is now also required to oversee the AI and correct it when necessary. My intention is to reduce the workload on the user and remove tedious parts of the job (searching, reading, writing) while shifting the work towards more critical thinking and editing. This setup has the potential to make it easier for users to harness the capabilities of AI in answering customer queries and to improve human-AI collaboration.

Finally, a connection can be drawn to the ten levels of automation laid out by Sheridan et al. (1978). While they are not able to exactly describe this prototype (i.e, there is only one option that the AI creates which is why there is no real selection process), the system resembles a system with level 7 automation. That means that the user requests the AI to generate an answer (by clicking a button), that is then produced and provided to the user (i.e., the AI 'tells' the action; Sheridan et al., 1978). Though the answer can be approved, edited or deleted afterwards, the AI will always produce an answer and print it in the textbox (without the user's approval) after the user clicks the button. In the following section, I describe how I "promoted" the position of the human to give them more control over the AI. This included giving users the ability to actively approve AI output which makes the system resemble a system with level 5 automation (cf. Sheridan et al., 1978). It also leads to more "back-and-forth" collaboration (cf., Clark et al., 2018).

Figure 3.3: This image shows the Answer text box and the "Generate from here button". Users could provide parts of the answer and then right-click to pop-up the button. After clicking on the button the AI would try to finish the answer from where the user left off.

3.2.3 From Buttons to Auto-Complete

In the first iteration of the interface I provided a simple *Generate Automatically* button that triggered the AI to produce an output given a certain input (see Figure 3.2). This made the interaction slow and procedural: the user would need to navigate towards the button, click the button, and wait for the AI to produce the output to then be able to edit it. In essence, clicking on a button to start the AI to generate an answer was not much different from running a Python script: there was no further facilitation of interaction between the user and the AI. An HCI expert noted that their behaviour upon reading the text and question was to always click the button. This meant for the expert that they had spent time to wait for the AI to generate an answer—time in which they were unproductive.

While I experimented with pre-generating an answer for the user for every query (i.e., a generated answer would already be provided to the user upon being confronted with the query) I decided against this design for several reasons: pre-generating is expensive (e.g., what if the user does not need the AI for a certain query?), it does not allow the user to collaborate interactively with the AI, and placing an answer in the text box makes it appear that the work is "done" and may make the user feel unneeded. Furthermore, such a design would be intrusive as the user has to engage with the output that the AI produced (i.e., either by editing or deleting it) (Clark et al., 2018).

Instead of being intrusive, I followed Clark et al.'s advice (2018) and gave control back to the user by enabling dynamic, back-and-forth collaboration. I achieved this collaboration by introducing a pop-up button that would appear if the user clicked on any text position within the text box (Figure 3.3). The button's label read "Generate from here". The advantage of the pop-up button over the static button was that users had to navigate less as the pop-up button appeared at the position where the user right-clicked. This can speed up the interaction frequency between the user and the AI as the AI generation can be triggered more often.

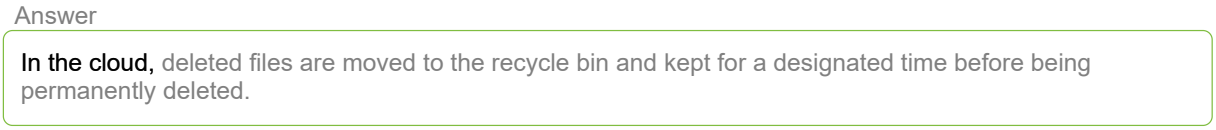

Figure 3.4: In this image, the AI auto-completed the text the user provided. The text the user provided is shown in black while the AI generated text is shown in grey. From here, the user could simply write over the text, pick parts of the auto-complete, or accept the whole AI generated text.

Furthermore, the provision of this button enabled users to write parts of the answer and then leverage the AI to generate the rest of it. That means that depending on what answer text the user provided, the AI had to consider this text when finishing the rest of the answer. Notably, this functionality gives the user the ability to effectively "guide" the AI to produce a more desirable answer. For example, if the user had an intuition about what the answer could be, they could provide the critical beginning of the answer for the AI to finish. While the functionality to finish an input, which is provided by my autoregressive AI model (i.e., the model takes any text input and tries to predict the next token), enables new interaction techniques, it can be confusing to users.

The problem with my design, noted by HCI experts, was that it was unclear how the position at which the user right-clicked impacted the auto-generated text. The confusion arose as users might not understand why certain text (the text after the trigger point) disappeared and was replaced by auto-generated text, as the trigger disappeared as well. This issue is a consequence of the sequentiality of my generative AI model as it takes whatever text input and generates from there. While I could have provided training to users I decided to find a more intuitive way to trigger AI generation which resulted in the development of the auto-complete feature. This feature is intended to satisfy the need for both interactive AI assistance and a high-level of human control by enabling users to approve AI output (cf., Clark et al., 2018; Sheridan et al., 1978).

Auto-Complete

I took inspiration from interfaces such as tab-completion in command-line interfaces and Gmail's Smart Compose system to address the design challenges that arose in my early design. I eliminated all buttons that triggered text generation by the AI and, instead, introduced an auto-complete functionality that reacts to the user's writing in real time. Every time the user enters text in the text box, the entire text is sent to the AI to generate

the follow-on text, until the end of a sentence is reached (or a max of around 50 words). The generated auto-complete recommendation is then displayed as grey text to the user following their input within the text box (Figure 3.4).

From there, the user can take multiple actions. First, they can keep writing their answer. If the recommended auto-complete does not match the text input anymore, it disappears and a new text generation is triggered after 400ms. Second, the user can press the TAB-button to accept the generated text which is then merged with the input of the user and the text cursor is moved to the end of the auto-completed text. This decision gives the user more control over the AI and reduces the level of automation from 7 to 5 as the user actively approves the 'action' by the AI (cf., Sheridan et al., 1978). Finally, the user can use the mouse cursor to click on the auto-completed text. This action accepts the auto-completed text up to the point of the cursor position, merging it with the user's text, while the text after the cursor continues to be displayed as an auto-complete recommendation (i.e., as grey text) to still be finalized. Unless the user clicks the 'Next' button to proceed to the next question, the answer is still editable after (parts of) the auto-complete was accepted. This means that users can 'rework' the suggestion (e.g., by pressing tab, then moving the cursor back to make edits).

My design intention for the auto-complete was to allow users to easily collaborate with the AI without requesting it (i.e., the AI automatically provides draft answers) but can also choose not to engage the AI (i.e., users can easily waive the auto-completed text by writing over it). That is, the AI provides real-time assistance that is tailored towards the user's input. Reducing the interaction complexity required to invoke AI assistance or to ignore it can provide more user control over the human-AI collaboration and increase the frequency of human-AI interaction.

Finally, I assume that the AI generates less text per successive generation with the auto-complete feature: the AI attempts to finish the sentence only, rather than generating a complete answer with given input (i.e., the AI decides when the answer is completed). I think this is preferable because this allows users to be able to identify if the output is flawed and guide the AI towards a better output with the help of the auto-complete functionality. This lowers the risk of cases in which the AI misunderstands the query and perpetuates this misunderstanding through flawed AI output. Catching the error early, therefore, can save time and resources in which the AI generates inappropriate output.

To support the user in evaluating the AI's output I developed a feature that enables users to trace back information through the AI model and highlight what parts of the text were relevant for the AI model to produce its output. I describe this "relevance check" in the next section.

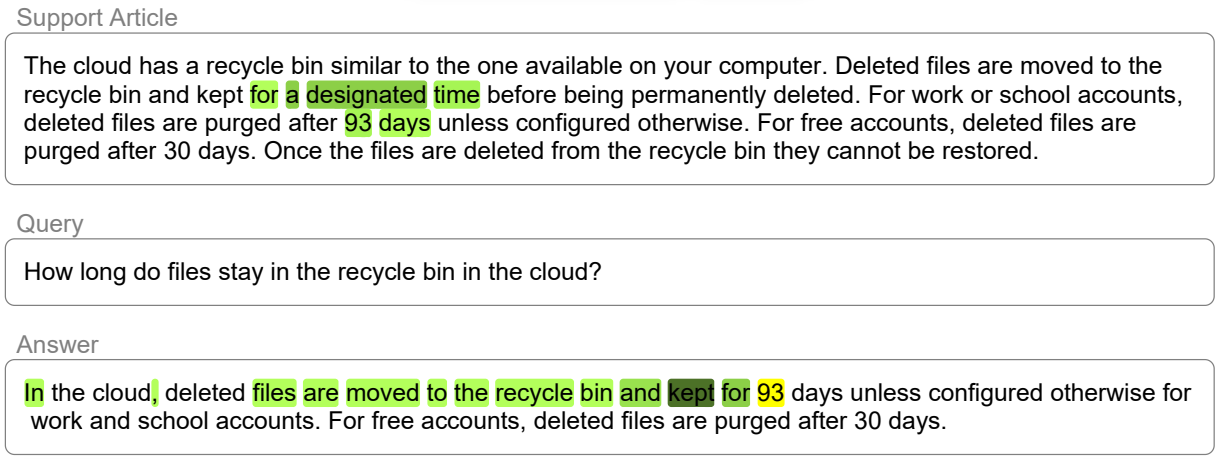

Figure 3.5: In this image, the user hovers over the number "93" to highlight all words that the AI considered relevant when it generated the number "93". Additionally, one can see that the AI paid high attention to the tokens right in front of the selected number. This makes sense as the AI relied on the previous words to predict the next best word.

3.2.4 From Visualizing Attention to Checking Text Origin

The Transformer that I used in this research follows a similar process to the one used by humans to generate an answer to a customer query: *find* relevant information in the source text that aligns with the query *to create* an answer. Technically, this is enabled by the attention mechanism intrinsic to the Transformer. In this section, I describe how I designed a feature that gives users access to this mechanism and allows them to see what information in the text the AI model *paid attention to* when it generated the answer to the query.

The attention visualization tool from Vig (2019) formed the basis of my visualization. While the tool can be used to locate relevant attention heads of the Transformer model (Vig, 2019), I wanted to design a feature that focuses on the input text alone (and not on attention heads; technical parts of the model) to reduce the number of relevance scores to visualize. For example, if the combined text (i.e., source text, query and answer written by user) consisted of 200 words that would pass through a 12-layer transformer with 64 heads for each layer, there would be over 30 million relevance scores to visualize.

To build this feature I averaged the relevance scores for all the layers and heads. This reduced the number of relevance scores by a factor of `number of layers` $\times$ `number of`

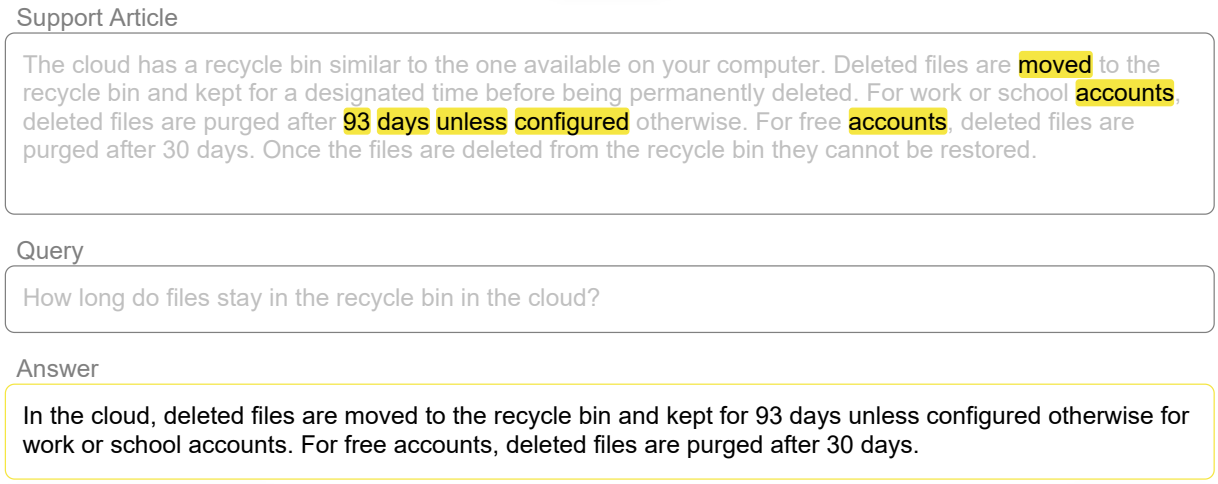

Figure 3.6: In this image, the users hovers over the Answer text box to highlight the averaged relevance scores for the whole box. The number of highlighted words is reduced to seven and only words in the Support Article text box are highlighted. I highlight the words in a yellow color to make the visualization easier to interpret.

`heads`. This resulted in a NUMBER OF WORDS $\times$ NUMBER OF WORDS matrix giving us the pairwise relevance scores for each word. In my 200 word example, this averaging results in 40,000 relevance scores. Note here that words that sequentially appear after a certain word have a relevance score of 0.

To give access to these scores, I built an interactive visualization for my first design prototype that highlighted the relevance scores in the Transformer model for a specific word. Users could activate this visualization by clicking on the button with the label "Visualize" appearing next to *Generate Automatically*. After, participants were able to hover over each word to highlight words that the AI considered relevant for this word (that the user hovers over) (Figure 3.5). This was achieved by grouping all relevance scores for a word into three equal-sized buckets which then were highlighted with different colours. The most relevant words for a word were highlighted in a dark green while words in the bucket in the middle were highlighted in a lighter green. Words in the bucket with the smallest relevance scores were not highlighted at all.

According to an HCI expert that gave me feedback on the design, this feature was fun to interact with but still very complex. This resonates with the contradiction presented in the related research: On the one side, research calls for more interactive explanations that give access to large neural nets (e.g., Abdul et al., 2018; Miller, 2017; Weld and Bansal,

2019). On the other side, explanations are hard to understand even for data scientists (Kaur et al., 2020). While some of the complexity can be mitigated with interactivity, it seems still difficult for participants to draw conclusions from these explanations (Kaur et al., 2020). In my study, depending on how relevance scores were bucketed for a specific word, many words (i.e., over 30) could be highlighted just for a single word. This complexity made it hard to derive the next user action (e.g., "Given what I see with this feature what should I do next?"). To make the visualization more accessible and useful to users I further simplified the visualization.

Based on feedback I received through the design process, I changed the relevance-visualization feature to average the attention weights for the answer text box and only displayed the relevance scores for words that appeared in that text box. The advantage of this visualization is that it enables users to highlight the relevance of words in the source text for all words in the answer text box at once. This reduces the amount of information and interactivity by highlighting the relevance to the whole answer instead of every word in the answer. More importantly in my opinion, it allows users to trace back the origin of text in the text box. For example, if the AI generated an answer to a text and query pair, users could activate the visualization, hover over the answer text box and highlight the most relevant words in input text that were relevant in generating the answer to the query. This crystallizes the purpose of the feature. By tracing back information through the AI model users can become aware of relevant parts of the source text that were relevant in generating the answer. This gives users the ability to orient themselves in the input text quickly and make edits to the answer if they consider the origin of the answer flawed (e.g., the AI model misunderstood the query and focused on the wrong words in the source text). In other words, the visualization works as an investigation tool: users can investigate AI output and more efficiently determine the quality of it.

To further simplify the visualization I adapted it to mimic the way humans read and highlight relevant information when reading articles. This includes that I only highlight the seven most relevant words in yellow for a text box (instead of different colours for different buckets and words). Furthermore, I greyed out words that are not highlighted (Figure 3.6). These changes were made to simplify the visualization (cf., Lombrozo, 2007) and to only give as much information as needed (i.e., the most relevant information) (cf., Grice, 1975). The intention of this design is to allow users to trace the origin of the answer even faster.

Finally, in my last design iteration, the visualization could be activated at any time to highlight relevant words for the answer text box. This is relevant because, with the auto-complete functionality described earlier, the answer text box could contain a combination of text that the user produced and/or that the AI generated. Once activated it would run the entire text through the AI model once to receive the relevance scores (and simply

discard the word that was produced in that run-through). In contrast to earlier designs that only provided the visualization after the user clicked on *Generate Automatically* and triggered the AI to produce an answer, this change enabled users to inquire relevant words of text at any point in time.

3.3 Summary

In this chapter, I started by describing the usage scenario in which the interface would be deployed. I then described how I leveraged the capabilities of the Transformer model and crystallized three design features in my iterative research through design (RtD) process: (1) the collocation of artificial intelligence (AI) and human workspaces to increase human-AI interaction and collaboration, (2) the auto-complete functionality to give users the ability to guide the AI towards more satisfactory output, and (3) the relevance visualization to give users insight into the AI and relate AI-generated answers back to the source article.

In the following chapter, I describe a mixed-methods, within-participant study to evaluate these assumptions.

Chapter 4

Pilot Study

I ran an online pilot study with eleven HCI experts from my research institute. The purpose of this study was to evaluate the design features that emerged from the research through design (RtD) process. These features have been iteratively improved upon through informal feedback from HCI experts, yet my goal was to evaluate them more thoroughly in a pilot study. In this sense, this pilot study was not a part of the RtD process, but an subsequent investigation of the interaction between the interface and human participants in a controlled environment. This separation allowed me to gain additional knowledge: while the RtD process itself generated knowledge (e.g., an exploration into 'ideal' human-artificial intelligence (AI) interaction), this study allowed me to investigate how participants actually interacted with and perceived the interface and the AI.

The original intention for this thesis was to conduct a more traditional in-person user study. The study preparations were completed and the study protocol had approval from the Research Ethics Office. However, this plan was disrupted due to the ongoing COVID-19 global pandemic, which is why the study was redesigned to be conducted online. Fortunately, I was able to maintain the original design and methodology of the study. This design included interacting with the interface in two experimental conditions and being interviewed after by the researchers. Due to the online aspect of the study I communicated through a voice chat with participants which did not change the mode of communication (e.g., the interview was still conducted orally). The interaction with the interface was also not changed as it was run in a web browser (and could thus be run offline or online). In the following, I describe the study design, questionnaires, the implemented system, the procedure, participants and the interview evaluation.

4.1 Study Design

Participants were exposed to two experimental conditions in a within-participant design after generating practice answers on their own (i.e., without the help of an AI). In the first study condition (see section 3.2.3), participants were able to generate answers to customer queries with the help of the auto-complete functionality. In the second condition (see section 7.3), participants could trace back written answers to relevant parts in the source article with the help of the visualization that I developed in my design. In this sense, the second condition is built on the first one as the interaction with the AI is the same, yet participants have access to the visualization as an additional feature.

I chose this within-participant design to avoid the problems Clark et al. (2018) ran into: separating participants between conditions (i.e., a between-participants design) would not allow participants to comment and contrast with other conditions and designs. With a focus on qualitative evaluation (through the semi-structured interview) and the introduction of four practice questions to answer by themselves, this design enabled participants to specifically comment on how they perceived and interacted with the AI to fulfill the task in general and how the addition of the relevance visualization impacts their interaction with the AI specifically. Note that I did not counter-balance the conditions as the relevance visualization was built on the interaction with the AI (provided by the auto-complete functionality) which I had to introduce first to participants. In future studies that have a stronger focus on quantitative analysis a between-participants design might be preferred to quantitatively measure differences between independent participants groups.

The study design follows a convergent, mixed-methods approach with both qualitative and quantitative components. My quantitative methods are the evaluation of log data on interface interactions, the user-experience questionnaire (UEQ), and the human-computer trust (HCT) scale (Laugwitz et al., 2008; Madsen & Gregor, 2000). My qualitative method is the evaluation of the semi-structured interviews. In the following section, I provide the background and rationale for the questionnaires I used in the study.

4.2 Questionnaires

The HCT scale was used to evaluate both cognitive and affective components of trust of humans in "intelligent systems which are designed to aid decision-making" (Madsen & Gregor, 2000). These components were the perceived reliability, technical competence and understandability of the system as well as the personal faith in and attachment to the

system by the human. The advantage of the instrument is that it is very general to intelligent information systems without requiring the user to have knowledge of the underlying technology. The scale provides a clear advantage over other scales I reviewed, even though the HCT scale was validated on the assumption that users had months of experience with the system (Atoyan et al., 2006; Madsen & Gregor, 2000) and was not validated by confirmatory factor analysis (Jeong et al., 2019b). Most importantly, it emphasizes different components of trust, such as understandability, which was crucial to this study. Furthermore, the HCT scale was designed for intelligent systems and is, therefore, different than the scales developed by Chien et al. (2014) and Jian et al. (2000) which were developed for automation. In my study, I argue that it will be easier for participants to relate AI to intelligent systems than to automation. Additionally, the latter two scales also lack a "systematic analysis to identify the dimensionality of the data and factor loading" and a lack of comprehensive validation tests respectively (Jeong et al., 2019b).

To contrast the levels of trust between my experimental conditions with the overall design of the interface, I used the UEQ to measure the user experience of my interactive interface (Laugwitz et al., 2008). The UEQ is a fast, easy-to-understand, 26-item scale that contrasts opposite adjectives on a 7-point Likert scale. It measures both usability aspects such as efficiency or dependability but also user experience aspects such as the the interface's attractiveness. I was able to adopt both the UEQ and the HCT scale without any changes which helps maintain their reliability and validity (Sousa et al., 2017).

Finally, I developed a small scale to self-report the level of expertise participants hold in AI. This includes not only questions about participants' programming skills and field knowledge in AI, but also their satisfaction in previous interaction with AI, what they think about the reputation of AI and if they think of AI as anthropomorphic. The questions about the reputation of AI were adapted from Mcknight et al. (2011) while questions about the anthropomorphism of AI were adapted from Lu et al. (2019). This means that these questions have not been validated in this exact form.

4.3 System Implementation

The Transformer model was based of the HuggingFace repository (Wolf et al., 2019) and trained on the MAchine Reading COmprehension (MS MARCO) dataset (Nguyen et al., 2016). The model was then deployed on a locally run Python server that participants could access with a web browser. The interface was built in HTML and Javascript.

4.4 Procedure

Participants were contacted by the researchers via video chat to set up the experiment and introduce them to the study. Once the experiment was set up, participants received a link to a website where they would find this study's systems. Upon arriving at the experiment website, participants completed the informed consent form, and then proceeded to the experiment page. Before interacting with the interface, participants were read the task instructions, which introduced them to the task of customer service workers (as I described in the usage scenario). These instructions included an example of what a good answer is and advised participants to include all relevant information from the source article in the answer. After reading these instructions, participants were exposed to four practice source articles which participants used to answer one customer query for each article without the help of AI.

Once participants created the answers to the four queries, they were exposed to two phases of question answering in which they were exposed to my experimental conditions. Each phase of question answering included ten different source articles that participants used to come up with an answer to ten provided queries. Before each phase, participants were exposed to a verbal tutorial given by the instructor through the video chat which introduced participants to the interface's features in that phase. Additionally, researchers would assist participants in answering the first two of the ten queries and answer participant questions about the interface.

After answering ten questions for both study conditions, participants filled out the UEQ and the HCT. At the end of the study, participants were asked about their experience with and expertise in AI in a questionnaire that I designed. Lastly, participants joined a semi-structured interview. The purpose of this interview was to contrast the two experimental conditions with each other, and inquire about participants' inner states (i.e., feelings and thoughts) about the designs.

4.5 Participants

Eleven participants were recruited from my research institute for this study. Participants were highly educated (ten graduate students + one undergraduate), young (between 20 and 36 years old; $median = 27$), and mostly identified as men (six male-identifying; three female-identifying; two preferred not to disclose). All participants knew the authors personally and might have been exposed to the interface (or earlier prototypes of it) or the idea of it in lab presentations and discussions, though had not interacted with it before.

4.6 Data Collection & Analysis

Data was collected from both quantitative and qualitative methods. To collect data from the questionnaires (HCT and UEQ) participants were guided to an external survey platform (i.e., Qualtrics). This data was then evaluated with statistical tests to find significant differences between study conditions for different constructs of the respective questionnaires. Additionally, I make use of the extensive tools that the creators of the UEQ, Laugwitz et al. (2008), provide. For any administered UEQ, these tools include programs to calculate metrics (e.g., mean, SD), internal validities and benchmarks (i.e., comparisons to other systems previously evaluated with the questionnaire; see Figure 6.3 for an example).

In addition to the questionnaire data, I collected log data from the interaction with the interface for each participant session. This collection included log entries when participants started the session, the amount of time spent on a question, when they edited an answer, when they triggered the AI to generate an answer, when they approved an answer, when they discarded an answer, and when they interacted with the relevance visualization. Besides this quantitative data, participants were asked to join a semi-structured interview which was audio-recorded and transcribed for further analysis. In the next sub-section, I describe this analysis.

4.6.1 Interview Analysis

I asked participants semi-structured questions in the interview to give enough flexibility to participants to express themselves (interviews lasted between 7 and 21 minutes depending on the participant). To evaluate the interviews I ran a thematic analysis as described by Braun and Clarke (2006). In this section I acknowledge my role as *active* researchers and describe my presumptions to this analysis.

In this analysis, I was particularly focused on how participants report how they interacted with AI, and how their ideas and conceptualizations of, and assumptions about AI impacted this interaction with regard to the interface and the relevance visualization. For this reason, the thematic analysis focused on latent themes (i.e., 'hidden' themes such as thought and behavior patterns) which I conceptualized and interpreted (Braun & Clarke, 2006). As this is a new field of research, this analysis followed an inductive thematic analysis which gives me the flexibility to cover a wide range of possible themes (Braun & Clarke, 2006).

I provide the prevalence of each theme in parentheses across all participants (i.e., the number of participants exhibiting that behaviour). My goal is to provide readers with "how

much a theme really existed in the data" (Braun & Clarke, 2006). I want to highlight, however, that the prevalence of themes is partly dependent on the questions that were asked in the interview. Fortunately, thematic analysis provided me with the flexibility to investigate themes that were exclusively brought up by a few interviewees and thus rarely appear in the data set (Braun & Clarke, 2006).

4.7 Summary

In this chapter, I described the design of the study that I evaluate and discuss in the following chapters. This study follows a within-participants design to give participants the chance to contrast the study conditions in the semi-structured interview. Furthermore, I administered the human-computer trust (HCT) scale and the user-experience questionnaire (UEQ) to measure participants trust in the system and their perception of usability and design of the system. Finally, I described my participants as very educated and the procedure they went through in the study. In the following chapters, I present the results of the thematic analysis of the interview data (chapter 5) and the evaluation of the questionnaire data (chapter 6).

Chapter 5

Thematic Analysis

In this chapter, I present the results of my thematic analysis. Thematic analysis requires themes to be defined and described, as well as analyzed within the context that they appear (Braun & Clarke, 2006). Thus, the relationship and hierarchy between themes as they emerged in my prototype and within my theoretical framework were analyzed. I split the analysis into two emergent themes: the interaction with the artificial intelligence (AI) itself and the interaction with the relevance visualization. For both, I describe and analyze what participants said they did and how they made sense of what they did.

5.1 Interacting with the AI

Almost all participants (10 of 11) said their primary strategy was to use the AI to provide a first draft of the answer that could be improved upon by cross-referencing the source article after reading the query and AI-generated answer. As a second strategy, 8 of 11 participants occasionally relied on the auto-complete functionality to finish their answer attempts. While the type of human-AI interaction is dependent on multiple factors, participants seemed to leverage the AI to get a answer draft for each question first before they moved over to the second strategy. On one hand, six participants reported that they "just tried to get the job done" (P2) by scanning instead of reading text as much as possible. This intention does not hinder the ambition to create "a fully elaborated answer" as participants (3 of 11) would simply "fluff up what [the AI] gave [them]" (P3). Here, the important contribution by the AI seemed to provide a first draft of an answer or "a pool of words [...] that [participants] can build on top [of]" (P1). On the other hand, three

participants criticized the speed of the auto-complete functionality as being too slow, as exemplified by the following comment:

> "I wait for it, and then I would be impatient because it is not showing, so I would type the rest of it." (P7)

Given this technical drawback, this might explain why only 1 of 11 participant reported that the AI helped them out occasionally when they did not know how to continue the answer. While it remains unclear whether a faster auto-completion would have caused more reliance on the auto-complete functionality, participants still found a purpose for it in the given prototype: 6 of 11 participants reported that they used the AI to "pull [more structured information such as] a list of things or concrete steps" (P3) from the source article. P1 described this pulling of information as "copy-pasting [from the source article]". This copy-pasting of specific information seemed intentional as participants reported that they "started to type out the beginning of sentences" (P3) and "assume[d] the AI [would] complete it" (P4). This finding indicates that participants (6 of 11) used the auto-complete functionality if they already knew *what* they wanted to write. Consequently, 3 of 11 participants described a learning effect they had while interacting with the AI, as illustrated by the following comment:

> "So, if I'm waiting for it to answer something I got a feeling for how much information I had to give it before it would come with the right answer. And it felt like once I figured out where that point was I could use it pretty consistently on the same type of question." (P9)

As the second sentence of this quote indicates, most participants (9 of 11) found that the AI generally helped them with the task by making the creation of answers more efficient (6 of 11) by creating a first draft (2 of 11). When asked about the AI's competence, participants reported the AI to be competent enough (6 of 11) or moderately competent (5 of 11). As noted by 4 of 11 participants, the AI seemed to be especially capable of creating the answer if "it was a simple question, and if there was a sentence [in the source article] that was clearly linked to that question" (P10). If the AI was not able to produce the right answer, P9 noted that they would:

> "continu[e] to write the right answer until the [AI] auto answered itself to be in line with what [they were] thinking." (P9)

Though this quote highlights the range of interaction possibilities the AI and the interface provided, most participants (6 of 11) described the AI (assistant) only as a tool "rather than an intelligence" (P9), and compared it to industry applications like GMail (https://www.google.com/gmail) or Grammerly (https://www.grammarly.com/). This conceptualization of AI seemed to have consequences on participants' perception on the issue of trust: rather than talking about the 'intentions' of AI participants (4 of 11) seemed to equate trust with the face validity of the AI's output. In other words, participants (4 of 11) would 'trust' the AI capabilities if it reliably produced good answers, as illustrated by the following comment:

> "Because the system was giving me some answers that I didn't trust I was always paying attention to see if the AI was right, and this is about trust with the system, right? If I saw that the first answers from the AI were super correct - no errors - I would just trust. [...] But because the AI was giving some not 100% correct answer [...] I would have to double-check to see if it was right." (P5)

5.2 Interacting with the Visualization

Participants (7 of 11) reported to have used the relevance visualization to verify AI-generated answers by "find[ing] where the AI thought it got the answer from" (P0) and "reading a little bit of the text of the support article around those [words highlighted by the visualization] to confirm whether the answer is correct" (P1). While the source articles were only between three and five sentences long, 5 of 11 participants imagined that the visualization would be especially useful if the articles were longer (e.g., "ten pages", P2). As such the relevance visualization can be useful to customer service workers as they are expected to answer queries from hundreds to thousands of pages of source materials (e.g., dozens or hundreds of different products being supported by a call centre). On the other hand, one participant noted that they would still read the entire source and having longer sources would, thus, not make a difference.

However, despite its potential usefulness, the visualization was generally criticized for a multitude of factors. 5 of 11 participants stated that interaction with the visualization broke the routine of the task because it was cumbersome. Some design issues listed by participants included the position of the "Visualize" button at the top of the screen (2 of 11), and the need to hover over the text box after clicking the button (4 of 11). To overcome this, 3 of 11 participants recommended highlighting text from the source article

automatically when the AI provided an output. While this could improve how participants interact with the visualization it still remained unclear to 6 of 11 participants how they could read and interpret the information that the visualization provided, as illustrated by the following comment:

> "If it hasn't highlighted all the words that it put into the suggested answer than I don't know, like, how to get from 'here is the question' to 'here is the suggested answer' like in the middle are those highlighted words and I don't know how you get from point A to B. Like it is telling me something but I don't know how to crack it." (P3)

This quote illustrates the difficulty participants (6 of 11) had in interpreting the highlighted words shown by the visualization, especially when the information was disparate and not related to the answer. The perception of how much the highlighted data was scattered (i.e., disparate and unrelated), thereby, varied from participant to participant. For example, P3 argued that "sometimes the information highlighted was, kind of, like disparate or not related" while P10 considered the visualization "completely random". Interestingly, participants (3 of 11) tried to overcome this lack of interpretability (i.e., how did the AI generate the answer given these highlighted words?) by matching their *expectation* of what the visualization might highlight with what it actually highlighted:

> "And like most of the time I didn't like the initial answer that it gave me, so then I would type my own, and then visualize it to see where would it take, like, in my mind I feel like I was highlighting stuff and then I wanted to see whether it was highlighting the same stuff that I was thinking." (P4)

While this mental verification made participants (2 of 11) more confident in their answer and the AI when it matched their expectation, it still did not help interpret disparate or non-related information. This did not seem to be much of a problem, however, for some participants (4 of 11) that took the underlying mechanism of the visualization as associative (i.e., it would try to make a connection between the answer and the source "as best as it can"). For example, 2 of 11 participants noted that the visualization was occasionally highlighting words that did not appear in the answer but were in the same category as the answer (e.g., names or numbers). Though this would make the visualization highlight words that were not related to the answer, it was still useful to participants (2 of 11) as they could find the correct answer among the highlighted words. According to P2, however, the underlying associative mechanism was generally problematic since it would try to highlight the AI-generated answer independently of its correctness:

> "It showed me what the AI used to answer the question but that doesn't help me answer the question if the question is being answered incorrectly. If the question was answered correctly I don't need the visualization tool at all. But if it is being answered incorrectly and it is visualizing certain segments that doesn't help me because it picked the wrong segment." (P2)

The severity of this issue was increased by my final design as I allowed participants not only to highlight AI-generated answers but any text (i.e., both AI and user generated) that was placed inside the Answer text box. While this allowed participants (2 of 11) to use the visualization as 'an advanced search' (P0), in which they would type words into the text box to actively look for key words, it confused participants that assumed the visualization would highlight words relevant to the answer generated by the AI, as illustrated by the following comment:

> "I don't understand why it needs to be using my answer to give me those relevant words. [...] Because if [...] just one sentence from me can change the output of relevant words, I don't know what relevance really is." (P10)

5.3 Summary

In this chapter I described my thematic analysis. I found that participants used the artificial intelligence (AI) to generate a first draft of the answer with the auto-complete feature. This helped them to get started more easily because they could simply edit the AI generation to their liking. In this interactive process participants perceived the AI mostly as a tool that can be leveraged. Because of this perception some participants took the AI's output at face value and trusted the AI if it produced satisfying answers. This finding had impact on the relevance visualization as participants did not use it much. The reason for this can be found in a cumbersome interaction but also in the inaccessibility of highlights being produced when they mismatched the mental model of what the participants thought the visualization "should" produce. Nonetheless, participants still found use-cases in the visualization such as using it as a general search and as a way to find the parts of the text the AI paid attention to.

The implications of these qualitative findings will be discussed and contextualized in the broader literature in chapter 7 (Discussion). In the next chapter, I describe the results of the questionnaires administered throughout the study. I will also look into the log data that gives more insight into how participants actually interacted with the AI.

Chapter 6

Quantitative Results

In this chapter, I will present the evaluation of the quantitative data collected in this study. This includes data from two questionnaires, the human-computer trust (HCT) and the user-experience questionnaire (UEQ), questions about the expertise of participants on artificial intelligence (AI) and their familiarity with it, and log data from the interface that I collected while participants interacted with it. While most of the evaluation will focus on descriptive statistics, I also apply inferential statistics to compare the study conditions auto-complete (see section 3.2.3) and relevance visualization (see section 7.3).

6.1 Questionnaires

When I administered the HCT and the UEQ I followed my impulse to standardize my results—against Gaver's advice (2012). While my intention was to expose participants to the AI with and without the relevance visualization to enable them to contrast the two designs in the interview, this conflicted with the attempt to measure differences in the designs with the help of the two questionnaires. Indeed, in this within-participants design, 3 of 11 participants reported learning effects they had with the AI. One participant specifically mentioned that this changed the way they filled out the questionnaires. Finally, one participant raised concerns about the wording of the HCT as it was unclear whether the 'system' referred to the AI or the interface.

Therefore, the following results have to be taken with a grain of salt. While it might be hard to compare the two study conditions, I nonetheless present the results as they can still provide insight into the general perception of the system and into the subscales

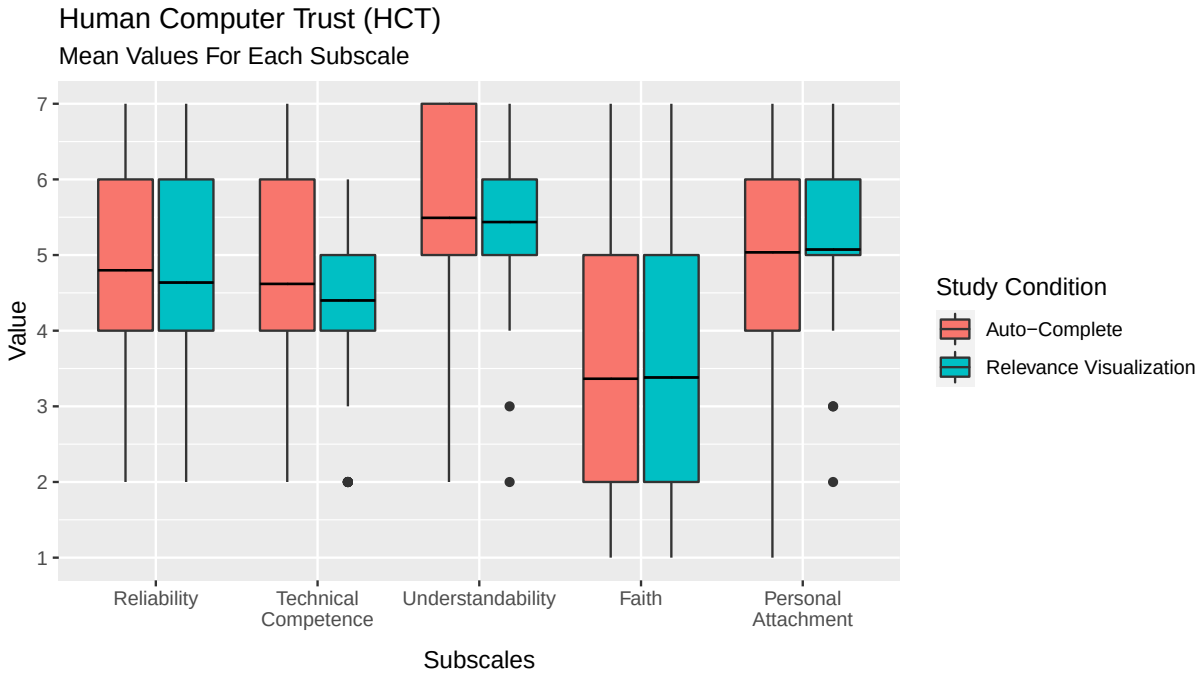

Figure 6.1: In this figure, a boxplot is presented for the two study conditions for each subscale of the HCT. These subscales are *Reliability*, *Technical Competence*, *Understandability*, *Faith* and *Personal Attachment* and can be found on the x-axis. On the y-axis the scores of the 7-point Likert scale (7: strongly agree) of the HCT can be found. For each boxplot I present the confidence interval, the range, and outliers. The black bar inside the box plot represents the mean.

of the HCT and the UEQ, such as the attractiveness, reliability, and understandability of the system as a whole.

6.1.1 HCT

The main results of the analysis of the HCT can be found in Figure 6.1. The boxplots for each subscale of the HCT are presented and compared pairwise for the two study conditions. In general, the boxplots are aligned for the two study conditions but vary accross the subscales. Specifically, the means for *reliability* ($M = 4.80$, $SD = 1.41$), *technical competence* ($M = 4.62$, $SD = 1.18$) and *personal attachment* ($M = 5.04$, $SD = 1.41$) vary between 4.5-5.0, while the means for *faith* are lower ($M = 3.36$, $SD = 1.65$) and for

Scale	Auto-Complete	Relevance Visualization
Reliability	4.80 (SD: 1.41, α: 0.75)	4.64 (SD: 1.34, α: 0.90)
Technical Competence	4.62 (SD: 1.18, α: 0.86)	4.40 (SD: 1.29, α: 0.69)
Understandability	5.49 (SD: 1.49, α: 0.79)	5.44 (SD: 0.96, α: 0.58)
Faith	3.36 (SD: 1.65, α: 0.91)	3.38 (SD: 1.65, α: 0.88)
Personal Attachment	5.04 (SD: 1.41, α: 0.74)	5.07 (SD: 0.98, α: 0.67)

Table 6.1: Means, SD (in parenthesis) and Cronbach's alpha (in parenthesis) for each subscale of the HCT for both study conditions.

understandability are higher ($M = 5.49$, $SD = 1.49$) for both study conditions. The exact means and their standard deviations can be found in Table 6.1.

In Table 6.1, I also provide Cronbach's alpha (Cronbach, 1951) which is lower for the relevance visualization condition than the auto-complete condition for all subscales but *reliability*. This decrease of internal consistency of the relevance condition (compared to the auto-complete condition) does not seem to accord with an increase in standard deviation. That could mean that participants changed their answers for individual items which did not, however, have an impact on the overall answer pattern for each subscale.

Comparing Means Of Study Conditions

Despite the concerns mentioned at the beginning of the chapter (re:comparability of the conditions), I still test if there is a difference in means between the two study conditions for each subscale. As participants went through both study conditions (i.e., a within-participants study design), I applied a dependent (i.e., paired) test for the all subscales. While the assumption of normality was satisfied for the differences between the values of the two study conditions for the subscales *reliability*, *technical competence*, *faith* and *personal attachment*, the differences of the subscale *understandability* was found to have significant kurtosis and skweness (Field et al., 2012). Therefore, I applied the non-parametric Wilcoxon signed-rank test for this subscale. As expected, none of the tests indicate a significant difference between the means (reliability: $t = 0.72$, $p > .1$; technical competence: $t = 1.32$, $p > .1$; faith: $t = -0.11$, $p > .1$; personal attachment: $t = -0.29$, $p > .1$; understandability: $V = 375$, $p > .1$).

Relevant Single Items

Though there was no significance between the two study conditions, it is worth highlighting single items within the subscales. To avoid 'cherry-picking' I highlight items that are 0.5 points from the mean for each study condition or from each other. First, participants neither agreed nor disagreed with the statement that the system produces advice as good as a highly competent person in both tests administered in the auto-complete ($M = 3.9$, $SD = 1.16$) and relevance visualization conditions ($M = 3.2$, $SD = 1.40$). Second, participants disagreed with the statement that they would believe the system rather then themselves if they were uncertain about a decision in both tests administered in the auto-complete ($M = 2.6$, $SD = 1.55$) and relevance visualization conditions ($M = 2.8$, $SD = 1.70$). This was also the statement that earned the most disagreement (i.e., lowest mean) of all statements. Third, participants neither agreed nor disagreed with the statement that they felt attached to the system in both tests administered in the auto-complete ($M = 4.3$, $SD = 1.35$) and relevance visualization conditions ($M = 4.4$, $SD = 0.88$). Fourth, participants state that they would feel a sense of loss if they could no longer use the system in the auto-complete condition ($M = 4.0$, $SD = 1.54$) which increased in the relevance visualization condition ($M = 4.8$, $SD = 1.11$)

6.1.2 UEQ

The main results of the analysis of the UEQ can be found in Figure 6.2. The boxplots for each subscale of the UEQ are presented and compared pairwise for the two study conditions. In general, the boxplots are aligned across the subscales but also vary for the two study conditions. Specifically, the means for all subscales but *perspicuity* ($M = 1.75$, $SD = 1.35$) and novelty ($M = 0.52$, $SD = 1.47$) vary between 1.16 and 1.48. The means for the subscales of *attractiveness*, *novelty*, *perspicuity* and *simulation* also differ visibly between the two study conditions.

The exact means and their standard deviations can be found in Table 6.2. On one side, participants found the system in the auto-complete condition more attractive, dependable, efficient, perspicuous, and stimulating than the system in the relevance visualization condition. On the other hand, participants describe the system as less novel in the auto-complete condition than in the relevance visualization condition.

In Table 6.2, I also provide Cronbach's alpha which increases for all subscales after administering the test in the relevance visualization condition (compared to the auto-complete condition) (Cronbach, 1951). This means that the UEQ for the relevance visualization condition showed higher internal consistency. The standard deviations for all subscales and

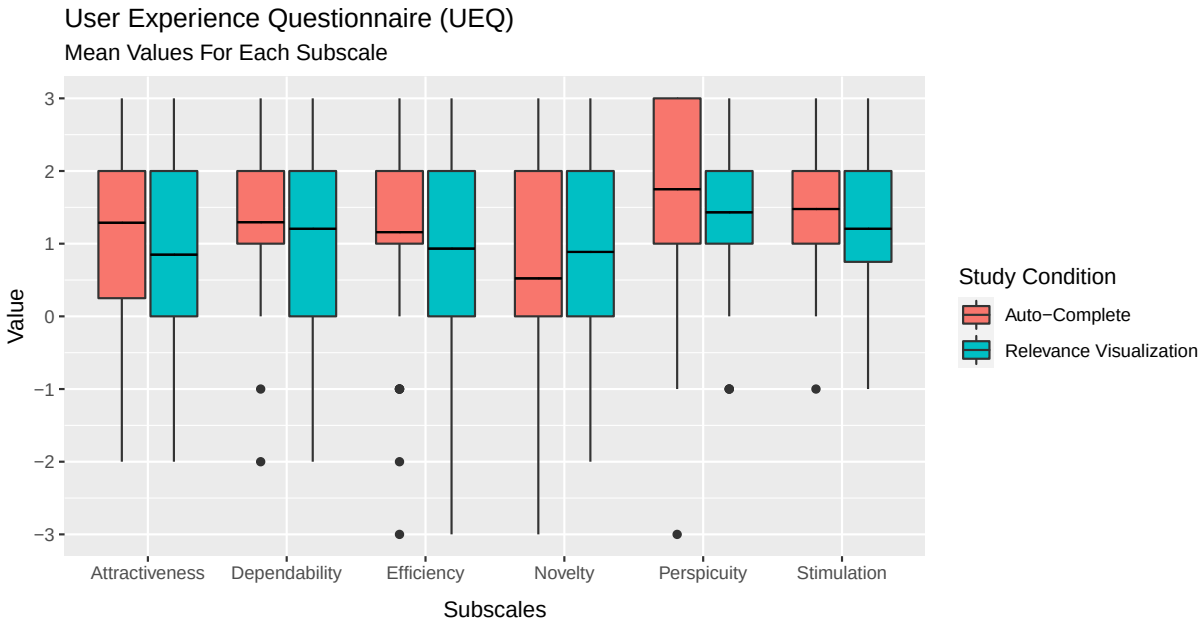

Figure 6.2: In this figure, a boxplot is presented for the two study conditions for each subscale of the UEQ. These subscales are *attractiveness*, *dependability*, *efficiency*, *novelty*, *perspicuity*, and *simulation*, and can be found on the x-axis. On the y-axis the scores of the 7-point Likert scale of the UEQ can be found. For each boxplot I present the confidence interval, the range, and outliers. The black bar inside the box plot represents the mean.

Scale	Auto-Complete	Relevance Visualization
Attractiveness	1.29 (SD: 1.08, α: 0.81)	0.85 (SD: 1.06, α: 0.90)
Dependability	1.30 (SD: 1.00, α: 0.65)	1.20 (SD: 1.17, α: 0.70)
Efficiency	1.16 (SD: 1.36, α: 0.50)	0.93 (SD: 1.44, α: 0.70)
Novelty	0.52 (SD: 1.47, α: 0.45)	0.89 (SD: 1.26, α: 0.79)
Perspicuity	1.75 (SD: 1.35, α: 0.56)	1.43 (SD: 1.02, α: 0.77)
Stimulation	1.48 (SD: 1.00, α: 0.57)	1.20 (SD: 0.95, α: 0.80)

Table 6.2: Means, SD (in parenthesis) and Cronbach's alpha (in parenthesis) for each subscale of the UEQ for both study conditions.

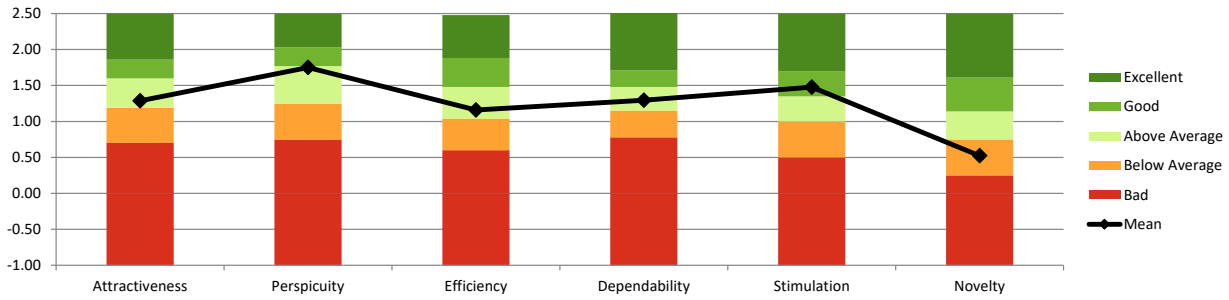

Figure 6.3: In this figure, a barplot is presented for the benchmark results for the UEQ administered after the participants were exposed to the system with the auto-complete functionality. The y-axis shows an excerpt of the normalized values of the 7 point Likert-scale (3.0: highest positive value on dimension; -3.0: highest negative value). The results show that the system achieves good results for *Stimulation* (10% of benchmarked systems better, 75% worse), results above average for *Attractiveness*, *Perspicuity*, *Efficiency* and *Dependability* (25% better,50% worse), and results below average for *Novelty* (50% better, 25% worse). This template for this plot was provided by Laugwitz et al. (2008).

study conditions vary between 0.96 and 1.47, and do not show a visible difference for the two study conditions.

The authors of the UEQ provide tools to benchmark the personal results of the UEQ compared to 452 existing values from other systems and prototypes (Laugwitz et al., 2008). For the UEQ that I administered in the auto-complete condition benchmark results can be found in Figure 6.3. The results show that the system with the auto-complete functionality achieves good results for *stimulation* (10% benchmark systems better, 75% worse), results above average for *attractiveness*, *perspicuity*, *efficiency* and *dependability* (25% better, 50% worse), and results below average for *novelty* (50% better, 25% worse). For the UEQ administered in relevance visualization condition the benchmark results show that the system achieves results above average for *perspicuity*, *dependability*, *stimulation*, and *novelty* (25% better, 50% worse), and results below average for *attractiveness* and *efficiency* (50% better, 25% worse).

Comparing Means Of Study Conditions

Despite of the concerns mentioned at the beginning of the chapter (re:comparability of the conditions), as with the HCT analysis, I still run dependent tests to find out if there are

differences in means between the two study conditions for each subscale. The assumption of normality for the differences between the values of the two study conditions was only satisfied for the *attractiveness* subscale. However, the differences for this scale only contain four different values, and so I used the non-parametric Wilcoxon signed-rank test for all subscales (Field et al., 2012). There was a significant difference between the *attractiveness* of the system in the auto-complete condition and in the relevance visualization condition ($V = 573$, $p < .001$). The effect size is small ($r = .19$). All other tests for other subscales returned non-significant results (all $V > 55$, $p > .05$).

Relevant Single Items

It is worth highlighting single items within the subscales. To avoid 'cherry-picking' I highlight items that are 0.5 points from the mean for each study condition or from each other. First, participants found the system easy to learn in the auto-complete condition ($M = 2.5$, $SD = 0.7$) which decreased in the relevance visualization condition ($M = 1.7$, $SD = 1.0$). Second, participants found the system interesting in the auto-complete condition ($M = 2.0$, $SD = 0.8$) which decreased in the relevance visualization condition ($M = 1.4$, $SD = 0.8$). Third, participants found the system neither fast nor slow in the auto-complete condition ($M = 0.2$, $SD = 1.7$) which increased in the relevance visualization condition ($M = 0.4$, $SD = 1.7$). Note here, this was the item with the highest standard deviation. Fourth, participants found the system pleasant in the auto-complete condition ($M = 1.5$, $SD = 0.9$) which decreased in the relevance visualization condition ($M = 0.9$, $SD = 1.0$). Fifth, participants found the system efficient in the auto-complete condition ($M = 1.2$, $SD = 1.5$) which decreased in the relevance visualization condition ($M = 0.5$, $SD = 1.6$). Sixth, participants found the system practical in the auto-complete condition ($M = 1.8$, $SD = 0.6$) which decreased in the relevance visualization condition ($M = 1.6$, $SD = 0.8$).

6.1.3 AI Expertise

Meanwhile, the self-report on the expertise and attitude towards AI reflected the findings of the thematic analysis. On a 7-point Likert scale (1 = *Strongly Disagree*, 7 = *Strongly Agree*) I found the following results. While most participants have not directly worked with AI frameworks like PyTorch or Tensorflow ($M = 2.82$, $SD = 2.04$), they have worked with programming languages like Python ($M = 5.00$, $SD = 2.45$). With regards to AI, participants agreed that AI can solve difficult problems ($M = 5.55$, $SD = 1.92$), but were undecided whether AI will ever reach human intelligence ($M = 3.45$, $SD = 1.44$).

Practice Question	1	2	3	4						
Average time in s	147	65	65	68						
Auto-Complete	5	6	7	8	9	10	11	12	13	14
Average time in s	152	30	57	41	48	112	59	38	60	52
Relevance Visualization	15	16	17	18	19	20	21	22	23	24
Average time in s	659	81	75	47	65	82	81	35	76	47

Table 6.3: Average time in seconds per question for all participants. In total, there were 24 questions. Questions 1, 5, and 15 introduced the new study condition.

Participant	Number of Auto-Complete	Accepted All	Accepted Parts
0	50	12 (24%)	20 (40%)
1	51	17 (33%)	6 (12%)
2	37	12 (32%)	11 (30%)
3	82	29 (35%)	18 (22%)
4	44	20 (45%)	8 (18%)
5	75	20 (27%)	11 (15%)
6	54	0 (0%)	31 (57%)
7	56	24 (43%)	7 (13%)
8	47	25 (53%)	1 (2%)
9	85	20 (24%)	5 (6%)
10	83	23 (28%)	4 (5%)

Table 6.4: Number of times each participant triggered the AI auto-complete functionality, how often they accepted the suggested text entirely (by pressing `tab`), and how often they accepted parts of the generated answer (by clicking on it with the cursor). The relative values are given in parentheses.

On average, participants did somewhat disagree with the notion that AI is human-like ($M = 3.00$, $SD = 1.79$) or has intentions ($M = 3.27$, $SD = 1.79$).

6.2 Log Data

Table 6.3 shows the average time per question that participants took. As can be seen, the average time is higher for questions 1, 5, and 15. The interface and its features were introduced to participants while asking these questions. Furthermore, participants also had the option to ask questions and receive feedback on the interface and the task while being exposed to question 2, 6, and 16. The average times for these questions do not visibly differ from the subsequent questions, however. The average time that each participant spent on each question (excluding 1, 2, 5, 6, 15, 16) is 61 seconds ($SD = 20.3$ seconds). The fastest participant took 26 seconds per question on average. The slowest participant took 88 seconds per question on average.

In total, participants triggered the AI 664 times to provide auto-completed text. Of those 664 times participants accepted the whole suggestion (by pressing `tab`) 202 times (30.4%). Furthermore, participants accepted parts of the suggested answer (by clicking on parts of the suggested text with the cursor) 122 times (18.4%). This means that participants ignored the AI (by 'writing over' it or finishing the questions) around half of the time (51.2%). In Table 6.4, viewers can see that the amount of acceptance of AI output varied between participants indicating different modes of user interaction with the system.

6.3 Summary

In this chapter, I presented the quantitative results of this study. Generally, the results show that the system was liked and found both reliable and easy to understand in both study conditions. Specifically, the system was more attractive to participants without the relevance visualization ($r = .19$). Finally, the log data shows that participants frequently triggered the artificial intelligence (AI) and used at least some parts of the generated answer around half of the time.

In the next chapter, I will discuss the findings of this and the previous chapter and connect them to my introduction and other related research.

Chapter 7

Discussion

My research through design (RtD) process, the thematic analysis of my HCI expert responses, and the evaluation of quantitative data, despite challenges due to COVID-19, revealed that the artificial intelligence (AI) assistant was generally able to support participants in the task of answering customer questions. In this chapter, I discuss *how* my design was able to achieve this, reflect on the design process and discuss how it can inform other designs and research in general. Specifically, I will look into the issue of 'interpretability': while the system was generally perceived as easy to understand (see HCT) and perspicuous (see UEQ), the addition of the relevance visualization dampened the positive perception of the system (i.e., the system was less attractive with the visualization). Before I discuss and speculate *why* the relevance visualization decreased the positive perception of the system (see Explaining Why Not) I will discuss how the system with the auto-complete functionality fulfills Parasuraman's (2000) evaluation criteria of automation design and discuss the need for 'invisible design'.

7.1 Evaluating Human-Centered Automation Design Criteria

In this section, I evaluate the interface and the auto-complete functionality with Parasuraman et al.'s (2000) evaluation criteria of automation design. These criteria include the mental workload for the human operator, the situation awareness of the operators, how complacent they become, their skill degradation for relying on the AI, the consequences of a wrong decision, and the reliability of the automation. While I provide strong evidence

for some of the criteria, I also speculate about others by leveraging what I learned in my design process.

Mental Workload. I did not directly measure the mental workload of participants when they interacted with the interface. Nonetheless, participants displayed two interaction strategies that I speculate reduced their mental workload. First, participants relied on the AI to provide an answer draft that could be improved upon rather than creating an answer from the ground up. Participants described this way of creating answers as easier and more efficient which indicates a reduction in the workload. Second, participants were able to simply ignore the auto-completed text when they wanted to write their own answer (around 50% of the time). This was enabled by my non-intrusive design (cf., Clark et al., 2018) which reduces the risk that users spend more time correcting or guiding the AI than benefiting from the AI's contributions (cf., Kocielnik et al., 2019). This further decreases the workload.

Situation Awareness. If systems and AI are not fully autonomous, human operators are required to take control in specific situations. This transition presents a challenge as users 'zoom out' and are not fully aware of the situation if they rely on automation. This issue is also critical in my system: users that rely on the AI to create answers to customer queries might not be fully aware of the source text. While this might not be an issue for experienced customer service workers (as they generally know the source text), inexperienced workers either have to read the entire source text (which diminishes the returns of deploying AI) or 'somewhat trust' the AI. My participants tried to overcome this issue by scanning the text for keywords they identified in the query and 'reading around' these keywords in relevant text passages. However, it is unclear how users would react to longer source texts and over time.

Complacency. Even in my short study, there was evidence for complacency. For example, P8 showed a high acceptance rate of the entire AI-generated answer (53%) which resulted in the second-lowest average time spent on each question (28.2 s). It can be argued here that P8 simply accepted what the AI provided without thoroughly checking if the AI-generated answer was correct and satisfying. Generally, this presents a challenge to all designers if the AI is 'good enough', but only 'most of the time'.

Skill Degradation. I speculate that the reliance on the AI does not constitute a significant threat to the general skill of answering questions. Because of a lack of situational awareness, however, I argue that users may develop less knowledge and expertise in customer service if they rely too much on the AI and do not familiarize themselves with the source text.

The Consequences of a Wrong Decision. The criticality of wrong decisions of text assistants like my customer query assistant is arguably lower than for life-threatening applications such as in transportation or medicine. Nonetheless, the number of wrong or unsatisfying answers should be reduced to increase customer satisfaction. To achieve this my design gives the final approval decision to the user which gives the user the chance to correct the AI if necessary (cf., level 5 of Sheridan et al.'s levels of automation, 1978). Furthermore, as users can simply ignore answers they think are incorrect or unsatisfactory the cost of wrong decisions by the AI is low. This enables users to interact with AI that sometimes produces unsatisfactory output (because of a lack of competence).

Reliability of Automation. Participants found the AI somewhat reliable (human-computer trust (HCT), $M = 4.80$, $SD = 1.41$) to provide a competent enough answer. While the reliability of the automation is a key factor in the understanding of and interaction with the AI (see Reliability over Interpretability in Low Risk Environments), my design enabled participants to test the AI noncommittally (cf., Mesbah et al., 2019) and gain first-hand experience (cf., Kocielnik et al., 2019). I argue that this back-and-forth interaction allowed participants to get an estimate of the reliability of the AI and decreased the time it took participants to learn how to successfully interact with the AI. For example, this could be observed when P9 mentions that they were able to learn how much information they had to provide to the AI for it to finish the answer.

After evaluating my design in the light of these human-centered automation criteria, I will synthesize and discuss the findings of participants' perception of the system and the auto-complete functionality.

7.2 Synthesis of the System and the Auto-Complete Functionality

I argue that the success of my system is the result of the design of the interface that allowed participants to collaborate with the AI in different ways. I draw this conclusion from the combined fact that participants described the AI as being useful for the task but would not rely on it blindly. Thus, it was the interface that allowed participants to leverage the AI to meet their needs, even though the AI would not have been 'good enough', had it directly faced customers. In this section, I discuss how these lessons learned can be generalized to other domains and prototypes.

My prototype is part of a group of AI assistants that live in low-risk environments. As described above, this means that the consequences of making errors are small for these assistants. These characteristics make them unique to other AI assistants (such as in medicine or finance) and fundamentally changes the human interaction with it. In my study, I confirmed the beliefs of Kocielnik et al. (2019): for passive systems in which the user makes the final decision participants prefer receiving sufficient suggestions by the AI that are satisfactory even if that means that the AI also provides unsatisfactory suggestions sometimes (i.e., optimization for high recall, low false negatives). This comes with the presumption that the (mental) workload of filtering and discarding false positives is low. With regards to my prototype, this would mean that the AI always tries to auto-complete the text input by the user while it is easy for the user to discard unsatisfactory answer completions. Here, always providing some auto-complete is preferable to only providing assistance if there is a high chance the auto-complete is correct.

I argue that the auto-complete functionality is ideal for this purpose. Users are able to receive a new suggestion every time they provide a new text input *because* it is easy to ignore the AI if the AI output is unsatisfactory. On the other hand, were the interaction intrusive and demanding, the amount of suggestions/auto-completes would likely need to be decreased. In contrast to Kocielnik et al. (2019), I speculate that low-risk systems with low costs to verify AI output are unique in that they are likely the only class of systems in which a high recall can be preferred over high precision. Note, that this statement has consequences for the interpretability of the system (cf., Explaining Why Not).

Before discussing the interpretability of my AI assistant it is worth looking deeper into the benefits of my design. While I argue that the design of the interface specifically enabled participants to leverage the AI, almost no participant explicitly stated that they were able to leverage the AI *because* the design allowed them to do so. Only one participant mentioned that they were satisfied with how they can edit the answer that the AI produced. Nevertheless, I believe the interface was well-designed as Norman writes in his popular book *The Design of Everyday Things* (2002): "Good design is actually a lot harder to notice than poor design, in part because good designs fit my needs so well that the design is invisible". In other words, as evidence for enabling design I can recognize the *lack* of criticism: none of the participants criticized the way the AI could be triggered, corrected, or discarded. This stands in contrast to the stark criticism of other design aspects of the interface (e.g., the interaction with the visualization) and highlights the usefulness of the auto-complete functionality.

For other human-AI collaborations, an 'invisible' design can similarly enable interaction with AI in different situations with different intentions. For example, in my study, whether participants used the AI to provide a draft as a starting point, or whether they leveraged

the AI to 'copy' parts of the source article, did not change the way in which participants accepted, edited or discarded the output given by the AI. Further strengthening its practicality, my design also collocates interactive text components and AI-generated content, which allows interaction with the AI *while* and *by* writing. Finally, the design of my interface allowed participants to leverage AI differently with regards to the problem they faced (i.e., creating vs. finishing the answer) while they were interacting with the same AI model.

In general, I argue that thinking about how to closely integrate responses from the AI with interactive components that the user can directly manipulate can help make the interaction with the AI 'disappear'. Note here, that this fundamentally changes the interaction with the AI: Instead of interacting with the AI through additional UI elements (e.g., buttons) users can interact with the AI through activities related to the fulfillment of the task (such as writing). While in language tasks this can be achieved relatively easily by an auto-complete functionality, I speculate that the idea of completing user input that is easy to accept, edit, and discard can be applied in other domains as well. For example, an AI assistant could try to finish the user's drawing directly on the canvas (Oh et al., 2018). Another example would be the composition of music (McCormack et al., 2019).

This advice may seem to contradict the idea of interpretability, as the intention seems to be to 'hide' the AI agent, but in my design process, I found that further integrating the AI into the task has the opposite effect of participants feeling more in control. This may be unique to generative tasks like answering customer queries and different in classification tasks that entail high risk (e.g., disease diagnosis). Therefore, it might be helpful to AI designers to analyze the type of task (e.g., generative or classification), the costs of wrong decisions, and the costs to discard them to decide whether users should 'just get the task done' as efficiently as possible or whether users should prioritize avoiding all wrong decisions. In the following section, I will analyze the interpretability of the system and compare it to other systems to conceptualize a better goal of interpretability.

7.3 Explaining Why Not

In my design, I attempted to integrate the idea of interpretability (of the model's behaviour) through the relevance visualization. The relevance visualization had mixed feedback with some significant criticism from my participants. Here, the findings of the HCT and the user-experience questionnaire (UEQ) questionnaires somewhat contradict the findings of the thematic analysis at first glance. Specifically, the results of the questionnaires do not differ much from the results before the visualization was added as a study condition,

while participants strongly criticized the visualization in the semi-structured interview. There may be many different reasons for this apparent contradiction. First, participants described learning effects when interacting with the system (see section 8.2). Second, both questionnaires asked participants to evaluate the whole system (and not just the visualization). This drawback makes it difficult to refer the feedback to the visualization itself as participants were able to make progress on the task without paying much attention to the visualization. Next, I will discuss the feedback I received on the visualization from the interview and interpret it given the context the design was deployed in and who interacted with it.

In hindsight, my relevance visualization design seems to be cumbersome. While my intention was that participants can, at any given time, stop to receive more information on the AI's output, the design choices that were made forced participants to break their task routine. This design stands in contrast to the 'seamless interaction' between the AI and users through the auto-complete functionality. Thus, a more integrated process of triggering the visualization (such as pressing a shortcut key or a toggle button to automatically display the visualization) could have allowed participants to stay in their routine. The possibility to turn off the automatic visualization would have also honoured the findings of Oh et al. (2018) who state that participants like to be in control of when the the AI 'explains its intentions'.

In addition to the cumbersome interaction with the relevance visualization, the critique of the interpretability of the visualization also reveals shortcomings of explainable AI more generally. In my design, interpretability actually suffered as participants were not able to understand how the AI used the highlighted words to derive an answer. The underlying conflict of this 'misunderstanding' is subtle but significant: understanding requires counterfactual reasoning that current neural networks such as the Transformer are not able to provide (Miller, 2017; Pearl & Mackenzie, 2018). According to Pearl and Mackenzie (2018), these 'learning machines' only operate by association of input and output. In contrast, understanding requires agents to 'imagine worlds that do not exist and infer reasons for observed phenomena'. Thus, to answer *why* my AI created a specific answer would require it to imagine a counterfactual case that lays open the cause for creating the answer (Miller, 2017). In other words, the AI needs to provide a counterfactual example to explain why it *did not* create some other answer. Equally, this problem might appear in 'white-box' designs of explainable artificial intelligence (XAI) in other domains. For example, given a saliency map of an image classifier designed to explain why an object in the image was classified a certain way by highlighting the relevant pixels in the image used by the classifier, I might ask, why did it not classify the object as something else?

The consequences of this 'over-promising of explainability' can be severe and damaging to the reputation of the field. In my study, I found that some participants simply claimed that the visualization would 'explain the AI' without giving a satisfactory answer to how this was achieved. Though the process of explanation might not be intuitive to participants, and therefore difficult to describe, these findings indicate how easy it was for some participants to trust an 'explanation' on no formidable grounds. This finding is in line with other research. For example, Kaur et al. (2020) found that people assume that an AI model can be trusted more if there is any 'explanation' provided. Similarly, Koehler (1991) found that simply presenting any explanation for a proposition strengthens people's belief that the proposition is true. Finally, in my study some participants tried to match highlighted words in the visualization with their expectation of them to infer something about the correctness of the answer for which there are no grounds based in causality.

These findings indicate that some participants trust the AI without evaluating the validity of the given explanation. The consequence is that people can misuse visualizations and other 'explanations' to come up with incorrect assumptions about the AI model, the dataset or interpretability itself (Kaur et al., 2020). This misuse could potentially lead to severe consequences if these assumptions lead to poor decisions. According to Kaur et al. (2020), it can be prevented if people 'think critically' and try to evaluate the visualization thoroughly. I argue that the field can facilitate this by avoiding over-promising XAI with regards to deep learning (DL). Ideally, this will increase the number of people that verify the explanation, but also decrease the number of people that are frustrated by and disappointed in the explanation (after they evaluate it critically).

7.3.1 Reliability over Interpretability in Low Risk Environments

My human-centred design process has showcased that the current state-of-the-art language model may not be capable of 'explaining' its behaviour as Miller (2017) lays out. While AI interpretability is fundamental in high-risk environments, my research indicates that full interpretability might not be necessary for successful human-AI collaboration in low-risk environments. This remedy is built on two factors. The first, rather trivial, factor is that the cost of making a poor decision is low. That means that the AI can 'over-compensate' by providing more output instead of ensuring the correctness of every output (see section 7.2). Second, in generative tasks (such as answering questions, drawing, etc.) participants can take the AI output at face value. For example, some participants in my research specifically reported that the output of the AI was 'enough' information to understand the AI's capabilities. This 'explanation' was enabled by first-hand experience (Kaur et al., 2020):

participants could freely experiment with the AI to receive an understanding of its general capabilities.

This finding also relates back to the concept of supervisory control (Sheridan et al., 1978). Without knowing how the machine fulfills a simple task, humans are able to ignore this lack of knowledge and to collaborate with the machine if it reliably fulfills the task (Sheridan et al., 1978). Knowing how to handle a machine is enough to interact with it: if you drive a car, for example, you need to know how to drive it but do not need to understand how the engine works. I argue this might be similar to generative AI in low-risk environments.

Despite this shift in design intent, I do still think feedback mechanisms from the field of XAI like the relevance visualization can be useful. My research has shown that participants used the relevance visualization to find relevant parts of text that the AI considered when creating the answer. In this interaction, the visualization served as a mechanism to help participants orient themselves in the source article. The purpose of this interaction is different from the original purpose of XAI: instead of explaining how and why the AI produced a specific output, the visualization helps participants browse data to verify whether the AI produced the right answer. Applying this paradigm shift to the task of image classification, saliency maps could be designed to help highlight the position of the object within the image, rather than to provide an explanation of an AI algorithm. I argue that this can help increase the situational awareness of participants by directing the intention of humans towards relevant data.

My findings suggest that the main benefit of a feedback mechanism like the relevance visualization is to ease the verification of AI output by browsing the source data in contrast to understanding AI output (i.e., why was this generated?). This also includes a shift in focus from the AI (and its 'intentions') towards the data and the task. I argue that this helps humans to collaborate with AI by directly interacting with its output rather than having to guess the AI's intentions. In my sample and task, this was already facilitated by the high technical competence of my participants and their perception of AI as being "just like a tool". In this sense, it might be beneficial for designers to create non-anthropomorphized AI to direct the focus of participants on system-like attributes like usefulness and reliability over human-like ones like integrity (Lankton et al., 2015).

7.4 Summary

In this chapter, I discussed and synthesized the findings of my study with regards to other research and contexts. I found that the strength of the auto-complete functionality

lays in the ability to easily trigger, edit, and discard artificial intelligence (AI) output. This integrates the interaction with the AI with the task and makes it 'seamless'. I then argue that this form of interaction can be beneficial to other AI assistants in low-risk environments. Consequently, I discussed the value and flaws of the relevance visualization and make a plea to end the over-promising of AI interpretability. Instead, I highlight the importance of reliable AI, supervisory control, and a focus on the data and the task in low-risk environments instead of intentions of AI.

In the next chapter, I will list the contributions made by this research, name limitations to it, and provide suggestions on how to overcome them in future research.

Chapter 8

Conclusion & Future Work

In this research, I followed a research through design (RtD) process to address some of the challenges of human-artificial intelligence (AI) collaboration and explored specific designs for leveraging AI to answer customer queries as customer service representatives. Specifically, I developed an auto-complete functionality to easily allow participants to invoke, guide, and dismiss AI assistance, collocated the workspace of the AI and human to speed up the interaction process and allow fast editing of AI output, and a relevance visualization to verify AI output. I found that participants liked leveraging the AI to provide a first draft of the answer or to finish a sentence, guided AI to produce better answers, and perceived AI "just like a tool" which allowed them to take AI output at face value. My findings suggest that thinking about how to integrate the AI into the task and to make the interaction with the AI "invisible" may actually improve the experience, rather than hinder explainability in low-risk environments. Finally, I discuss the need for reliable, non-anthropomorphized AI over fully explainable AI.

8.1 Contributions

My work produced the following main contributions to research (listed extensively):

1. I followed a RtD process to design and develop an interface that enables users to collaborate with AI to answer customer queries. In this process, I conceptualized the capabilities of the Transformer model, trained the model, described 'ideal' human-AI collaboration, and produced interaction designs to collaborate with the model

(i.e., auto-complete functionality) through many cycles of design, development, and evaluation. Specifically, I co-located the workspaces of the AI and the human to improve the editing of AI output and created an auto-complete functionality to trigger the AI and accept or discard its output. I describe the RtD process and my design decisions in detail.

2. I present a study with HCI experts that demonstrates that participants leveraged the AI through the design to provide a first draft to their answer and to finish sentences that they intended to write. In doing so, I found that most participants perceived the AI "just like a tool" which allowed them to have supervisory control over the AI. This control was enabled by integrating the interaction with the AI into the task itself. I then discussed the benefit of invoking the AI through activities related to the fulfillment of the task (e.g., writing) and how this integration with the task can make the interaction with the AI 'invisible'. I also discuss the advantage of high-recall systems in low-risk environments if the costs of recovering from failure is low and how these systems can mitigate unsatisfactory AI output.

3. I developed a post-hoc explanation to give insight into the behaviour of the AI. I presented and discussed the results of a thematic analysis and quantitative evaluation of questionnaire data, which showed that users tried to match their mental model of what the visualization is supposed to highlight with what it actually highlighted. If the mental model matched, some participants became more confident. If it did not match, some participants became confused and frustrated. I then discussed how this explanation failed to 'explain' the AI and made a call to stop the over-promising of explainable artificial intelligence (XAI) related to neural nets. Instead, I advocated for reliable AI in low-risk environments that can be taken at face value and the need to give tools to users to easily verify AI behavior by browsing input data.

8.2 Limitations & Future Work

The generalizabilty of this research was limited by the homogeneous, highly technical sample of participants. Though I would have preferred an in-person study with a a more diverse sample to increase the study's external validity, the ongoing COVID-19 global pandemic, which halted all on-campus, in-person research activities such as user studies, and disrupted or delayed many other non-essential research activities (i.e. non-COVID-related research) blocked my preference's realization. Nonetheless, I believe the feedback of these HCI experts was valuable, yet the positive sentiment towards AI as well as the

ability to perceive AI as a tool might be exclusive to this group of people. In addition to this specific attitude towards AI, participants were part of the same research group as the researchers. Thus, replies from participants might have been biased, especially if participants knew "the research was about XAI". Throughout this research I tried to be critical of this aspect. For example, I did not blindly accept the reasoning of one participant who claimed that the relevance visualization *explained* AI without providing any reasonable explanation of *how* this was achieved. In future studies, it will be interesting to see how a more heterogeneous sample perceives AI and how this perception changes human-AI collaboration.

Another limiting factor is the potential learning effect that impacted the interaction between participants and AI. Specifically, three participants noted that they improved their ability to interact with the AI. One participant mentioned that this might have polarized answers given in the second questionnaire. This might explain why the values for the human-computer trust (HCT) and the user-experience questionnaire (UEQ) were generally not significantly different. While my study design was a conscious decision to give participants the ability to contrast two different designs in the interview (having the AI vs. having the AI plus relevance visualization) I recognize these learning effects. Future, between-participant studies may be needed to determine whether my findings can be measured quantitatively in effect sizes between independent study groups.

8.3 Closing Remarks

This thesis describes an attempt to evaluate human-AI collaboration in a realistic task setting. While AI can be leveraged in successful interaction design today, the interaction with it remains mechanical. Though this is not disadvantageous for simple, low-risk tasks, we must develop different AI models for more complex, 'human-like' tasks. These models will likely rely on deep learning (DL) (e.g., supervisory training of neural nets), but also rely on other technologies and models as well. To improve current AI models it might be worth looking into how knowledge can be represented in different ways for AI to reason about it in more abstract ways.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda, In *Proceedings of the 2018 chi conference on human factors in computing systems*, Montreal QC, Canada, Association for Computing Machinery. https://doi.org/10.1145/3173574.3174156

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-ai interaction, In *Proceedings of the 2019 chi conference on human factors in computing systems*, Glasgow, Scotland Uk, Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Arnold, K. C., Chauncey, K., & Gajos, K. Z. (2020). Predictive text encourages predictable writing, In *Proceedings of the 25th international conference on intelligent user interfaces*, Cagliari, Italy, Association for Computing Machinery. https://doi.org/10.1145/3377325.3377523

Atoyan, H., Duquet, J.-R., & Robert, J.-M. (2006). Trust in new decision aid systems, In *Proceedings of the 18th conference on l'interaction homme-machine*, Montreal, Canada, Association for Computing Machinery. https://doi.org/10.1145/1132736.1132751

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate (Y. Bengio & Y. LeCun, Eds.). In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings*. http://arxiv.org/abs/1409.0473

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., & et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Berge, S. H. (2018). *Rise of the chatbots: Trust in artificial intelligence during extreme weather events* (Master's thesis). http://urn.nb.no/URN:NBN:no-65934

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Józefowicz, R., & Bengio, S. (2015). Generating sentences from a continuous space. *CoRR*, *abs/1511.06349*arXiv 1511.06349. http://arxiv.org/abs/1511.06349

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chien, S.-Y., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014). Towards the development of an inter-cultural scale to measure trust in automation (P. L. P. Rau, Ed.). In P. L. P. Rau (Ed.), *Cross-cultural design*, Cham, Springer International Publishing.

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. *CoRR*, *abs/1904.10509*arXiv 1904.10509. http://arxiv.org/abs/1904.10509

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human–Computer Interaction*, *31*(10), https://doi.org/10.1080/10447318.2015.1070549, 692–702. https://doi.org/10.1080/10447318.2015.1070549

Clark, E., Ross, A. S., Tan, C., Ji, Y., & Smith, N. A. (2018). Creative writing with a machine in the loop: Case studies on slogans and stories, In *23rd international conference on intelligent user interfaces*, Tokyo, Japan, Association for Computing Machinery. https://doi.org/10.1145/3172944.3172983

Company, M.
bibinitperiod. (2020). Ai adoption advances, but foundational barriers remain. https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r*. Sage publications.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*.

Garfield, M. J. (2008). Creativity support systems. In *Handbook on decision support systems 2: Variations* (pp. 745–758). Berlin, Heidelberg, Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-48716-6_34

Gaver, W. (2012). What should we expect from research through design?, In *Proceedings of the sigchi conference on human factors in computing systems*, Austin, Texas, USA, Association for Computing Machinery. https://doi.org/10.1145/2207676.2208538

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When will AI exceed human performance? evidence from AI experts. *CoRR*, *abs/1705.08807*arXiv 1705.08807. http://arxiv.org/abs/1705.08807

Grice, H. P. (1975). Logic and conversation, In *Speech acts*. New York: Academic Press.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, *2*.

Ha, D., & Eck, D. (2017). A neural representation of sketch drawings. *CoRR*, *abs/1704.03477*. http://arxiv.org/abs/1704.03477

Halpern, J. Y., & Pearl, J. (2005a). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887. https://doi.org/10.1093/bjps/axi147

Halpern, J. Y., & Pearl, J. (2005b). Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, *56*(4), 889–911. https://doi.org/10.1093/bjps/axi148

Harris, J. J. (2019). Leveraging asymmetry and interdependence to enhance social connectedness in cooperative digital games.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., ... Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. *CoRR*, *abs/1803.05567*arXiv 1803.05567. http://arxiv.org/abs/1803.05567

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, *abs/1502.01852*arXiv 1502.01852. http://arxiv.org/abs/1502.01852

Hofstadter, D. R. (1979). *Godel, escher, bach: An eternal golden braid*. USA, Basic Books, Inc.

Holmquist, L. E. (2017). Intelligence on tap: Artificial intelligence as a new design material. *Interactions*, *24*(4), 28–33. https://doi.org/10.1145/3085571

Horvitz, E. (1999). Principles of mixed-initiative user interfaces, In *Proceedings of the sigchi conference on human factors in computing systems*, Pittsburgh, Pennsylvania, USA, Association for Computing Machinery. https://doi.org/10.1145/302979.303030

Jeong, H., Park, J., Park, J., Pham, T., & Lee, B. C. (2019a). Analysis of trust in automation survey instruments using semantic network analysis (I. L. Nunes, Ed.). In I. L. Nunes (Ed.), *Advances in human factors and systems interaction*, Cham, Springer International Publishing.

Jeong, H., Park, J., Park, J., Pham, T., & Lee, B. C. (2019b). Analysis of trust in automation survey instruments using semantic network analysis (I. L. Nunes, Ed.). In I. L.

Nunes (Ed.), *Advances in human factors and systems interaction*, Cham, Springer International Publishing.

Jian, J.-Y., Bisantz, A., & Drury, C. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*, 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning, In *Chi 2020*. https://www.microsoft.com/en-us/research/publication/interpreting-interpretability-understanding-data-scientists-use-of-interpretability-tools-for-machine-learning/

Kim, B. (2015). *Interactive and interpretable machine learning models for human machine collaboration* (Doctoral dissertation). Massachusetts Institute of Technology.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013). The future of crowd work, In *Proceedings of the 2013 conference on computer supported cooperative work*, San Antonio, Texas, USA, Association for Computing Machinery. https://doi.org/10.1145/2441776.2441923

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems, In *Proceedings of the 2019 chi conference on human factors in computing systems*, Glasgow, Scotland Uk, Association for Computing Machinery. https://doi.org/10.1145/3290605.3300641

Koehler, D. (1991). Explanation, imagination, and confidence in judgment. *Psychological bulletin*, *110*(3), 499–519. https://doi.org/10.1037/0033-2909.110.3.499

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations.

Lankton, N., Mcknight, D., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*, 880–918. https://doi.org/10.17705/1jais.00411

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire (A. Holzinger, Ed.). In A. Holzinger (Ed.), *Hci and usability for education and work*, Berlin, Heidelberg, Springer Berlin Heidelberg.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–44. https://doi.org/10.1038/nature14539

Linegang, M. P., Stoner, H. A., Patterson, M. J., Seppelt, B. D., Hoffman, J. D., Crittendon, Z. B., & Lee, J. D. (2006). Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(23), 2482–2486. https://doi.org/10.1177/154193120605002304

Lipton, Z. C. (2016). The mythos of model interpretability. *CoRR*, *abs/1606.03490*arXiv 1606.03490. http://arxiv.org/abs/1606.03490

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, *55*(3), 232–257.

Lu, L., Cai, R., & Gursoy, D. (2019). Developing and validating a service robot integration willingness scale. *International Journal of Hospitality Management*, *80*, 36–51. https://doi.org/https://doi.org/10.1016/j.ijhm.2019.01.005

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust, In *11th australasian conference on information systems*. Citeseer.

McCormack, J., Gifford, T., Hutchings, P., Llano Rodriguez, M. T., Yee-King, M., & d'Inverno, M. (2019). In a silent way: Communication between ai and improvising musicians beyond sound, In *Proceedings of the 2019 chi conference on human factors in computing systems*, Glasgow, Scotland Uk, Association for Computing Machinery. https://doi.org/10.1145/3290605.3300268

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manage. Inf. Syst.*, *2*(2). https://doi.org/10.1145/1985347.1985353

Mesbah, N., Tauchert, C., Olt, C. M., & Buxmann, P. (2019). *Promoting Trust in AI-based Expert Systems* (Publications of Darmstadt Technical University, Institute for Business Studies (BWL) No. 113208). Darmstadt Technical University, Department of Business Administration, Economics and Law, Institute for Business Studies (BWL). https://ideas.repec.org/p/dar/wpaper/113208.html

Miller, C. A., & Parasuraman, R. (2003). Beyond levels of automation: An architecture for more flexible human-automation collaboration. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(1), 182–186. https://doi.org/10.1177/154193120304700138

Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, *abs/1706.07269*arXiv 1706.07269. http://arxiv.org/abs/1706.07269

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *CoRR*, *abs/1712.00547*arXiv 1712.00547. http://arxiv.org/abs/1712.00547

Morgan, B. (2020). Customer service is a $350 billion industry, and it's a mess. https://www.forbes.com/sites/blakemorgan/2017/09/25/customer-service-is-a-350b-industry-and-its-a-mess/#4ccebe4211be

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, *abs/1611.09268*arXiv 1611.09268. http://arxiv.org/abs/1611.09268

Norman, D. A. (2002). *The design of everyday things*. USA, Basic Books, Inc.

Oh, C., Song, J., Choi, J., Kim, S., Lee, S., & Suh, B. (2018). I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence, In *Proceedings of the 2018 chi conference on human factors in computing systems*, Montreal QC, Canada, Association for Computing Machinery. https://doi.org/10.1145/3173574.3174223

Omale, G. (2020). Gartner predicts 25 percent of digital workers will use virtual employee assistants daily by 2021. https://gtnr.it/3b5C46K

Parasuraman, R., Sheridan, T., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297.

Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce*, *7*(3), https://doi.org/10.1080/10864415.2003.11044275, 101–134. https://doi.org/10.1080/10864415.2003.11044275

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (1st). USA, Basic Books, Inc.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. M. (2018). Manipulating and measuring model interpretability. *CoRR*, *abs/1802.07810*. http://arxiv.org/abs/1802.07810

Qiu, Z., Ren, Y., Li, C., Liu, H., Huang, Y., Yang, Y., Wu, S., Zheng, H., Ji, J., Yu, J., & Zhang, K. (2019). Mind band: A crossmedia ai music composing platform, In *Proceedings of the 27th acm international conference on multimedia*, Nice, France, Association for Computing Machinery. https://doi.org/10.1145/3343031.3350610

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018a). Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018b). Language models are unsupervised multitask learners. https://openai.com/blog/better-language-models/

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, *abs/1910.10683*. https://arxiv.org/abs/1910.10683

Ramanishka, V., Das, A., Zhang, J., & Saenko, K. (2016). Top-down visual saliency guided by captions. *CoRR*, *abs/1612.07360*arXiv 1612.07360. http://arxiv.org/abs/1612.07360

Reed, F. (2020). Promise of ai not so bright. https://www.washingtontimes.com/news/2006/apr/13/20060413-105217-7645r/

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. http://www.jstor.org/stable/4531523

Russell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd). USA, Prentice Hall Press.

Samek, W., Wiegand, T., & Müller, K. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, *abs/1708.08296*. http://arxiv.org/abs/1708.08296

Sheridan, T. (1992). *Telerobotics, automation, and human supervisory control*. Cambridge, MA, USA, MIT Press.

Sheridan, T., Verplank, W., & Brooks, T. (1978). Human and computer control of undersea teleoperators.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. interface agents. *Interactions*, *4*(6), 42–61. https://doi.org/10.1145/267505.267514

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). Grand challenges for hci researchers. *Interactions*, *23*(5), 24–25. https://doi.org/10.1145/2977645

Sousa, V. E. C., Matson, J., & Lopez, K. D. (2017). Questionnaire adapting: Little changes mean a lot. *Western Journal of Nursing Research*, *39*(9), 1289–1300. https://doi.org/10.1177/0193945916678212

Steiner, D., MacDonald, B., Liu, Y., Truszkowski, P., Hipp, J., Gammage, C. L., Thng, F., Peng, L., & Stumpe, M. (2018). Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *American Journal of Surgical Pathology*. https://journals.lww.com/ajsp/Fulltext/2018/12000/Impact_of_Deep_Learning_Assistance_on_the.7.aspx

Stolterman, E. (2008). The nature of design practice and implications for interaction design research. *International Journal of Design*, *2*(1). https://www.jstor.org/stable/4531523

Sutskever, I., Martens, J., & Hinton, G. (2011). Generating text with recurrent neural networks, In *Proceedings of the 28th international conference on international conference on machine learning*, Bellevue, Washington, USA, Omnipress. https://dl.acm.org/doi/10.5555/3104482.3104610

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Wein-

berger, Eds.). In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27*. Curran Associates, Inc. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

Tecuci, G., Boicu, M., & Cox, M. T. (2007). Seven aspects of mixed-initiative reasoning:an introduction to this special issue on mixed-initiative assistants. *AI Magazine*, *28*(2), 11. https://doi.org/10.1609/aimag.v28i2.2035

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*arXiv 1706.03762. http://arxiv.org/abs/1706.03762

Vig, J. (2019). A multiscale visualization of attention in the transformer model. *CoRR*, *abs/1906.05714*arXiv 1906.05714. http://arxiv.org/abs/1906.05714

Wadhwani, P., & Gankar, S. (2020). Outsourced customer care services market size by service. https://www.gminsights.com/industry-analysis/outsourced-customer-care-services-market

Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Commun. ACM*, *62*(6), 70–79. https://doi.org/10.1145/3282486

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*. https://arxiv.org/abs/1910.03771

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., & Zweig, G. (2016). Achieving human parity in conversational speech recognition. *CoRR*, *abs/1610.05256*arXiv 1610.05256. http://arxiv.org/abs/1610.05256

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media, In *Proceedings of the 2017 chi conference on human factors in computing systems*, Denver, Colorado, USA, Association for Computing Machinery. https://doi.org/10.1145/3025453.3025496

Yang, Q., Scuito, A., Zimmerman, J., Forlizzi, J., & Steinfeld, A. (2018). Investigating how experienced ux designers effectively work with machine learning, In *Proceedings of the 2018 designing interactive systems conference*, Hong Kong, China, Association for Computing Machinery. https://doi.org/10.1145/3196709.3196730

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-ai interaction is uniquely difficult to design, In *Proceedings of the 2020 chi conference on human factors in computing systems*, Honolulu, HI, USA, Association for Computing Machinery. https://doi.org/10.1145/3313831.3376301

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research through design as a method for interaction design research in hci, In *Proceedings of the sigchi conference on*

human factors in computing systems, San Jose, California, USA, Association for Computing Machinery. https://doi.org/10.1145/1240624.1240704

APPENDICES

Appendix A

Additional Study Materials

A.1 Information Letter

INFORMATION LETTER

Date: Jan 29th, 2020 – Jan 30th, 2021
Project: Including the human in Explainable Artificial Intelligence (XAI): Exploring visualizations of in- and output to neural networks in an HCI study

Principal Investigator
Dr. Mark Hancock
Management Sciences
University of Waterloo, Canada
mark.hancock@uwaterloo.ca

Student Investigator
Marvin Pafla
Systems Design Engineering
University of Waterloo, Canada
mpafla@uwaterloo.ca

Study Overview
You are invited to participate in a remote research study that aims to explore different AI feedback visualizations in a human-AI collaboration. The goal of this collaboration is to answer questions together with the AI system. The goal of this study is to improve human-AI collaboration and create designs in which AI can explain itself and its actions. This research is conducted in line with a Mitacs Accelerate cooperation with our industry partner Deltcho Valtchanov from Axonify Inc. and with the Master's thesis of Marvin Pafla.

What you will be asked to do
This is an online study. You will be contacted via video chat and are asked to visit a website. You need a computer with a microphone. We will ask you for your IP address to establish an 1-1 online connection with the researcher (we will help you find it). As a participant, you will be asked to answer questions about 28 small text passages in collaboration with an AI system. This requires you to read those passages and then answer questions about those paragraphs. The paragraphs will be short text passages about products and services of different companies. Your role is to act as a service customer representative that answers customer questions about services/products of your company. An AI can be used to generate an answer for you which you can select, edit or delete to come up with your own question. Throughout the study, the AI will give you different feedback that gives you insight how the AI generated the question it generated. To evaluate this feedback, we will ask you to complete a short series of questionnaires which will be computer-administered. After you worked with the AI to generate questions, we will ask you to join us in a semi-structured interview. This interview will take around 20 minutes in which we try to inquire about your experience with the interaction with the AI systems. You will be asked about your personal preferences and experiences. Don't worry if you are unsure about some of the questions. You are not expected to know everything. Hopefully the questions will trigger ideas and thoughts you may want to share. You can agree in the consent form whether the interview is audio or not. In the end of the study, we will ask you some short demographic questions to control for confounders in our study.

Participation and Remuneration
Participants in the study should be comfortable reading short passages. Participation in this study is completely voluntary and will take approximately 60 minutes of your time. You may decline to answer any questions presented during the study if you so wish. Furthermore, you may decide to withdraw from this study by advising the researcher and may do so without any penalty. $15 as an E-Transfer or an Amazon electronic gift card (sent vial email) will be given to you for your time invested in this important step of this research. The amount received is taxable. It is your responsibility to report this amount for income tax purposes.

Benefits of the Study
There are multiple benefits to society. The big question is how we can build AI that is able to explain its own behavior in a way that humans can make sense of it. This study will help us understand how AI explanations are perceived by humans and how AI should be designed so that humans can engage in collaborations with trustworthy AI.

Risks to Participation in the Study
The AI model you will interact with is a probabilistic model. Therefore, there is a very small chance for it to generate racist, sexist or other offensive language. If you feel emotionally harmed throughout the experiment (e.g., feeling demeaned or distressed), please advise the investigator to stop the study at any time. If you feel that the AI has generated an offensive remark, please alert the researcher.

Confidentiality and Security of Data
Your identity will be confidential. Your name will not be included in any thesis or report resulting from this study. Furthermore, because the interest of this study is in the average responses of the entire group of participants, you will not be identified individually in any way in any written reports of this research. All records of data collected during this study will be retained in a locked office at the University of Waterloo for a minimum of 8 years, to which only researchers associated with this study have access. Electronic data and audio recordings will be kept for a minimum of 8 years on a password-protected computer in a locked office at the University of Waterloo, to which only researchers associated with this study have access to. All identifying information will be removed from the records prior to storage. Anonymized data collected from this study might be shared with our industry partner Deltcho Valtchanov from Axonify Inc. to improve the usability of AI technology in real products. Any data shared will not include identifying information (e.g., name, email, etc.). You will be only identified by a number only. Data will be transferred using a secure network or a physical password protected hard drive.

When information is transmitted over the internet privacy cannot be guaranteed. There is always a risk your responses may be intercepted by a third party (e.g., government agencies, hackers). We will ask you for your IP address to establish an 1-1 online connection with the researcher. We won't save your IP address and will discard it once the study is completed. University of Waterloo researchers will not collect or use internet protocol (IP) addresses or other information which could link your participation to your computer or electronic device without first informing you.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #41468). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005. For all other questions regarding this research study, please feel free to contact any member of the research team listed on this form.

A.2 Consent Form

Including the human in Explainable Artificial Intelligence (XAI): Exploring visualizations of in- and output to neural networks in an HCI study

Participant Consent Form

By signing this consent form, you are NOT waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities. I have read the Information Letter regarding the study being conducted by the M.A.Sc. Student Marvin Pafla in Systems Design Engineering at the University of Waterloo. I have had the opportunity to ask questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted.

I am aware that I have the option of allowing my interview to be audio recorded to ensure an accurate acquisition of my responses. I am also aware that excerpts from interview may be included in the thesis and/or publications to come from this research, with the understanding that the quotations will be anonymous.

I was informed that I may withdraw my consent at any time during the study without penalty by advising the researcher.

Only researchers associated with this study will have access to the password-protected study records. Paper files will be locked securely. We will keep your data for a minimum of **8** years. All your data will be transcribed and anonymized after two weeks. You can withdraw your consent to participate and ask that your data be destroyed by contacting one of the researchers within this time. All data will be destroyed according to University of Waterloo policy.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #41468). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca. If you have any questions regarding this study, you can also contact the principal investigator, Dr. Mark Hancock at mark.hancock@uwaterloo.ca or 519-888-4567 ext. 36587.

With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

☐YES ☐NO

I agree to have my interview audio recorded.

☐YES ☐NO

I agree to the use of anonymous quotations in any thesis or publication that comes of this research.

☐YES ☐NO

I agree to have my transcribed and anonymized data be uploaded to a database for future use.

☐YES ☐NO

A.3 Task Instruction

You are a customer service representative for a company. Your job is to answer customer questions online. To do that you will read short paragraphs of text that contain information on your company's product/service and answer the questions relating to this information.

For example, given the text passage

> Apple released the Macintosh Plus on January 10, 1986, for a price of US$2,600. It offered one megabyte of RAM, easily expandable to four megabytes by the use of socketed RAM boards. It also featured a SCSI parallel interface, allowing up to seven peripherals - such as hard drives and scanners - to be attached to the machine. Its floppy drive was increased to an 800 kB capacity. The Mac Plus was an immediate success and remained in production, unchanged, until October 15, 1990.

one could receive the following customer question:

> How much RAM memory did the Macintosh Plus provide when it released in 1986?

Your job is to answer this question after reading the text. A good answer could be:

> The Macintosh Plus provided one megabyte of RAM when it released in 1986.

As seen in this example, the answer should provide all the relevant information asked for in the question.

In this experiment, you will answer 28 questions in total. To get used to the task, you will answer 4 questions without AI and 4 questions with the help of an AI. Don't worry we will show you how the AI works and answer all questions you might have. These are practice questions but try to answer them correctly. The answer should be of high enough quality to be released to customers. After the practice questions, you will go through two stages in the experiment in which you will answer 10 questions in each stage (20 in total). In each stage you will be assisted by AI while having access to different feedback mechanism that allow you to explore the behavior of the AI.

A.4 AI Expertise Questionnaire

AI Expertise

Artificial Intelligence (AI) is a technology that can be found in many products today. For example, you use AI…

- when your email program tries to predict what you will write next
- when you unlock your phone by scanning your face with a camera
- when content is recommended to you by social media

Many researchers assume that AI will become more sophisticated and independent from humans in the future. Under this assumption, please answer the following questions:

AI reputation
- I am worried about AI.
- I believe AI can solve difficult problems.
- I think AI will never reach human intelligence.
- I am afraid AI will cost many jobs.

Anthropomorphism
- I think of AI as computer software.
- I think of AI as human-like.
- AI has a mind of its own.
- AI has intentions.

AI expertise
- I know how AI works.
- I have developed AI in the past.
- I have worked with Tensorflow, Pytorch, Keras or any other Machine Learning library.
- I have worked with Python, R, or any other programming language.

A.5 Interview Guide

Purpose

The purpose of this study is to investigate human-AI collaboration. The goal is to evaluate different AI designs.

Confidentiality

As stated in the consent form and the information later all your data will be fully anonymized. If you agreed in your consent form, your interview will be audio recorded and then transcribed to enable anonymization.

Form

In this interview you will be asked semi-structured interview questions. You don't have to know the answer to all the questions, but we appreciate your thoughts and your comments.

Time

The interview will take around 20 minutes. You can stop the interview at any time.

Feedback & Doubts

After the interview you will be given a feedback letter that contains the researcher's contact. Feel free to contact us anytime if you have remaining questions or doubts.

General Structure

1. Topic
    a. Question
        i. Hint

Interview Guide

1. Introduction
    a. How did it go?
    b. Good/bad/disturbing things?
    c. Clear/unclear things?
2. Question Generation
    a. How was it do create questions by yourself?
        i. Difficulty
        ii. Complexity
    b. What was your strategy to design questions?
        i. What is a good question?
3. AI assistance
    a. How competent was the AI?
        i. Perceived technical competence
    b. How reliable was the AI?
        i. Perceived reliability
    c. How did you understand the AI?
        i. Perceived understandability

    d. How did you believe in the AI?
        i. Faith
    e. How close did you feel to the AI?
        i. Personal attachment
4. Attention visualization
    a. Can you explain the attention visualization to me?
    b. How did the visualization change the interaction?
        i. Reliability
        ii. Competence
        iii. Understandability
        iv. Faith
        v. Attachment
    c. How did the visualization change your trust towards the AI?
    d. Can you contrast having the attention visualization with not having it?
5. Selection of alternatives
    a. How was it to select alternatives?
    b. How did it change the interaction?
        i. Control
        ii. Feedback (trial and error)
        iii. Reliability
        iv. Competence
        v. Understandability
        vi. Faith
        vii. Attachment
    c. Did you feel like the AI explained its behavior?
    d. Can you contrast selecting alternatives with not doing being able to select them?
6. Together
    a. How was the combination of the feedback mechanism?
        i. Confusion
        ii. Added value
    b. Can you contrast what was more helpful: selection of alternatives or attention visualization?
7. Final
    a. How did your attitude towards AI changed throughout this experiment?
        i. This system vs general
    b. Do you have any final comments/concerns/doubts you want to express?

A.6 Feedback Letter

FEEDBACK LETTER

Date:
Project: Including the human in Explainable Artificial Intelligence (XAI): Exploring visualizations of in- and output to neural networks in an HCI study

Principal Investigator
Dr. Mark Hancock
Management Sciences
University of Waterloo, Canada
mark.hancock@uwaterloo.ca

Student Investigator
Marvin Pafla
Systems Design Engineering
University of Waterloo, Canada
mpafla@uwaterloo.ca

We appreciate your participation in our study and thank you for spending the time to help us with our research!

Study Overview
This study explores the collaboration of humans and Artificial Intelligence (AI). The purpose is to find out how AI can explain itself to become more trustworthy to humans. The goal of this study is to improve human-AI collaboration and create designs in which AI can explain itself and its actions. If you have any questions or concerns, please feel free to contact the student investigator Marvin Pafla at mpafla@uwaterloo.ca.

Confidentiality and Security of Data
Your identity will be confidential. Your name will not be included in any thesis or report resulting from this study. Furthermore, because the interest of this study is in the average responses of the entire group of participants, you will not be identified individually in any way in any written reports of this research. Paper records of data collected during this study will be retained in a locked office at the University of Waterloo for a minimum of 8 years, to which only researchers associated with this study have access. Electronic data and audio-video recordings will be kept for a minimum of 8 years on a password-protected computer in a locked office at the University of Waterloo, to which only researchers associated with this study have access to. All identifying information will be removed from the records prior to storage.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE #41468). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005. For all other questions regarding this research study, please feel free to contact any member of the research team listed on this form.

Psychological or Emotional Harm

If the study caused you psychological or emotional harm, please inform the principal investigator. Counseling and help is available at https://uwaterloo.ca/campus-wellness/urgent-help-and-emergency-contacts.

A.7 Code

Depending on the Transformer model and its vocabulary the output of the model might vary. Our model generated tokens which can represent single words or added together can form a single word. To not confuse participants we averaged the relevance values for words that consisted out of multiple tokens. The following function achieves this in Python:

```
def get_averaged_attentions_per_word(self, attentions,
    tokens_decoded):

        #average layers
        attentions_averaged = self.average_attentions(attentions)

        #the current word that is built out of tokens
        word = ""
        last_token_is_unique_word = False
        words = []

        #index in attentions_averaged that describes the start of the
            current word
        index_word_start = 0

        #number of colums and rows that have been averaged out
        number_of_averages = 0


        #save last token for later as it does not have attention
           weights
        last_token = tokens_decoded[-1]
        tokens_decoded = tokens_decoded[:-1]

        for index, token in enumerate(tokens_decoded):

            #adjusted index describes the end of word index in the
               averaged_attentions after adjustment of averaged
               columns/rows
            adjusted_index = index - number_of_averages
```

```
#basically number of tokens for current word
index_difference = adjusted_index - index_word_start

#append last token to word or create new word for last
    token
if (len(tokens_decoded) - index) < 2:
    if any(c in token for c in " ,.?!<"):
        last_token_is_unique_word = True
    else:
        word = word + token
        adjusted_index = adjusted_index + 2


# start new word for tokens with special symbols
# but not the first token or for the last token
if any(c in token for c in " ,.?!<") and word is not ""
    or (len(tokens_decoded) - index) < 2:

    if (index_difference) > 1 or len(tokens_decoded) -
        index < 2:
    #if (index_difference) > 1:

        #average matrix column-wise
        attentions_averaged_column = np.mean(
            attentions_averaged[index_word_start:
            adjusted_index], axis=0)[np.newaxis]
        attentions_averaged = np.concatenate([
            attentions_averaged[:index_word_start],
            attentions_averaged_column,
            attentions_averaged[adjusted_index:]])
        attentions_averaged_row = np.expand_dims(np.mean(
            attentions_averaged[:,index_word_start:
            adjusted_index], axis=1), axis=1)

        #average matrix row-wise
        attentions_averaged = np.concatenate(
            [attentions_averaged[:, :index_word_start],
                attentions_averaged_row,
```

```
                        attentions_averaged[:, adjusted_index:]],
                            axis=1)

                    #-1 because all tokens were averaged into 1 word
                    number_of_averages = number_of_averages +
                        index_difference - 1

                words.append(word)

                # start of a new word
                index_word_start = index_word_start + 1
                word = token

                if last_token_is_unique_word:
                    words.append(word)

            else:
                #no new word, just move forward
                word = word + token

        words = words + [last_token]

        assert len(words) == attentions_averaged.shape[0] + 1 ==
            attentions_averaged.shape[1] + 1

        return words, attentions_averaged
```

This project was created in cooperation with Axonify Inc. which provided the code repository for this project. While the interface was developed by myself, it relies on IP-protected parts of the Axonify repository. Therefore, the application can not be run by itself which is why I will not publicly share the interface code.

Nomenclature

HCI expert A person with at least one year of experience in HCI as a post-graduate researcher from my research institute. 5, 21, 24–26, 29, 32, 52, 62

Transformer Current state-of-the-art language model (i.e., neural net) that uses hundreds of millions of parameters to predict the next best word given any input. To do that it scores the relevance of each input word with its intrinsic attention mechanism. These scores can be used to indicate what input the model 'paid attention to'. 4, 5, 8, 13, 21, 22, 28, 29, 31, 34, 57, 61