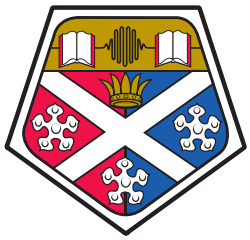
Exploring the Impact of Conversational Strategies on User Search Experience in Goal-Oriented Tasks in a Voice-Only Domain



Mateusz Dubiel
Computer and Information Sciences
University of Strathclyde

A thesis submitted for the degree of
Doctor of Philosophy

Glasgow 2020

Declaration of Authenticity and Author's Rights

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: *Mateusz Dubiel*

Date: 09/11/2020

Acknowledgements

Firstly, I would like to thank my supervisors Dr Martin Halvey and Dr Leif Azzopardi. Martin, thank you for your continued encouragement and faith in my abilities, your mentorship made the successful completion of this project possible and helped me to become a better researcher in the process. Leif, thank you for your keen interest in my work, for challenging my ideas and reminding me that each research problem can be approached from many different angles.

A big thanks goes to my friend and fellow-researcher Dr Damien Anderson for being an excellent wizard and offering feedback that helped to improve this work.

I would like to thank all the members of DASSI and SiSRG, and their respective founders, Dr Mark Dunlop and Prof. Ian Ruthven. Both groups were great forums for discussions that fostered the development of this PhD.

To Prof. Simon King, thank you for being a great teacher and a source of inspiration.

Additionally, I am grateful to Prof. Giuseppe Riccardi and Prof. Minoru Nakayama who hosted me during my research visits. Your guidance during my placements contributed to my personal growth and helped me to broaden my research horizons.

I am much indebted to my friends and fellow denizens of the Livingstone Tower. These include. Amine for his ever-optimistic outlook on life. David for his enthusiasm and help with recruitment of participants. Diane for scenic walks in Ardrossan. Dominika for her support and great taste in movies. Linda for our chatty runs and informative stats tutorials. Mohammed for many coffee breaks. Niall for helping me to improve my writing. Olivia for her inspiring stories and reading my manuscripts. Revathy for her composure and always being ready to help. Vasilis for his upbeat personality and many discussions about conversational agents. Zuzanna for being able to see the bright side and cheering me up. Alessandra and Pilar for being fantastic co-authors. Sylvain for always having time for me, constantly helping me to improve my work, for many great hiking trips and research collaborations.

My sister and my parents have always been there for me. Lena, thank you for believing in me. Dad, thank you for always encouraging me to learn new things and for introducing me to the Amiga 500, our first 'speaking computer'. Mom, thank you for teaching me how to appreciate the beauty of this world and for your emotional support. This one is for you guys!

‘Here I will say only that world injected its patterns into human language at very inception of that language, mathematics sleeps in every utterance and can only be discovered, never invented.’

Stanisław Lem, ‘His Master’s Voice’

–To my family

Contents

List of Figures

List of Tables

List of Abbreviations

ASR	**A**utomatic **S**peech **R**ecognition
CA	**C**onversational **A**gent
CSA	**C**onversational **S**earch **A**gent
DA	**D**ialogue **A**ct
DM	**D**ialogue **M**anager
DS	**D**ialogue **S**ystem
GOT	**G**oal-**O**riented **T**ask
HCI	**H**uman-**C**omputer **I**nteraction
IR	**I**nformation **R**etrieval
IIR	**I**nteractive **I**nformation **R**etrieval
NLG	**N**atural **L**anguage **G**eneration
SCS	**S**poken **C**onversational **S**earch
SLT	**S**poken **L**anguage **U**nderstanding
SSES	**S**earch **S**pace **E**xploration **S**atisfaction
SUS	**S**ystem **U**sability **S**cale
TLX	**T**ask **L**oad Inde**x**
TTS	**T**ext-**T**o-**S**peech
VSS	**V**oice-**S**earch **S**ystem
WOZ	**W**izard **o**f **O**z

Glossary

Cognitive Workload is a level of measurable mental effort of an individual in response to a cognitive task. In the current work cognitive workload is measured based on the interaction of an information seeker with different conversational agents.

Conversational State Tracking is an act of determining at each turn of a conversation what action the user wants to accomplish at this point of conversation. Examples of user actions include requesting information or confirming a time for a booking.

Decision Theory is an approach to judgement and decision making. It focuses on the subjective expected utility of a person's actions. In this thesis it is considered in terms of benefit that different conversational agents can bring to information seeker.

Elicitation is a question-asking activity in conversation. Its explicit function is to request information.

Goal-Oriented Task is a task in which a participant is expected to accomplish a certain goal. For example, booking a flight is considered as a goal-oriented task.

Interlocutor is someone who takes part in a conversation.

Mixed Initiative (in a conversation) is an act of switching initiative between participants where they decide what action to take next.

Revealment is an activity of presenting (revealing) information.

Search Space is the space of all available options that a participant can choose from during a simulated search scenario.

Short Term Memory is an individual's capacity for holding a limited amount of information in mind in a readily available state over a short period of time. The estimated span of short term memory is considered to be around 18 seconds [129].

Slot-Filling System is a system that extracts values of certain types from user utterances that are required for attributes of a query (slots). An example of a slot-filing action for a restaurant booking is finding the date and time or the reservation and number of people who will be dining.

Statefulness (in the context of conversation) is the ability of an agent to maintain conversational state and attend to the information provided over the course of the conversation.

Usability the capability of the system to be used easily and effectively by a user in order to fulfill the specified certain task in a given scenario. The term was coined by Shackel (cf. [142]).

User Search Experience is the measure of interaction with a conversational agent that is based on participants' perception of the agent and its performance in a goal-oriented task. In the current research, user search experience is evaluated via subjective measures (cognitive workload and satisfaction) and objective measures (interaction metrics and performance).

Voice-only (in the context of search activity) is a type of interaction that is conducted exclusively though the medium of speech without any graphical interface.

Wizard of Oz (in the context of human-computer interaction) is an interactive system simulation technique where a human acts as a system while the user thinks they are interacting with a software agent. The term is attributed to Dahlbäck et al. [38].

Working Memory is a model proposed by Baddeley and Hitch [13] that focuses on how memory is applied towards achieving goals and problem solving. The model contains three elements: (1) central executive, (2) phonological loop which stores acoustic information and (3) visuospatial sketchpad which holds visual information. The role of the central executive is to manipulate the information stored in the phonological loop and visuospatial sketchpad.

Abstract

Conversational search has recently become established as a new information search paradigm. Increasing numbers of people are using smart speakers and virtual personal assistants to access information online. Given the substantial increase in the usage of devices that feature conversational agents, it is timely to explore the implications that they have on user search behaviour.

This thesis explores how conversational search agents impact on user search experience and outcomes for goal-oriented tasks. Our enquiry is inspired by the premise that a fully conversational agent equipped with memory could lead to a better user search experience than an agent that is based on a slot-filling paradigm which represents the current state of the art. This dissertation presents three Wizard of Oz, lab-based studies that explore a series of conversational agents that differ in their ability to preserve conversational state and employ a different degree of conversational initiative. The goal of this research is to get a better understanding of how different ways of presenting information impact on user search experience in terms of cognitive load, satisfaction with the agent, time required to complete the task and overall performance.

In the first study, we address the problem of an agent's statefulness, which is the ability to preserve and account for the information presented by the user without the need to repeat it. Specifically, through a series of interactive search tasks, we evaluate how agent's memory improves user search experience. In the second study, we focus on mixed-initiative in conversation, by evaluating how different ways of eliciting information and presenting it back to the user (revealment) impact upon user search experience. In the third study, we further explore the role of mixed-initiative by evaluating agent's 'proactivity' (i.e. the search involvement that goes beyond the original scope of the query) on user search experience. Finally, we compare the insights from all of the three studies and assess how differences in search task design and the mode of delivering the studies (lab-based vs. deployed online) impact upon experimental outcomes.

Overall, the contribution of the research presented in the current thesis is threefold. Firstly, it provides a validation of theoretical frameworks of conversational search by empirically investigating the impact of agent's memory and mixed initiative on user

search experience. Secondly, it provides insights into the efficacy of alternative ways of eliciting information and presenting search results during conversational search. Thirdly, it contributes towards a more robust evaluation of conversational agents by developing a spectrum of conversational strategies and reflecting on the impact of the design of goal-oriented search tasks.

PART I:

Thesis Outline, Background and Methodology

Chapter 1

Introduction

Conversational Search has recently emerged as a new information retrieval paradigm [154]. Although previous research considered theoretical aspects of accessing information via voice [126, 156, 162], little work has been dedicated to a practical evaluation of how to present information back to the user and how to evaluate the impact of different information presentation techniques on user search experience. This thesis addresses this gap by investigating how different conversational agents differ in their potential to maximise user performance while minimising required cognitive effort. In particular, it aims to answer the question: '**How much would a truly conversational agent with the ability to preserve state (memory) and conversational initiative improve user search experience compared to a stateless, passive voice search system?'** The answer to this question will be provided based on the results of three Wizard of Oz studies (presented in Chapters 4, 5 and 6) where different conversational agents were evaluated interactively via simulated search scenarios. We will further compare different experimental setups and discuss the impact of task complexity (see Chapter 7) on user search experience.

1.1 Motivation

There is a general assumption that natural speech conversational search agents are more usable and user friendly than slot-filling based agents (current state-of-the-art) that provide preprogrammed responses (cf. [106]). Regardless of its prevalence, however, this assumption is rarely investigated empirically. Recent theories [126, 156] outline the qualities that a search agent should have in order to be considered 'conversational', however, due to technological limitations (e.g. reinforcement learning limits agent's capabilities to user simulations that it was trained on [32]) a full and robust implementation of such agents is still infeasible. Consequently, due to the lack of empirical evaluations, little is known regarding the potential impact of conversational agents on user search experience regarding: cognitive workload, satisfaction and task performance. In this PhD, I seek to narrow this gap in knowledge by conducting a series of interactive user experiments where participants interact with conversational search

agents, simulated by a human, to accomplish a series of goal-oriented search tasks. The three studies (presented in Chapters 4, 5 and 6) provide both quantitative and qualitative insights into the impact of different conversational agents on user search experience.

As recently noted by Trippas: [154, p.16] '[...]the IR community lacks a broader insight into how users will engage with these highly interactive search systems and which components may be involved. These insights in turn are required in order to make a highly interactive and usable system.' This thesis seeks to address this gap via three interactive studies that evaluate the impact of a conversational agent's memory, and aspects of mixed initiative (i.e. information presentation strategies and pro-activness) on user search experience.

1.2 Context

We consider conversational agents in the context of goal-oriented search scenarios, i.e. scenarios where the user has to accomplish a specific goal such as booking a flight. In the current work, the term 'user search experience' comprises cognitive impact, the user's satisfaction with a search system and task performance (a more detailed explanation of search experience is provided in Section 3.4). The purpose of interaction with the system is to achieve a certain goal (e.g. booking a service or finding a product etc.) rather than an information seeking task. In all studies presented in the current thesis, the tasks are conducted exclusively through the medium of voice, with no visual feedback being provided.

1.3 Research Questions

In order to provide the answer to the overarching question: '**How much would a truly conversational agent with ability to preserve state (memory) and conversational initiative improve user search experience compared to a stateless, passive voice search system?**', the current thesis explores the following sub-questions:

1. How does a stateful conversational agent[1] (with a memory component) differ from a stateless conversational agent[2] with regards to: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?

2. How do different information elicitation (passive vs. active) and revealment (summary vs. listing) techniques impact on user search search experience in a goal-oriented task with regards to: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?

[1] An agent that keeps track of the state of interaction, by setting values in a designated storage field.
[2] An agent with no record of previous interactions, it handles each request individually, based only on information that comes with it.

3. How do agents that proactively elicit search criteria (proactive elicitation) and proactively recommend search results that are outside the original scope of the user's query (proactive recommendation) vary in terms of their impact on: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?

1.4 Contributions

The principal contributions of the current PhD are based on Wizard of Oz studies. Through three interactive user experiments, we demonstrate how conversational agents with different levels of conversational involvement (passive, active and proactive) impact on user search experience in goal-oriented tasks. The contributions of this thesis are summarised below.
This thesis:

1. Investigates the impact of agent's memory on user search experience by comparing a conversational agent that maintains conversational state (stateful CA) with an agent that does not (stateless CA) - based on the answer to research question 1.

2. Investigates the impact of the degree of initiative and conversational involvement of the agent on user behaviour - based on the answer to research question 2 and 3.

3. Offers qualitative insights regarding user expectations of conversational agents - obtained via semi-structured interviews with participants.

4. Proposes a set of goal-oriented search scenarios (provided in Appendix E) to evaluate the impact of conversational agent behaviour on user search experience.

5. Proposes a spectrum of conversational search strategies (discussed in Section 3.3.1). The strategies aim to provide a set of detailed rules on how to elicit and reveal information in a goal-oriented task - based on research question 2 and 3.

The current work has several implications. Firstly, it illustrates, in a practical way, how a stateful conversational agent could impact user search experience compared to an agent based on slot-filling architecture. Secondly, conceptually, it proposes a standardised method of presenting information that addresses the problem of the transitory nature of speech and highlights trade-offs between different information presentation strategies. The aim is to ensure that the user gets the best possible understanding of the search space without getting overwhelmed with information. Finally, methodological implications include: the development of goal-oriented search tasks and the design of a balanced space of search options that can be used in future experiments, as well as the Pareto-frontier evaluation metric (further discussed in Section 3).

In summary, the findings of the current thesis should enable conversational designers to develop more usable conversational agents and provide them with tools to make their evaluation more robust and standardised.

1.5 Thesis Outline

The current thesis is organised into the following parts and corresponding chapters.

PART I: Thesis Outline, Background and Methodology

Chapter 1 - Introduction: provides the outline of the thesis, explains challenges in evaluation of conversational agents, and discusses contributions of the thesis.

Chapter 2 - Background: contextualises research that relates to conversational search agents, including insights from the fields of human-computer interaction, signal processing and information retrieval. Chapter 2:

(1) provides a definition of a conversational search paradigm as an instance of an interactive information retrieval and contrasts it with the traditional, static information retrieval,

(2) provides challenges for implementing conversational agents that arise from the complexity of human language,

(3) introduces Radlinski and Craswell's Theoretical Framework of Conversational Search [126] and its expansion by Azzopardi, Dubiel, Halvey and Dalton [11], and

(4) summarises recent interactive user studies that focus on the evaluation of conversational agents.

Chapter 3 - Methodology: explains Wizard of Oz methodology, provides details of the general experimental setup for all of the studies, characteristics of the search tasks and an explanation of both qualitative and quantitative metrics used for evaluating conversational agents.

PART II: Three Wizard of Oz studies, investigating aspects of agent's memory (Study 1) and mixed-initiative (focusing on information presentation strategies (Study 2) and agent's pro-activness (Study 3)).

Chapter 4 - Investigating Memory Component of Conversational Agents: The first Wizard of Oz study provides an empirical comparison between a slot-based, stateless conversational agent (that represents the state-of-the art) and a stateful conversational agent, and evaluates their impact on user search experience. User search experience is evaluated both quantitatively (user performance metrics) and qualitatively (insights from semi-structured interviews). – The focus of this experiment is on the impact of the 'Memory' element of the Theoretical Framework of Conversational Search [126].

Chapter 5 - Eliciting and Presenting information via Audio-only channel: The second Wizard of Oz study where the focus is on eliciting and presenting information by a conversational agent. We propose a spectrum of conversational search strategies and empirically evaluate them in a series of interactive search scenarios. – The focus of the experiment is on the impact of the 'Mixed Initiative' element of the Theoretical Framework of Conversational Search [126] .

Chapter 6 - Proactive Conversational Agents The third Wizard of Oz study focuses on the impact of proactive conversational agents. The study focuses on how proactive elicitation and recommendation of search results by a conversational agent impacts on user search experience – The study further explores the 'Mixed-initiative' in search, where users delegate more of their search agency to the agent.

Chapter 7 – Comparison of 3 Wizard of Oz Studies: - this chapter compares the impact of different experimental setups and search task design on user search experience. The chapter also makes recommendations for design and development of conversational search tasks agents based on the results of three interactive WOZ studies.

Venn diagram 1.1 illustrates the three aspects of conversational agents and their impact on user search experience that are covered in Chapters 4-7.

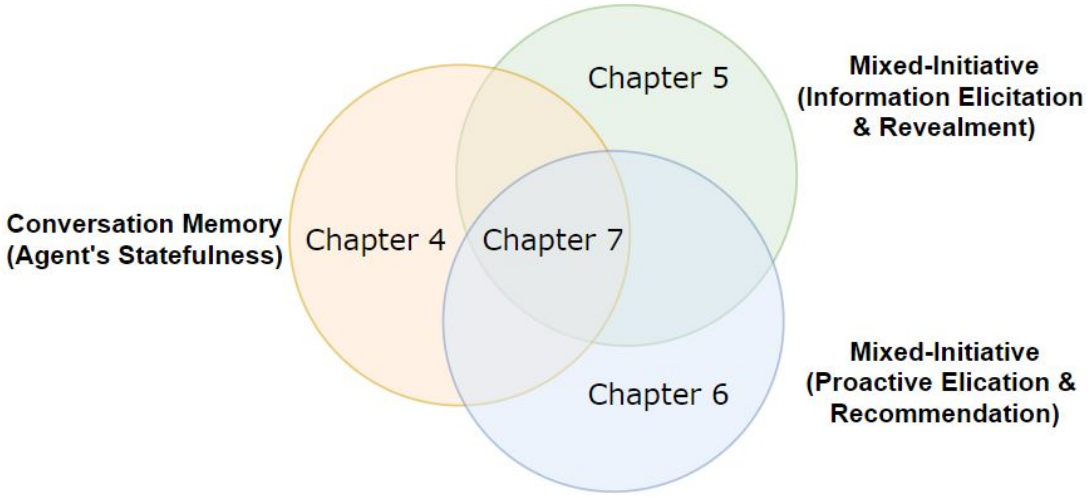


Figure 1.1: Contribution of each chapter to addressing the three main components of a conversational agent.

PART III: Conclusions

Chapter 8 – Conclusions - concludes the research presented in this PhD thesis, acknowledges its limitations and makes recommendations for future work.

Finally, the thesis contains five appendices with complementary information about participant information sheets and consent forms (Appendix A), questionnaires for three studies featured in the thesis (Appendix B), interactive dialogue excerpts (Appendix C) additional material (Appendix D) and search scenarios used in all of the studies (Appendix E).

Publications

Publications Used in this Thesis

Research that resulted from this PhD has been published at the following peer-reviewed venues, using only the parts of these papers that are directly attributable to the author. For each paper, we refer to the corresponding chapter where the content of the paper is included:

1. Dubiel, M., Halvey, M., Azzopardi, L., and Daronnat, S. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *The Second International Workshop on Conversational Approaches to Information Retrieval* (2018) [52] - The content of this paper is discussed in Chapter 4.

2. Dubiel, M. Towards human-like conversational search systems. *CHIIR '18: 2018 Conference on Human Information Interaction & Retrieval* (2018). [46] - The content of this paper is discussed in Chapter 4.

3. Dubiel, M., Halvey, M., Azzopardi, L., Anderson, D., and Daronnat, S. Conversational strategies: Impact on search performance in a goal-oriented task. *In The Third International Workshop on Conversational Approaches to Information Retrieval* (2020) [50] - The content of this paper is discussed in Chapter 5.

4. Dubiel, M., Halvey, M., Azzopardi, L., and Daronnat. Interactive evaluation of conversational agents: reflections on the impact of search task design. *ICTIR '20: ACM SIGIR International Conference on the Theory of Information Retrieval* [53] - The content of this paper is discussed in Chapter 7.

5. Azzopardi, L., Dubiel, M., Halvey, M., and Dalton, J. Conceptualizing agent-human interactions during the conversational search process. *In The Second International Workshop on Conversational Approaches to Information Retrieval* (2018) [11] - The content of this paper is discussed in Chapter 3.

Additional Publications

Over the course of my PhD, I have also published and co-authored the following papers. However, they are not included in the current thesis as they are not directly connected to the topic of this PhD.:

1. Dubiel, M., Halvey, M., and Azzopardi, L. A survey investigating usage of virtual personal assistants. *arXiv preprint arXiv: 1807.04606 (2018)* [49].

2. Dubiel M., Halvey, M., Azzopardi, L., Aylett, M., Wester, M., and Braude, D. A. Improving conversational dynamics with reactive speech synthesis. *In Voice-based Conversational UX Studies and Design Workshop* (2018) [51].

3. Dubiel, M., Oplustil, P., Halvey, M., and King, S. Persuasive synthetic speech: Voice perception and user behaviour *In Proceedings of the 2nd International Conference on Conversational User Interfaces* (2020) [55]. – The paper received a honourable mention award.

4. Dubiel, M., Cervone, A., and Riccardi, G. Inquisitive mind: a conversational news companion. *In Proceedings of the 1st International Conference on Conversational User Interfaces* (2019) [48].

5. Dubiel, M., Nakayama, M., and Wang, X. Combining Oculo-motor Indices to Measure Cognitive Load of Synthetic Speech in Noisy Listening Conditions. *To Appear In Proceedings of ETRA '21: 2021 Symposium on Eye Tracking Research and Applications* (2021).

6. Dubiel, M., Nakayama, M., and Wang, X. Evaluating synthetic speech workload with oculo-motor indices: preliminary observations for Japanese speech. *In Proceedings 15th International Joint Conference on Biomedical Engineering Systems and Technologies.* (2021) [54].

7. Dubiel, M. Becoming digital: Toward a post-internet society. *Journal of Enabling Technologies (2018)* [45]. - Book Review.

8. Daronnat, S, Azzopardi, L., Halvey, M. and Dubiel, M. Impact of agent reliability and predictability on trust in real time human-agent collaboration. In *Proceedings of the 8th International Conference on Human-Agent Interaction (HAI)* (2020) [39].

9. Tortoreto, G., Stepanov, E. A., Cervone, A., Dubiel, M., and Riccardi, G.Affective behaviour analysis of on-line user interactions: Are on-line support groups more therapeutic than twitter? *ACL 2019* (2019), 79 [153].

Chapter 2

Background

The development and evaluation of conversational agents spans a number of areas, including Information Retrieval (IR), Human-Computer Interaction (HCI), computational linguistics and cognitive science. This chapter introduces the key terminology and background information required to understand the mechanics of goal-oriented, conversational agents and methods of their evaluation. First, Section 2.1 provides a brief historical context for the concept of a conversational agent. Next, Section 2.2 presents 'conversational search' as an interactive search paradigm and explains how it differs from traditional information retrieval. In Section 2.3, we discuss search models with a special focus on dialogue acts theory and its relevance to conversational search. Then, in Section 2.4 we look at the challenges that the complexity of human language poses to conversational search systems - and illustrate how spoken language understanding works based on a goal oriented task. Next, in Section 2.5, we present several conversational search frameworks. Section 2.6 presents relevant work on evaluation of conversational agents. In Section 2.7, we explain how the current thesis combines different evaluation metrics used in previous work to answer the research question. Finally, Section 2.8 provides a summary of the current chapter.

2.1 Inception of Conversational Systems

The idea of seamlessly communicating with a computer system via voice has a long history. Conversational systems have been featured in sci-fi movies such as Space Odyssey 2001, Star Trek and Interstellar, just to name a few examples. Futuristic visions presented such systems as intelligent companions capable of supporting humans with a variety of navigation and planning tasks. In 1987, Apple proposed Knowledge Navigator, a device that allowed natural language interaction for carrying out searching, planning and communication actions [140]. The concept was received with great enthusiasm [140]. However, implementation proved impossible due to technological limitations of the time, including (among the other issues) low accuracy of Automatic Speech Recognition (ASR) and limited computing power. 24 years later, the project returned as intelligent assistant, Siri, but the range of tasks that it was capable of was

still nowhere near the vision laid out in the Knowledge Navigator project. Siri had problems dealing with complexity of language, i.e.: maintaining coherent conversation and co-reference resolution [36, 98].

Continued technological development allowed conversational agents to improve. In May 2018, during the Google I/O developer conference [1], the company showcased Duplex [92], its automated conversational system designed to make restaurant bookings via phone. Duplex stunned the audience by its ability to hold a natural language conversation with a human operator in a realistic and smooth manner. Not only was the system able to perfectly understand the interlocutor (a high accuracy ASR) but it also responded in a perfectly natural sounding way (intelligible text-to-speech (TTS)). Although, it was later revelled that Duplex was not completely autonomous, with human operators still monitoring the phone line and ready to intervene when the conversation did not go the right way [26]. Regardless of its limitations, Duplex demonstrated that for clearly defined tasks that do not require excessive exploration and negotiation, the conversational agent is commercially feasible.

2.1.1 Conversational Search Systems of the Past

Although the wide-spread availability of commercial conversational agents is a relatively recent phenomenon it was underpinned by decades of intensive research on dialogue systems [32]. Early examples of conversational search systems include THOMAS [112], a reference retrieval program that supported users in selecting documents without the need to explicitly formulate the queries. THOMAS supported the search process by asking user questions and presenting suggestions. At the end of the question-answering process, the program presented the user with relevant results. In the 1980s Croft and Thompson proposed Intelligent Intermediary for Information Retrieval (IR^3) [37] that was modelled to reflect an expert intermediary. Contrary to a traditional information retrieval system in which the user was limited to one retrieval strategy (a query), the IR^3 helped the user to expand their domain knowledge, provided explanation, enabled browsing, retrieval and evaluation. The system could also confirm or request more information from the user.

2.2 Conversational Search Paradigm

As stated by Anand et al.: 'The conversational search paradigm promises to satisfy information need using human-like dialogues, this kind of 'information-providing' dialogues will increasingly happen en passant and spontaneously.' [3, p.34] In this section, we will provide a definition and characteristics of a conversational search paradigm (Section 2.2.1), highlight how it differs from a traditional information retrieval paradigm (Section 2.2.2) and explain how conversational

[1] https://events.google.com/io/ (last accessed on the 30th October 2020)

search systems can be evaluated using an interactive information retrieval approach (Section 2.2.3).

2.2.1 Defining Conversational Search

In this thesis, we use the definition of a conversational agent by Radlinski and Craswell who define it as 'a system for retrieving information that permits a mixed-initiative back and forth between a user and agent, where the agent's actions are chosen in response to a model of current user needs within the current conversation, using both short- and long-term knowledge of the user.' [126, p.4]. The authors further highlight that such a system needs to take an active part in a conversation with the user (mixed-initiative) and keep track of information provided (memory) to enable the efficient execution of various information retrieval tasks.

From the perspective of IR, the purpose of conversational search is to accomplish a specific goal such as retrieving information, booking a flight, checking news via means of conversation, etc. The notion of an information exchange between interlocutors (conversational parties) is crucial. In conversational search the agent and the user alternate turns between each other to accomplish a specific task. During conversational search there are two main actions that a user or an agent can take, namely:

1. Elicitation - which concerns how to clarify the information need by asking questions and checking current understanding.

2. Revealment - which concerns how the knowledge or information should be presented.

Conversational search departs from a traditional information retrieval paradigm, where user queries are submitted to a search engine which provides answers and allows for query refinement. Unlike a traditional interaction paradigm, conversational search allows users to formulate their queries via voice. Due to its dynamic and interactive nature, as well as the need to deal with ambiguities of language, conversational search presents a number of challenges (cf. [32]).

In the following Section (2.2.2), we will provide a more detailed explanation of the differences between a traditional information retrieval paradigm and conversational search, and then discuss interactive information studies as relevant tools to evaluate conversational systems (Section 2.2.3). In the current thesis, the terms Conversational Agent (CA) and Conversational System will be used interchangeably.

2.2.2 Conversational Search vs. Traditional Information Retrieval

Conversational search (CS) differs from the traditional browser-based search paradigm (illustrated in Figure 2.1) due to its interactive nature. In CS, the system can be actively involved in the search process and proactively resolve a user's information needs via conversation. Contrary to a query-response paradigm where results are displayed as a list, CS enables users to interactively clarify their information needs and to critique the results provided.

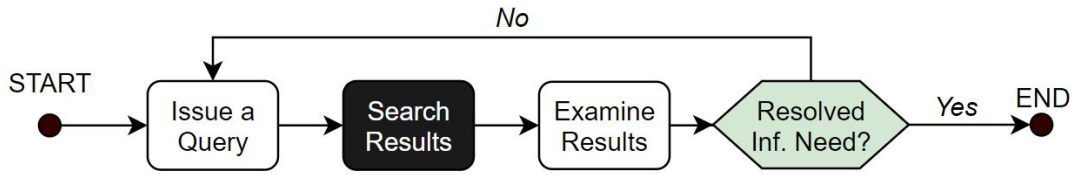


Figure 2.1: A simplified diagram of the information access model. User actions are represented by white squares, the black square symbolises system action and the green hexagon denotes a decision point.

Conversational search assumes a multi-turn, verbal interaction between the user and the system. Both the user and the system can take the initiative, while the system keeps track of what happened earlier in conversation (state tracking). A user's information need can be expressed, formalised and elicited through the means of natural language conversational interactions. The system proactively supports the user's search process and provides responses that are relevant to the context and cognitively processable. Contrary to CS, traditional information retrieval paradigm is more rigid in its form.

2.2.2.1 Traditional Information Retrieval

Information Retrieval (IR) concerns how information (in the form of documents) is stored, organised and accessed (cf. [14, chapter 1]). The challenge for the user is to present the information need in the form of a query that can be understood by the system. In its basic form, the information need is typically presented as a set of key words which summarises the user's information need. The goal of the IR system is to relay the information back to the user in a timely and concise manner to satisfy their information need [14].

Ad-hoc retrieval is a standard IR task. In such a task, the system aims to provide documents from within the collection that are relevant to an arbitrary user information need that was provided to the system in the form of a one-off, user-initiated query. The user is required to translate their information need into a query language supported by the system. In a conventional IR system, the user provides a set of words that convey the meaning (semantics) of the information need.

The spectrum of IR studies is presented in Figure 2.2. At one end of the spectrum, IR research concerns the investigation of computer algorithms and system architectures (system focus), at the other end, there is a human-centered approach that is focused on trying to understand how people use information search systems and how they interpret the information provided by the system.

Since the topic of the current PhD concerns conversational search, which can be considered as an instance of interactive information retrieval where a user interacts with a search system via voice, the discussion of traditional information retrieval measures will not be considered in

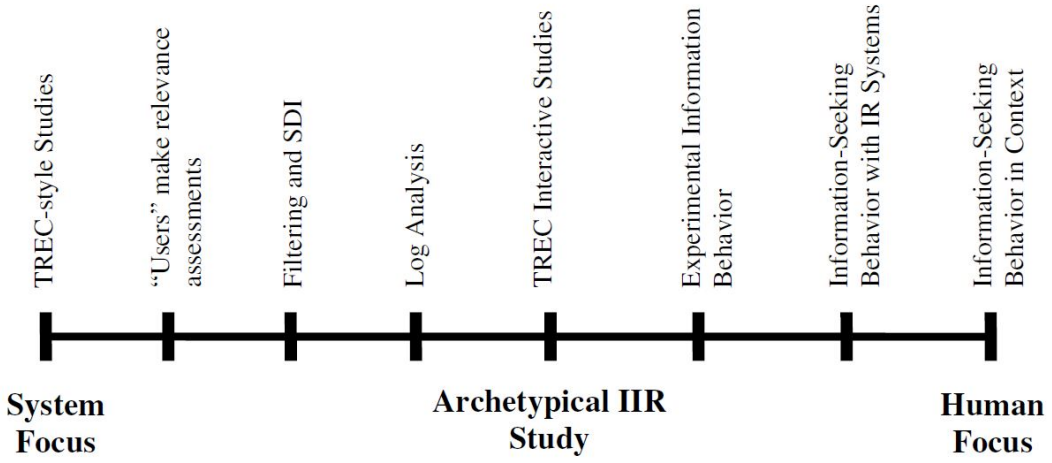


Figure 2.2: Spectrum of interactive information retrieval studies figure adapted from [81].

the current thesis. Instead, in the following subsection, we will provide a detailed discussion of interactive information retrieval studies that can serve as a means for testing conversational systems.

2.2.3 Interactive Information Retrieval

While traditional IR studies focus on the performance of retrieval systems, in interactive information retrieval (IIR) the focus is on user behaviour and experiences: including physical, cognitive and affective features of interaction [81]. IIR encompasses many disciplines, including psychology, information and library science and human-computer interaction (HCI). The need for human evaluation of IR systems was recognised early by pioneers of the field such as Salton [133] who emphasised the importance of user perceptions and attitudes, or Cleverdon who [30] identified user effort as a viable measure of IR system performance that can complement precision and recall.

The need for an interdisciplinary approach to IIR is highlighted by Kelly, who argues: 'The inclusion of users into any study necessarily makes IIR, in part, a behavioural science. As a result, appropriate methods for studying interactive IR systems must unite research traditions in two traditions which can be challenging.' [81]. Consequently, due to its multidisciplinary nature, one of the major challenges in IIR research is the lack of standardised metrics and limited guidance on how to conduct research studies.

Toms [152] noted that the development of IIR into a separate research discipline was gradual and included the following stages:

1. The focus of online searching shifted from an evaluation of search outcomes to testing experimental variables and their impact on user performance. This was reflected in a more systematic design of research protocols of human-based experiments.

2. HCI studies emerged in the 1980s giving rise to user-center design, bringing more attention to user requirements and intentions. Consequently, task-based design gained importance in the process of system development and put more emphasis on system usability.

3. In 2000 Borlund [21] introduced the concept of the simulated work-task. This kind of task provided a contextual situation which motivated the search problem.

4. Information retrieval shifted from satisfying topical relevance to meeting user specific relevance. Overall, it was not the outcome itself but rather the process of getting to that outcome that defined the final success. [152]

The complexity and difficulty of evaluating IIR studies was highlighted by Järvelin, who pointed out that: 'when one brings a human-actor (the user) into the information retrieval setting, all standardisation of evaluation disappears. There is no single experimental design to follow' [74]. Järvelin highlighted that people use retrieval systems as they offer them support in carrying out a specific task, arriving at a particular goal. Therefore, it is most relevant to assess the subjective and objective benefits that such a system brings to their performance - these are used as benchmarks of system quality. Due to the variability of real-life scenarios and the large number of factors that need to be taken into consideration, easily measurable goals need to be employed as proxies for any evaluation of information retrieval systems [74].

Järvelin proposed that in IIR experiments, the following variables should be examined:

- dependent variables: which are examined to observe the effects of independent variables (e.g. evaluation metrics)

- independent variables: which are systematically varied in order to observe the responses in dependent ones (e.g. varying information retrieval techniques to see the effect on evaluation metrics)

- concomitant variables: variables that are fixed to prevent uncontrolled variation in the results (e.g. test collections and topics)

- confounded variables (differ between users): knowledge, motivation and skills

According to Ruthven, IIR is shaped by two factors: (1) research on information seeking and search behaviour and (2) research on new methods of interacting with electronic resources [131]. The merit of information seeking research is that it provides a broader context and sheds light on decisions that are involved in finding information. These insights can then be used in the design of systems that facilitate information access. Ruthven claimed that two aspects of search task are most important to IIR: (1) background activity that motivates the need to search and (2) the task to be fulfilled (e.g. obtaining resources or information). The support that the system offers a searcher in completing their task while also helping them understand how it

operates, are the pivotal factors in success and satisfaction. The key reason that IIR studies should continue as a dedicated field of research is the unceasing creativity which characterises human interaction with IIR systems [131].

In the current thesis interactive scenarios involving contextualised search tasks, inspired by Borlund's simulated-work tasks [21], are used to evaluate conversational search agents. In line with the recommendations proposed by Järvelin [74], our evaluation considers both objective (performance) and subjective (a user's perception of the agent) metrics, and is referred to as user search experience (see Section 3.4 for a more detailed explanation of evaluation metrics). In order to provide a theoretical underpinning for interactive information retrieval, in the following section we will discuss several information search models.

2.3 Information Search Models

Models of information seeking are used to explain information seeking actions, their motivation and outcomes [175]. Models are derived from observations on how people searched for information in an interactive task. Searching can either be done on their own or with intermediaries (reference librarians, automated systems, etc.)

There is a core set of actions within the model that a user of an information system needs to use to express their information need to a system. In IIR, the information seeking search model process includes: submitting a query, examining the results and reiterating the search process up until user's information need is satisfied. The most-well known models include: Belkin 's anomalous states of knowledge [17], Saracevic's stratified model of information retrieval [134], Elilis' behavioural approach to information retrieval system design [56], Kuhlthau's model of library search process [89] and Marchionini's Information seeking in Electric Environments [101].

In conversational search, an information need is resolved via dialogue between the information seeker and information provider (intermediary). The following section provides the explanation of the concept of dialogue acts that is pertinent to conversational search.

2.3.1 Modelling Information Search through Dialogue Acts

According to Austin, each utterance spoken in dialogue can be considered as a dialogue act that is performed by the speaker in order to communicate their intentions (cf. [8]).

The COnversational Roles (COR) model, proposed by Sitter and Stein [144], incorporates dialogue acts into the information seeking process. In the COR model, information seeker and intermediary interact through a set of dialogue acts, i.e. directive, commissive and assertive in order to: ask, make and reject offers until the information need is resolved. The model, presented in Figure 2.3, determines all legitimate types and sequences of possible dialogue acts that correspond to predefined actions used in information seeking dialogue.

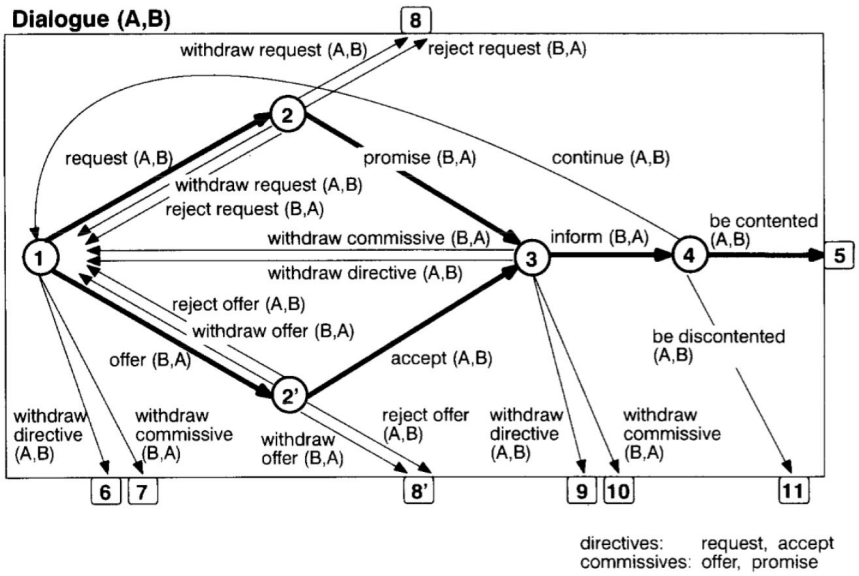


Figure 2.3: COR Model - Diagram adapted from [144]. Seeker/intermediary roles are represented by A/B respectively. Circles represent dialogue states and are linked by arrows that represent state transitions. Each transition corresponds to action taken by the seeker or intermediary. Squares represent terminal states which mark the end of a dialogue.

In order to apply the COR model to information search activity two pre-requisites must be met: (1) there needs to be a preexisting information need to be resolved and (2) both seeker and intermediary must participate in dialogue until the information need is resolved [144]. The COR model relies on the assumption that both conversation participants are able to fulfill each other's expectations completely and immediately [144]. This implies that the seeker (A) is fully able to express their information need and the intermediary (B) can correctly interpret it an take appropriate action. The inherent complexity of human language makes information seeking though dialogue challenging as it requires clarifications and checking of understanding. The following section will illustrate the complexity of human language based on an example dialogue.

2.4 Why is Conversational Search Challenging?

The dialogue below illustrates why interpreting the meaning of natural language can be a challenging task for machines.

TARS: [*as Cooper repairs him*] Settings. General settings. Security settings.

COOPER: Honesty, new setting: ninety-five percent.

TARS: Confirmed. Additional customisation.

COOPER: Humor, seventy-five percent.

TARS: Confirmed. Auto self-destruct sequence in t minus ten, nine...

COOPER: Let's make that sixty percent.

TARS: Sixty percent, confirmed. Knock, knock.

COOPER: You want fifty-five?

[*a long, pronounced silence*]

(INTERSTELLAR [109])

A successful conversation requires its participants to understand language not only at the level of literal meaning of words and sentences (semantics) but also to incorporate the knowledge about the world and contextual information, in order to be able to interpret implicit clues and a speaker's intentions (pragmatics). The example above is a dialogue from a sci-fi movie Interstellar, where a NASA pilot Joseph Cooper calibrates the configuration settings of an intelligent robot, TARS. The robot shows pragmatic understanding as it correctly interprets the intention of Cooper. By asking: 'You want fifty-five?' Cooper implicitly tells TARS to stop cracking jokes and, effectively, orders him to 'shut up'. Although such a dialogue would not pose any comprehension difficulty to an average human interlocutor, it is something that present day Conversational Agents (CA) are still incapable of. However, in recent years we have seen a lot of progress with CAs becoming more human-like. In order to understand why this is the case let us consider the features of human conversation, and illustrate how dialogue systems are designed to incorporate them.

2.4.1 Human Conversation

As summarised by Jurafsky and Martin [79], these are the following features of human conversation:

- **Turns**: conversational partners make contributions sequentially by taking turns.
- **Speech Acts:** are used to communicate certain intentions: request, enquire, negate etc.
- **Grounding:** participants need to establish what they both agree on - common ground [146]. It means that a speaker acknowledges what their conversational partner is saying.

- **Initiative:** it is quite common in human-human conversation for the initiative to switch between participants.

- **Dialogue structure:** in conversation there are adjacency pairs, a QUESTION is followed up by an ANSWER, a PROPOSAL can be followed by an ACCEPTANCE or a REJECTION. They can help an interlocutor to decide which action to take next. Dialogue acts are sometimes separated by side sequences [75] or sub-dialogues. These interjections may not be directly related to the main dialogue but request clarification or refer to a new direction of conversation.

- **Inference and Implicature:** a speaker expects their conversational partner to draw some inferences. In his theory of conversational implicature, Grice [67] proposed that what enables hearers to draw inferences is that conversation is guided by a set of maxims, general heuristics that play a guiding role in interpreting conversational utterances. One such maxim, is the maxim of relevance which says that speakers attempt to be relevant, rather than uttering random speech acts.

The fact that conversational partners need to seamlessly navigate all of the above features in order for communication to be successful and natural, makes the design and implementation of conversational agents a challenging task [79]. In the following section we will illustrate how dialogue systems (fully automated instantiations of conversational agents) are implemented for simple goal-oriented tasks.

2.4.2 Goal-oriented Dialogue Agents

Goal-oriented dialogue agents are designed to complete a particular task (booking a service, looking up information) via conversation with a user. They include digital assistants on cellphones (e.g. Apple Siri, Microsoft Cortana etc.) or stand-alone devices (e.g. Amazon Echo, Google Home etc.) The architecture of a goal-oriented dialogue agent is presented in Figure 2.4. In order to navigate the complexity of human language, goal-oriented agents rely on a combination of different speech processing and dialogue management modules. At the beginning of the interaction, the user utters a request. This utterance in the audio form is then converted to text by the automatic speech recognition (ASR) module. Next, the recognised words are transformed into meaning representation by the spoken language understanding module (SLU). The output of the SLU is forwarded to the dialogue state tracker (DST) whose goal is to estimate the current state of dialogue. The new dialogue state is then passed on to the dialogue policy manager that decides on a suitable action to take. Finally, based on the output of the policy manager, natural language generation (NLG) and text-to-speech (TTS) the user with the agent's response.

Overall, the goal of the dialogue agent is to extract three elements form the user's utterance: (1) domain classification - what is the user talking about, (2) intent determination - what task do they want to accomplish and (3) slot filling - extract the information that the user wants the system to understand (cf. [79]). The concept of the dialogue state tracker is central to conversational search as the CA needs to keep track of the dialogue history and correctly estimate the current state of dialogue (e.g. keeping track of a user's preference for type of food or price range).

Dialogue state tracking is challenging due to the fact that errors at the beginning of the conversation process - ASR, and SLU - are propagated down the pipeline and can lead to the agent misunderstanding the user. Incorrect interpretation may lead to suggesting a different service or making inadequate recommendations. For instance, the incorrect recognition of a user's utterance as Killermont street instead of Richmond street would result in an incorrect assignment of departure place (FROM: slot) and consequently in the taxi being sent to the wrong pickup point.

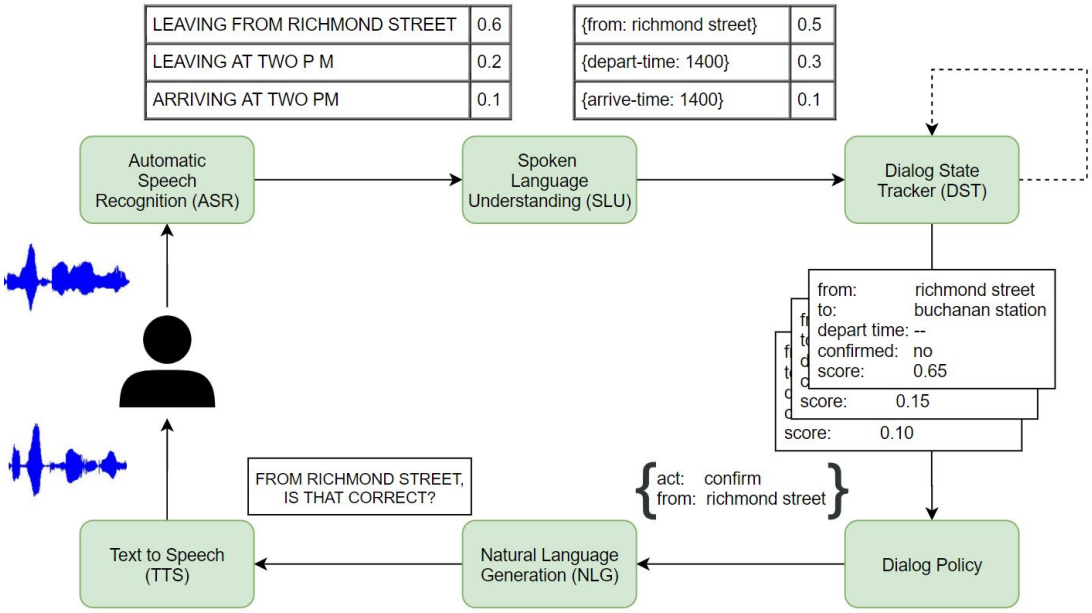


Figure 2.4: Architecture of a dialogue-state system for goal-oriented dialogue. This is a modified version of diagram featured in Williams et al. [174]

Of course, thanks to significant improvements in ASR [181], dialogue agents are now capable of handling simple tasks such as one-shot factoid queries that do not involve many conversational turns and preserving a conversational state over a long period [3]. Table 2.1 provides an overview of different conversational search tasks along with the performance of currently available systems. As shown in Table 2.1, exploratory search and tasks that require more system initiative and more turn-taking are challenging for the dialogue system.

Table 2.1: Examples of Conversational Search Tasks.

Task	Description	Example Query	Performance
Factoid Queries [78]	Queries that do not require user to followup.	'How tall is Ben Nevis?'	Good
Exploratory Search [161]	Search task that requires more thorough exploration of the topic. Comparing information from different sources.	'Tell me about the impact of the 2008 financial crisis on Scottish Economy'	Poor
Open Domain Conversation [130]	Open discussion about interests.	'What was your favourite subject at school and why?'	Poor
Product Search [177]	Finding a product/service and comparing it with available alternatives.	'I am looking for good hiking boots.'	Mixed - depending on how much exploration is required

The importance that a conversational agent can handle a multi-turn conversation with a user is reflected in recent theoretical frameworks that build up on the COR model. The following section discusses the conversational search frameworks that are most relevant to this thesis.

2.5 Conversational Search Frameworks

There are only a few models that account for the 'active' role of the system in the information seeking process [11, 126, 156, 162]. In this section we will focus on the frameworks developed by Radlinski and Craswell [126], and Azzopardi et al. [11] as they are the most closely related to the current thesis.

2.5.1 Theoretical Framework of Conversational Search - Radlinski and Craswell

Radlinski and Craswell [126] posit that in order to answer a variety of users' information needs in an efficient manner, the conversational system should: allow natural language requests, propose search results and ask users for clarification if necessary. The user should be able to provide feedback on the system's suggestions, including negative feedback. Over time, the process should allow the system to build a cumulative picture of the user's information need over the course of conversation, based on their query statements and relevance feedback. Consequently, users do not have to repeat their information need.

Radlinski and Craswell put forward five properties that a search agent needs to have in order to be conversational. The properties are:

1. *user revealment* where the user discloses to the agent their information needs,
2. *agent revealment* where the agent reveals what the agent understands, what actions it can perform, and what options are available to the user,
3. *mixed initiative* where both the agent and the user can share the initiative and direct the conversation,
4. *memory* where the agent tracks and manages the state of the conversation, the user's information need, etc., and,

5. *set retrieval* where agent is able to work with, manipulate and explain the sets of options/objects which are retrieved given the conversational context.

The above properties are required so that the agent can facilitate a search by helping the user formulate their information need and build expectations regarding its capabilities. During the search process, the agent takes the initiative and uses memory to retain information relevant to the query. Before presenting the results back to the user, the agent needs to reason about the usefulness of the retrieved information and decide on what to present to the user. Crucially, for the system to be truly conversational it must have memory and be able to maintain state. This allows information to be elicited from the user in a piecemeal fashion as the conversation progresses, rather than requiring the user to provide all the information upfront. The memory component: (1) allows the agent to remember what was previously said by the user, or itself, to help resolve the information need and (2) gives the user the option to explicitly reference past information to indicate what statements are correct and what information should be forgotten [126, p.4].

The conversation comprises a series of back and forth exchanges that consist of elicitation and revealment actions, i.e. asking and providing information. Each time it is the agent's turn to speak, it will: (1) select an action to perform and (2) request that the user provide a specific type of response. While Radlinski and Craswell [126] provided a general overview of possible actions they did not specify the exact type of actions that are available to agents and users in the interaction space.

2.5.2 Conceptualizing agent-human interactions during the conversational search process - Azzopardi et al.

Building on Radlinski and Craswell's framework [126], Azzopardi et al.'s framework [11] provides specific examples of how information can be conveyed in conversational search. Before presenting the results back to the user, the agent needs consider the usefulness of the retrieved results and decide what it needs to present. Azzopardi et al. [11] assume that a conversational agent has two goals: (1) to help the user resolve their information need and (2) to help the user to understand the space of options that are available to the user. The goals are achieved through specific actions and interactions that are available to a user and the agent.

The overview of actions and interactions is presented in Table 2.2. Azzopardi et al. [11] assume that any information need has a related sub-set of information needs that are associated with it. The 'inquire actions' are requests for revealing part of search space by agent to the user, while 'reveal actions' concern how the information is presented between the interlocutors; the agent discloses the information that it found to the user and the user makes their query more specific as requested by the agent. Over the course of the conversation, the agent will need to decide which way of presenting information is the most suitable (e.g. listing vs. summarising

the results). The agent needs to actively update the state of the conversation, and, at some point make a decision about whether to (1) continue eliciting requirements or (2) reveal the options given the information need.

Table 2.2: An Overview of the Actions and Interactions available to User and Agent - as proposed by Azzopardi et al. [11]

		User	**Agent**		
	Query Formulation [156]	**Reveal** Disclose Non-Disclose Revise Refine Expand	**Inquire** Extract Elicit Clarify	User Revealment [126]	Memory [126]
Set Retrieval [126]	Result Exploration [156]	**Inquire** List Summarize Compare Subset Similar	**Reveal** List Summarize Compare Subset Similar	System Revealment [126]	
		Navigate Repeat Back More ... Note	**Traverse** Repeat Back More ... Record		
Mixed Initiative [126]		**Interrupt**	**Suggest**		
		Interrupt	Recommend Hypothesize		
		Interrogate Understand Explain	**Explain** Report Reason		

Azzopardi et al. [11] argue that conversational actions and interactions should be employed by the conversational agent so as to (1) 'Minimise the Conversational Effort' and (2) 'Maximise the number of the options provided so that users have a good overview of the search space and can make an informed choice'. There is a tension between the two above goals as the capacity of human working memory imposes limits on how information can be provided verbally without overloading the user. In this thesis we consider the impact of different conversational actions in terms of their cognitive implications, as well at the impact on satisfaction with presented options and search performance. These factors are collectively referred to as user search experience (discussed in Section 3.4 in more detail). Before progressing to a discussion of the methodology employed in current research, let us briefly discuss relevant studies that have focused on an empirical evaluation of conversational agents.

2.6 Relevant Studies

In this subsection, we present relevant work that focuses on the evaluation of CAs from three different perspectives, namely: (1) automated dialogue systems (Section 2.6.1), (2) sociolinguistics (Section 2.6.2), and (3) interactive conversational search studies where: (a) the CA is simulated by a human or (b) two people who collaboratively seek information.

2.6.1 Dialogue Systems Evaluation

In current thesis, we consider a dialogue system (DS) to be a fully automated instance of a CA that is based on a slot-filling architecture (cf. [79]). In the 1980s, the widespread understanding and application of the theory of Hidden Markov Models to speech processing fuelled major improvements in the automatic recognition of continuous speech [125] that made spoken interaction with computer systems feasible. Since then we have seen a rapid rise in research on the development and evaluation of DS's.

Task completion success is a core measure for evaluating the performance of dialogue systems. Related to DS performance are the efficiency cost (the measure of a system's support in resolving user queries) and the quality cost (the impact of speech processing modules that affect users' perception of the DS). Traditionally, efficiency metrics used for DS evaluation were: total time of task completion (the shorter the better) and number of conversational turns (the fewer the better), while the quality cost metric was the performance of the automatic speech recognition (ASR) module [118, 141, 184]. PARADISE (PARAdigm for DIalogue System Evaluation) is a seminal evaluation framework by Walker et al. [167] that combines a set of performance measures (i.e. user satisfaction, task success and dialogue cost) into a single evaluation function. The PARADISE framework featured a survey with questions regarding: (1) Text-to-Speech (TTS) Performance, (2) ASR Performance, (3) Task Ease, (4) Interaction Pace, (5) User Expertise, (6) System Response, (7) Expected Behaviour and (8) Future Use - that provided fine-grained insights into user satisfaction with the CA. Based on evaluations conducted on the DARPA Communicator corpus [93] which contains a collection of dialogues in a flight booking domain, the 'Future Use' question proved to be the best indicator of user satisfaction, while the system turn duration was found to be the best predictor of performance (cf. [166]). It should be noted that while the PARADISE framework enables a comparison of different CAs based on task complexity, it does not reflect that some solutions may be better than others. For instance, while a certain flight may get a participant to their destination faster it would be considered as an equally good option as a longer flight. In order to address this limitation, in the current thesis performance is evaluated in terms of 'Pareto Optimality' (explained in Section 3.4.2.1).

As the natural language processing capabilities of CAs improved (esp. ASR and natural language understanding) the focus of evaluation shifted to the impact of different information

presentation strategies on task success. Three notable approaches to CA information presentation are: (1) Summarise and Refine (SR) [117] where a large number of options is grouped into small clusters that are summarised and presented to the user for further refinement , (2) the User-Model based (UM) approach [104] which relies on decision theory techniques that are used to present a small number of options that best match a user's preferences and (3) the User-Model based Summarize and Refine (UMSR) [40] that orders options for step-wise refinement based on ranking attributes in the ranking model. Unlike UM, after presenting all the relevant options to the user, USMR also provides the user with a brief account of the irrelevant options to increase users' confidence that they have thoroughly explored the search space. Demberg et al. [41] conducted an interactive experiment in which they evaluated the USRM against the SR approach. The experiment combined the use of performance metrics (task completion success, duration of interaction and number of dialogue turns) with a subjective questionnaire regarding the overview of the search space (see Figure B.3 for reference). Demberg et al. [41] found that USRM led to fewer dialogue turns and a quicker completion time than SR. In terms of subjective metrics, although participants indicated a preference for USRM, the difference in ratings did not reach a statistically significant level. In a related study, Polifroni and Walker [119] developed different variants of the SR approach that differed in clustering methods and the presentation of results. Polifroni and Walker [119] found that when many options were available, users preferred utterances that were generated based on the User-Model (i.e. tailored specifically to requirements based on user profile: a tourist vs. person with knowledge of the local area) rather than a generic summary of results.

Another strand of research explored the use of machine learning approaches for simulating user dialogue behaviour (e.g. [63, 64, 70, 137]). In these approaches, models of user behaviour are learned from corpora of human-computer dialogues in order to develop an optimal DS interaction strategy [136]. DS interaction strategy is tested through trial-and-error against different models of user behaviour. The performance of a dialogue system can be evaluated by measuring the quality of dialogue or success of generated dialogue (e.g. in terms of the number of filled and grounded slots) [136]. The score obtained with the learned strategy could be then compared to the handcrafted interaction strategy or by running comparative studies between different modelling techniques [137]. Over the last decade, statistical user simulation has played an important role in improving the performance of dialogue systems and making them more natural [1].

As discussed above, measures of evaluating DSs were mostly quantitative in nature and concerned with the performance of system modules (e.g. ASR, TTS, dialogue manager etc.). We will now turn to the field of socioliguistics which provides qualitative, user-focused methods that can be used to explore how people communicate with conversational agents by analysing semantic and/or pragmatic aspects of language.

2.6.2 Sociolinguistics

Sociolinguistics is a discipline that is concerned with the relationship between language and the context of its use [72]. Within sociolinguistics, among other approaches, we can distinguish conversation analysis [132] and discourse analysis [169]. Both of these approaches draw from the tradition of ethnomethodology [62], are qualitative in nature, and focus on analysing the functional and sense-making properties of language [179]. While conversation analysis is focused on observing the use of language in social interactions (i.e. how language is organised), discourse analysis is concerned with the function of language (i.e. how language is used to accomplish various goals) [179]. In this subsection, we will present previous research that employed conversational analysis, and briefly comment on relevance of a discourse analysis approach for the evaluation of CAs.

Apart from Gilbert et al. [65] who examined the potential of conversation analysis in the design of human-computer interaction, until quite recently, research in this area was scarce. Examples of recent work on conversation analysis include: an exploration of interactions with CAs on mobile devices [121], explication of how users attend to problems encountered during interaction with Amazon Echo in a home setting [120], and investigation of interaction strategies in human-robot interaction [114]. Using conversational analysis is helpful to understand what strategies users employ to overcome problems that they encounter when communicating with CAs, such as adjusting speaking style to improve the performance of automatic speech recognition [114] or rephrasing and repeating requests to overcome communication problems [120].

While conversation analysis is based on the view that human language is neither formalisable nor bound by rules [25], discourse analysis considers language in terms of 'purposive joint activities between two speech partners' [169, p.29] that can be defined based on the goals of the speech partners [99]. Walton and Krabbe [168] proposed a taxonomy that distinguishes six types of dialogue based on its purpose, namely: (1) Persuasion, (2) Negotiation, (3) Inquiry, (4) Deliberation, (5) Information-seeking, and (6) Eristics. Building on Walton and Krabbe's taxonomy [168], Macagno and Bigi [100] introduced the notion of 'dialogue move' to account for the dynamic nature of individual goals that change over the course of the dialogue. Discourse analysis could be used in information seeking tasks with CAs to better understand how a user's goals change over the course of the interaction. In particular, a product search where user is required to explore a space of available options in order to make an informed choice, would be suitable for this method of analysis.

We have presented relevant literature on the evaluation of conversational agents from the perspectives of dialogue systems and sociolinguistics. The final section of the literature review will cover interactive information seeking studies, the research area that is most closely related to the topic of this thesis.

2.6.3 Interactive Information Seeking

The summary of recent studies (since 2017, the beginning of my PhD) that focus on how people interact with conversational search agents is provided in Table 2.3. The selection of the studies was informed by a literature review conducted in the domain of conversational agents for information retrieval. The studies are categories based on their type as: Wizard of Oz (WOZ) - where the agent is simulated by a human [9, 15, 50, 52, 60, 82, 83, 164, 165], or Human-Human (H-H) interactive search studies where one person acts as seeker and the other is an intermediary who has access to a search engine [150, 157]. Each type of study can be delivered via text, voice or a combination of these two modalities.

The focus of interactive conversational studies is on ascertaining how people behave and interact with the agent, without imposing a pre-defined interaction policy regarding how participants should act; instead participants are provided with search tasks and roles but not explicitly instructed on how to complete them (cf. [102]).

It should be noted that while the studies presented in Table 2.3 provide important insights into users' interactions with conversational agents, they only account for some aspects of user search experience (i.e. cognitive load, satisfaction and performance) but do not (with the exception of [50, 52]) explore the interplay between these aspects *while* also accounting for user feedback. In our research [50, 52], we provide a broader overview by exploring both the impact of the system on a user's cognitive load and satisfaction, as well as evaluating how well the agent supports user performance and the execution of tasks. We also identify the user's requirements and expectations regarding conversational agents, by using semi-structured interviews that follow the interactive part of each experiment (discussed in Sections 4.4.3, 5.5 and 6.5). Another contribution of the research presented in the current thesis are the insights on the impact of variations in conversational task complexity and search criteria on the outcome of the task - i.e. flight options selected by the users (presented in Chapter 7).

Table 2.3: Summary of Recent Interactive Information Retrieval Studies in the Conversational Search Domain.

Authors (Year)	Study Type	Modality	Description
Avula et al. (2018) [9]	WOZ	Text	Avula et al. explored user engagement in a collaborative search scenario. In the interactive search experiment, participants were provided with information either by a dynamically intervening chatbot or a chatbot that monitored the conversational channel. The results indicated that dynamic chatbots enhanced users' collaborative experience and that the bot's intervention type did not hugely impact users' perception or engagement levels.
Barko-Sherif et al. (2020) [15]	WOZ	Voice/ Text	Barko-Sherif et al. developed a conversational framework and evaluated it empirically by similating a conversational agent for recipe recommendation tasks. The participants who interacted with the agent tend to use more human-like language when speaking to the agent as compared to typing.
Dubiel et al. (2018) [52]	WOZ	Voice	Dubiel et al. compared the impact of a passive vs. active conversational agent on users' cognitive load, satisfaction and performance in a goal oriented task. They found that an active agent led to higher user satisfaction, less cognitive strain and promoted more positive sentiment towards the agent.
Dubiel et al. (2020) [50]	WOZ	Voice	The study explored the role of different elicitation and revealment strategies on user search performance. While there were no statistically significant differences between how the strategies were perceived by users, there was a difference in performance with active agents helping users to find cheaper and faster flight options.
Frummet et al. (2020) [60]	H-H	Voice	Frummet et al. conducted an in situ experiment to collect conversationally described information needs in a home-cooking scenario. The CA was simulated by a human to facilitate a cooking task. The authors used word embeddings to obtain deeper semantic representation of participants' information need that goes beyond the surface structure of the utterance. Interestingly, using stop-words provided pragmatic insights that led to higher accuracy in predicting the category of information need.
Kiesel et al. (2019) [82]	WOZ	Voice	Kiesel et al. explored how a conversational agent can suggest query corrections to users during an interactive task. They found that users were more satisfied when the system corrected them, even when correction attempts were unsuccessful. The tone of clarification had an impact on the user satisfaction as well.
Kiesel et al. (2020) [83]	WOZ	Voice	Kiesel et al. conducted a focused user study to explore what an argument search should look like. Participants interacted with a simulated voice search system to fulfil their information needs. The results indicated that, to trust the system, participants required the system to provide rich information on the retrieved arguments along with source of information, supporting evidence as well as have a background knowledge on the presented events or entities.
Thomas et al. (2017) [150]	H-H	Voice	Thomas et al. provided pairs of volunteers with information-seeking tasks acting out the role of intermediary and seeker. The task was conducted via VPN. The authors recorded the interactions of volunteers and used video and audio feedback to analyse their effort, engagement and satisfaction.
Trippas et al. (2017) [157]	H-H	Voice	Trippas et al. conducted a study in which participants completed three search tasks of different complexity in an acoustic only setting . As the task complexity increased participants spend more time interacting.
Vtyurina et.al. (2017) [165]	WOZ	Text	In an interactive search scenario Vtyurina et al. compared how user satisfaction, ability to find information, and ability to answer topical quiz questions varies when interacting with a commercial chatbot compared to a human or a simulated system. They found that participants were significantly more satisfied when interacting with a human or a wizard as opposed to an automated system.
Vtyurina and Fourney (2018) [164]	WOZ	Voice	Vtyurina and Fourney explored the role of implicit conversational cues such as 'ok' and 'yeah' on task completion in a cooking scenario. Although the current generation of conversational assistants does not tackle these cues they are a natural way in which users communicate and should be identifiable by agents to make conversations with them more natural

Regardless of the recent explosion of interest, conversational search is still a nascent research area with many open research questions. In particular, interactive outcome based search evaluations are required to provide insights on the impact of different conversational search systems on users' search performance and the implications on cognitive workload (especially important in the context of voice only interaction [182]). Also, to date, little work has addressed the problem of the design of goal-oriented search tasks that could be used to evaluate conversational assistants [53]. The current thesis starts to address these gaps by seeking to provide an answer on how to design more robust and user friendly voice-only CAs for goal oriented tasks.

2.7 Research Goal

Conversational search frameworks proposed by Radlinski and Craswell [126] and Azzopardi et al. [11], presented in Section 2.5, postulate that a CA should actively support a user during the conversation search process to better understand the search space and resolve the information need. The current thesis aims to provide an empirical bridge to the theory of conversational search by investigating the impact of CAs with different levels of support on user search experience. Specifically, through three interactive information seeking studies (discussed in Chapter 4, 5 and 6), we investigate the impact of CAs on four key aspects of user search experience, namely:

1. Cognitive Workload
2. Satisfaction with the Agent
3. Task Performance
4. Interaction Time

The selection of the four aspects of user search experience was informed by the literature review, presented in Section 2.6. By exploring the impact of these four aspects we seek to answer our overarching research question: **'How much would a truly conversational agent with the ability to preserve state (memory) and conversational initiative improve user search experience compared to a passive, non-interactive slot-filling system?'** Answering this question will provide empirical evidence to validate theoretical frameworks [11,126] and, in turn, help to improve the design of voice-only conversational agents. Firstly, investigating cognitive workload will provide a better understanding of the impact of different methods of presenting information on users and explore the trade-off between the amount of information presented and the effort in using the CA [158]. Secondly, satisfaction (understood as the 'fulfillment of a specified desire or goal' [81, p.120]) will serve as a proxy for overall contentment with the CA [86,87,167] and an indicator of future use [49]. Thirdly, task performance will be used to assess if a higher level of CA support could allow users to achieve better results (i.e. to maximise their performance) [11,167]. Finally, by looking at interaction time, we will explore the impact of CAs on the efficiency of task execution (i.e. CA impact on minimising interaction cost) [11,167].

The four aspects of user search experience are investigated through a combination of subjective and objective metrics (described in detail in Section 3.4). To provide a fuller understanding of user search experience, in addition to the quantitative data, we also analyse participants' feedback, gathered during semi-structured interviews that followed each interactive study. Qualitative feedback will serve as means of contextualising the obtained quantitative results and of aiding our interpretation of participants' behaviour during interactive tasks [42].

All interactive studies presented in the current thesis were conducted using a Wizard of Oz (WOZ) framework [38] (explained in Section 3.2). The selection of the WOZ framework was motivated by its flexibility and potential to simulate multi-turn information seeking conversations that require a thorough exploration of the search space and a comparison of results. Unlike an automated dialogue systems approach, WOZ is not bound by the limitations of the individual components of a system, such as Automatic Speech Recognition or Dialogue Manager. We are mindful that our evaluation method is also subject to limitations. In Section 3.2, we will weigh the WOZ up against alternative approaches to evaluating conversational agents.

2.8 Chapter Summary

This chapter provided background information about previous research on conversational search and contextualised it within the area of information retrieval. This chapter:

- Provided a historical background on research on conversational systems; highlighting improvements in speech processing technology that led to the development of dialogue systems.
- Defined conversational search and explained how it differs from the traditional IR paradigm. We explained that the interactive character of conversational search requires the active participation of both intermediary and seeker.
- Introduced the Conversational Roles Model (COR) which conceptualised a set of actions that could be taken by both intermediary and seeker in an information seeking dialogue.
- Explained challenges created for conversational agents by the complexity of human language, arising from the fact that the meaning in conversation is built incrementally and requires interlocutors to attend to information that is communicated at different stages of conversation (memory) and act on it (initiative).
- Illustrated how dialogue systems work in practice by showing how different modules allow to process spoken language and execute queries.
- Presented recent conceptual frameworks that build on the COR model and provide theoretical underpinning for the conversational search process; further highlighting the importance of memory and initiative in conversation.
- Provided an overview of relevant studies with their findings and identified research gaps that are addressed in the current thesis.

In the next chapter we will provide an explanation of the methodology that was used to design and run our three studies, and discuss the concept of user search experience used to evaluate conversational agents along with the corresponding metrics used in evaluation.

Chapter 3

Methodology

This chapter provides a description of experimental methods used in this thesis. First, Section 3.1 introduces the concept of interactive system evaluation - providing insights from HCI and IIR perspectives. Next, Section 3.2 describes a Wizard of Oz framework [38] as a tool for evaluating conversational agents. Then, in Section 3.3 we provide an overview of the global experimental pipeline used in the three Wizard of Oz studies featured in this thesis, along with a description of search tasks, types of conversational agents used, and evaluation metrics. Section 3.4 explains objective and subjective measures that were used to evaluate user search experience. Finally, Section 3.5 provides a summary of the current chapter.

3.1 Interactive System Evaluation

Kelly argues that laboratory studies are useful in the process of development and evaluation of information retrieval systems as they provide the researcher with a good level of control [81, p.28] 'Laboratory studies are good with respect to the amount of control researchers have over the study situation. This is particularly useful when trying to isolate the impact of one or more variables. Of course, one perennial criticism of laboratory studies is that they are too artificial, do not represent real life and have limited generalizability.' In our experimental setup we aim to investigate the impact of particular features of a conversational agent such as memory and specific information presentation techniques. Specifically, we will use explanatory studies, which are a type of laboratory study, to examine the relationship between different types of conversational agents and their impact on user search experience. As defined by Kelly et al., 'Explanatory studies examine the relationship between two or more variables with the goal of prediction and explanation. Explanatory studies are often concerned with establishing causality and because of this require variables of interest to be isolated and studied systematically. Explanatory studies use more structured and focused methods than exploratory or descriptive studies and involve hypothesis testing. Despite the name, it is important to note that not all explanatory studies offer explanations - many just report observations and statistics without offering any explanation. It is also important to distinguish between prediction and

explanation: it is possible to build predictive models of events without actually understanding anything about why such events occur. Very often researchers stop at prediction and do not pursue explanation, but it is actually explanation that is tied most closely to the theoretical development.' [81, p.26]

The design of conversational search agents can be considered from both HCI and and IR perspectives. Both approaches will be discussed in turn in the following subsections.

3.1.1 HCI Perspective

Go and Carroll underline the importance of using scenarios in the process of system design and evaluation as they allow for dealing with several aspects of the problem simultaneously and serve as aids to user imagination [66, p.46]. HCI takes a human perspective to provide views on system usability and cognition. 'Human-computer interaction uses scenarios to describe the use of systems and to envision more usable computer systems. To observe and then analyze the current usage of a system, it is necessary to involve authentic users. In this approach, actors in a scenario are specific people who carry out real or realistic tasks.' [66, p.48]. HCI uses scenarios to analyse currently available systems and to envision more usable systems in the future.

In the current thesis, we use scenario-based techniques to provide context for goal-oriented search tasks and make the design of conversational search agents more concrete. Since our conversational search agents are tested in the context of information retrieval the design of our search task will draw from the IIR tradition.

3.1.2 IIR Perspective

Borlund [22] proposed an IIR evaluation model which is a mix of system-driven and user-focused evaluation approaches. The components of the model listed by Borlund, are: 'potential users as test persons; the application of individual and potentially dynamic information need interpretations e.g. sub-component of a work-related simulation, and: the assignment of multidimensional and dynamic relevance judgements, structured post-search interview to follow-up on the test person's expectations and perceptions on their participation in the experiment.'

Borlund [22] emphasised that the use of simulated work-tasks helps to ensure that the experiment has both realism and control. The IIR evaluation model is more user-focused by accounting for the dynamic nature of relevance rather than accounting for only system-based metrics of precision and recall. In this thesis, simulated search scenarios 3.3.2 are used to increase participant engagement and provide a benchmark to compare performance across participants.

3.1.3 Focus of Evaluation

Kelly [81, p.100] enumerated four sets of measures that became the standard in IIR, namely: '(1) demographics, (2) the second set of measures includes those used to characterise the interaction

between the user and the system and user's search behaviours, such as number of queries issued, number of documents retrieved and query length - these measures are typically extracted from data; (3) the third set of measures are performance-based measures related to the outcome of the interaction, such as number of relevant documents saved, average mean precision, discounted cumulative gain; (4) based on evaluative feedback elicited from subjects - such measures often probe subjects about their attitudes and feelings about the system and their interactions with it.' In the current thesis, the combination of different cognitive and performance metrics is used to evaluate user search experience (see section 3.4 for a discussion of all metrics used). All of the interactive experiments featured in this thesis were run using a Wizard of Oz (WOZ) framework [38].

3.2 Wizard of Oz Studies

The phrase WOZ is used to describe a testing methodology where participants are interacting with an experimenter who simulates a software agent - the element of deceit is employed in order to manage participants' expectations and encourage natural behaviour (cf. [38]). The merit of using a WOZ framework is that it enables 'acquisition of causal knowledge through controlled variation' cf. [35] - hence why we used a WOZ setup to address our main research question i.e. 'How much would a truly conversational agent improve user search experience compared to a passive, non-interactive slot-filing system?', in a controlled lab environment. The illustration of individual experimental setups employed in each of the interactive studies is presented the corresponding sections: Study 1 (Section 4.3.3), Study 2 (Section 5.3.3) and Study 3 (6.3.3).

A WOZ framework, like any other research method, is subject to limitations. Table 3.1 lists some of the merits and limitations of WOZ along with the characteristics of two alternatives, i.e. automated evaluation (dialogue systems approach) and conversation analysis (sociolinguistic approach). A more detailed discussion of relevant evaluation approaches is presented in Section 2.6.

Table 3.1: Merits and Limitations of CA Evaluation Approaches.

Approach	Merits	Limitations
Wizard of Oz	+ High flexibility + Ability to simulate complex conversations + Allows for improvisation to avoid dead-ends in conversation	- Consistent implementation of interaction strategy is challenging - Requires extensive training and piloting - CA interaction policy needs to be detailed and clearly defined
Interactive testing with dialogue systems	+ High consistency + Faithful representation of performance of the system + Not subject to human error	- Long development time - Limited ability to preserve state (esp. in long conversations) - Likelihood of speech recognition errors
Conversation analysis	+ High realism (esp. if applied in a home-setting) + Interactions are not limited by pre-defined tasks + Experiments can be conducted in parallel (not limited by lab space)	- Limited comparability of linguistic data between individuals - Researcher has no control over the environment of interaction with CA - Interaction with CA may be prone to interruption by external events

While a fully automated evaluation approach to CA evaluation could have enabled realistic interactions, it was subject to speech recognition errors and problems with state tracking. As for a conversation analysis of CAs in home-setting, despite having the potential to offer rich linguistic data, it does not provide control of the interaction environment, which is required to ensure that all participants are subject to the same experimental conditions. On balance, taking into consideration the nature of the project, the available resources and time constraints, the WOZ approach was selected as our evaluation method. This choice allowed us to simulate fully conversational agents that are not constrained by the technical limitations of an automated system, while also ensuring consistency and comparability across participants, thanks to a controlled lab environment.

3.3 Overview of the Experimental Pipeline

We conducted three WOZ studies to explore different types of conversational agents and their impact on user search experience. Although the selection of questionnaires varied slightly between each, all of the studies followed a similar pattern and consisted of the same experimental stages. Chapters 4, 5 and 6 provide a specific discussion of metrics used in each of the studies.

A global overview of experimental stages is presented in Figure 3.1. Each of our 3 experiments followed this structure. During the first stage participants were welcomed, briefed about the objectives of the study and asked to provide their written consent in order to participate in the experiment. Next, participants were asked to complete a series of interactive search tasks and fill in corresponding questionnaires to evaluate their search experience (discussed in subsection 3.4) until they completed all of the allocated search tasks. In the next stage participants were invited to a post-study interview and finally to attend a debriefing session.

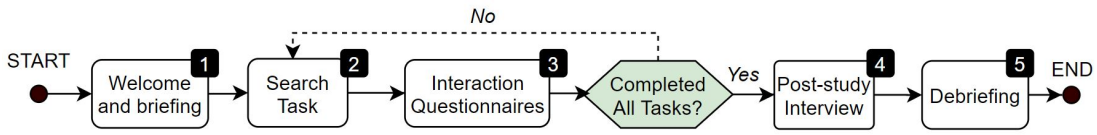


Figure 3.1: A global overview of experimental stages. Stages 2 & 3 were repeated until participant completed all search tasks.

3.3.1 Types of Agents Used in Interactive Studies

The main focus of our investigation is on how the type of conversational agent impacts on user search experience. In the first study (Chapter 4) we establish a baseline by investigating the impact of the agent's memory component (the ability to maintain conversational state). Next (Chapter 5 & 6), we move on to investigate different elicitation and revealment techniques and their impact on user search experience. The selection of features under investigation is based on two crucial elements of a conversational agent identified by Radlinski and Craswell [126],

namely: (1) memory and (2) mixed initiative. These two features are then discussed in terms of their ability to maximise task performance and minimise interaction cost (two goals of a conversational agent identified by Azzopardi et al. [11]).

Based on the conceptual framework provided by Azzopardi et al. [11], we propose a spectrum of conversational strategies (illustrated in Figure 3.2) which determines how conversational agents elicit information (elicitation) and present it back to the user (revealment).

The agent's elicitation involvement increases from left (passive) to right (pro-active). As involvement increases, the CA asks more clarifying questions, helping the user to specify their query (active CA) or makes proactive recommendations to amend the original query. The agent's revealment increases from top (no revealment) to bottom (listing results). As the degree of revealment increases, the CA will provide more detailed results back to the user.

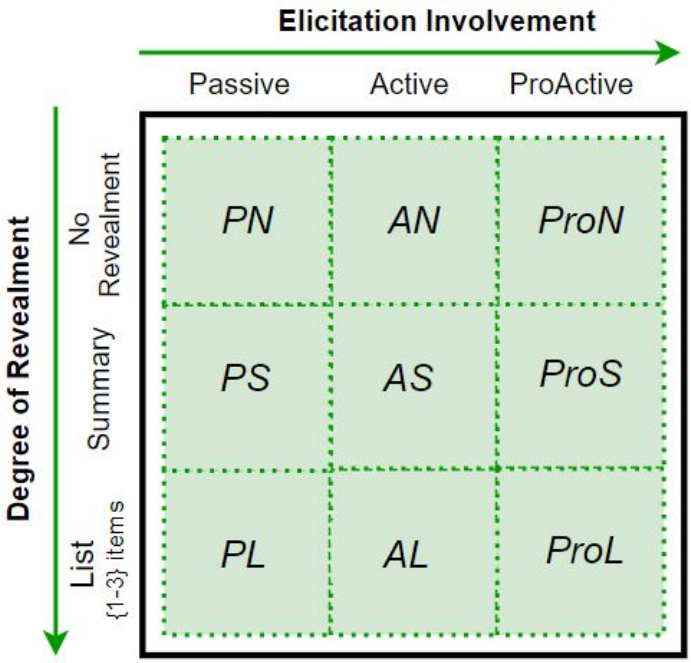


Figure 3.2: Spectrum of Conversational Agent Strategies. The agent's conversational involvement increases from left to right and top to bottom.

The impact of the agent's memory, as well as different conversational strategies used by agents, is evaluated in terms of user search experience based on interactive search scenarios.

3.3.2 Search Scenarios

Our search scenarios were inspired by simulated work tasks proposed by Borlund [22]. We chose simulated search scenarios as they provide participants with an information need that requires several steps to be completed in order to be satisfied. This is in contrast with simple factoid queries, frequently carried out by users [47], that can be resolved with a single query (e.g. 'What is the population of Chile?'). The premise behind using simulated search scenarios was to provide a background context that our participants could easily relate to.

An example search scenario is provided text box 3.1 below. Search scenarios were used to provide a relatable background story and motivate the participants to engage in search tasks. Although phrasing of search tasks differed between each study (more details are provided in the corresponding chapters), each task always contained 3 core elements, namely: (1) Description of the task, providing information regarding destination and the purpose of travel; (2) Indicative request, providing scenario-specific constraints with regards to flight specification and budget constraints (e.g. desired time of arrival and maximum spending allowance), and (3) a note containing additional instructions on the interaction etiquette (e.g. 'Please wait for the conversational agent to finish its turn before you start to speak).

Description of Task: You are traveling to [Destination] to [Purpose of Travel].
Indicative Request: You need to [e.g. required departure time, maximum budget].
Note: Additional information about the task.

Search Scenario 3.1: An example Search Scenario.

3.4 Evaluating Conversational Search Experience

Figure 3.3 illustrates measures used to evaluate user search experience. We used elements of decision theory framework [43] to specify how different factors contribute to the overall *user search experience*. Decision theory is based on the concept of economic rationality, where decisions taken by agents (rational human beings) are considered in terms of meeting a set of objectives that are evaluated based on corresponding measures [43]. In the current work, the objective is to find the cheapest and fastest flight that meets criteria outlined in the search scenario (maximising performance) and to do so in the quickest and least cognitively taxing manner (minimising effort). The evaluation of CAs focuses on four aspects: (I) cognitive workload, (II) satisfaction, (III) task performance and (IV) interaction metrics. Aspects I and II provide subjective insights on how different CAs are perceived by participants, while aspects III and IV provide objective insights regarding the outcome of the search task.

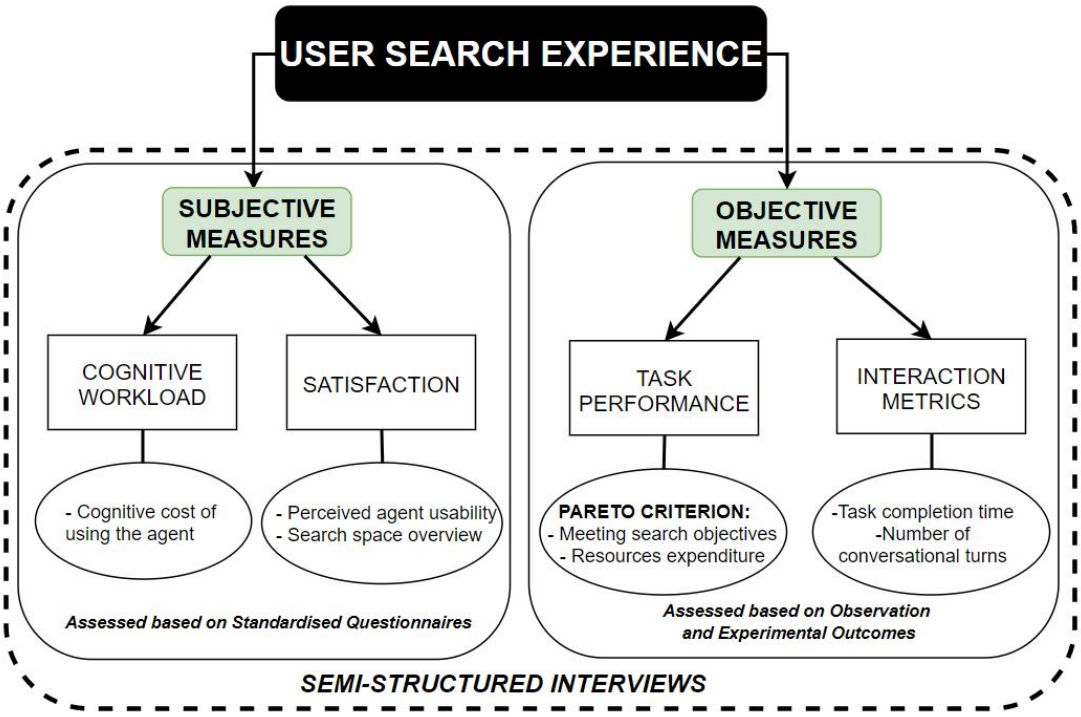


Figure 3.3: User Search Experience divided into objective and subjective measures. Aspects of user experience are presented in white squares together with corresponding metrics that are used to evaluate them (white ovals).

The selection of the four constituent aspects of user search experience was informed by the literature review carried out in the field of evaluation of dialogue system and conversational agents evaluation [85,87,96,98,103,119,167,176]. The aspects were chosen, as their combination provides a holistic overview of the search experience. Our approach builds on core evaluation metrics used in HCI (i.e. usability and mental workload) by adding insights on the agent's retrieval performance and overview of search space (IIR metrics). The essential role that usability and mental workload play in the development and evaluation of interactive systems is highlighted by Longo:

> 'Widely employed in the broader field of HCI, usability and mental workload are two constructs from ergonomics, with no crystal and generally applicable definitions. There is an acute debate on their assessment and measurement. Although ill-defined, they remain extremely important for describing the user experience and improving interface and system design.' [96, p.2]

The importance of employing usability and mental workload concepts while accounting for the context of system usage is paramount for a robust and reliable assessment of any agent or computer system. Accounting for the demands of the executed task is also crucial for describing user experience. In the current thesis, in order to account for all the constituent elements of user search experience, we combine metrics of cognitive workload, satisfaction and performance.

The following subsections provide a description of all the constituent aspects of search experience and their corresponding metrics as presented in Figure 3.3.

3.4.1 Subjective Metrics

3.4.1.1 Cognitive Workload

Participants' cognitive workload while interacting with conversational agents was measured with NASA TLX questionnaire [68]. The questionnaire consists of six items, measured on a Likert scale (20 points per item), that provide insights into the overall cognitive workload. The constituent items of the questionnaire are: mental demand, physical demand, temporal demand, performance, effort and frustration. The final score is calculated by averaging the items out and providing a raw score for the system. The higher the score, the more cognitively taxing the system is perceived to be. In our studies we did not report the physical demand item because it was not relevant to our search tasks. NASA-TLX questionnaire is provided in the Appendix B.1 for reference.

Measuring the cognitive workload of a conversational agent is important because the serial nature of speech makes presenting results over an audio-only channel challenging and taxing to process [182]. The amount of information that can be stored and processed in working-memory is limited [12]. Therefore, using NASA TLX questionnaire is pertinent in a voice-only search task to measure the impact of different ways of presenting information on users' cognitive workload.

3.4.1.2 Satisfaction

To measure participants' satisfaction we combined System Usability Scale (SUS) [24] with additional proxies such as sentiment towards the system (measured with sentiment analysis tool VADER [73]) and satisfaction search results provided by a CA (measured by Search Space Exploration Satisfaction (SSES) [176] questionnaire). Although the use of the instruments measuring satisfaction varied between our studies (full details will be provided experimental Chapters 4-6), SUS was featured in all.

SUS is a standardised questionnaire that measures system usability on a scale from 0 to 100, where 100 signifies the most usable system. SUS has been used to evaluate interactive systems in order to determine their potential for adoption, acceptability, performance and usability. Figure 3.4 provides a visual guide to interpreting SUS scores across these four categories: Net Promoter Score (NPS) [128] which is the likelihood of recommending a system to a colleague or dissuading them from using it; system acceptability; adjectives describing user experience, and a graded score marked on a F-A scale. SUS questionnaire is provided in the Appendix B.2 for reference.

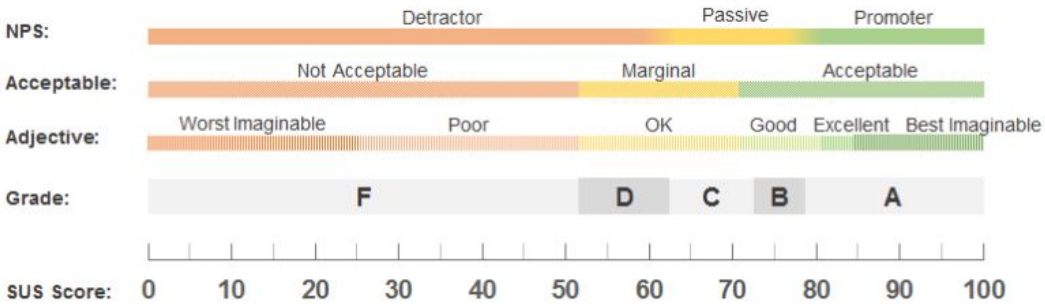


Figure 3.4: Grades, adjectives, acceptability, and NPS categories associated with raw SUS scores. Figure adapted from [147].

SSES questionnaire (featured in studies 2 and 3) consists of four questions that are measured on a 5-point Likert scale - from 1 (strongly disagree) to 5 (strongly agree). The questions concern user satisfaction with: (1) clarity of results presentation, (2) overview of available flights, (3) completeness of presentation (i.e. proportion of relevant options that were presented to the user) and (4) speed of presentation. SSES questionnaire is provided in Appendix B.3 for reference. In Study 1, we used VADER [73] to analyse the sentiment of participants towards the system, based on their utterances. The result of the sentiment analysis was a set of polarities composed of 'positivity', 'negativity','neutral' and 'compound' scores. The compound score is a normalised, weighted composite score that is useful for giving a single measure of sentiment for a specific utterance. A more extensive discussion of sentiment scores is provided in Section 4.3.8.

3.4.2 Objective Metrics

3.4.2.1 Task Performance

Task performance was evaluated in terms of flight options that were selected by the participants. We introduced the concept of Pareto Optimality to evaluate if flight options chosen by participants were optimal or not. The concept of Pareto Optimality is illustrated in Figure 3.5.

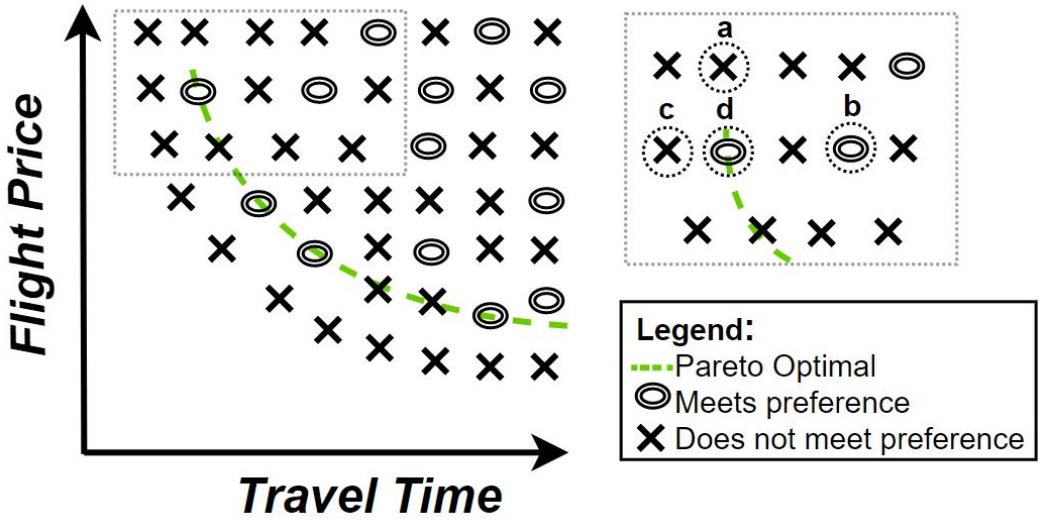


Figure 3.5: Trade-off between flight cost and travel time (left) and a close-up with some example flight selections (right). All selected flights (a, b, c, d) are considered in reference to the closest flight on the Pareto Optimal: a – is an option that wastes money and does not meet the arrival preference, b – is an option that wastes time but meets the arrival preference, c – is an option that saves time but does not meet the arrival preference, and d – is the optimal option.

The Pareto Frontier is formally represented by equations 3.1 and 3.2, where R^{m} is considered to be a space of flights, X represents feasible flight options in this space, and Y is a set of decision vectors such that given our preferred decisions criteria (short travel time and low price), all points on the Pareto frontier dominate over other points in the search space.

$$Y = \{y' \epsilon\, \mathbb{R}^m : y = f(x),\ x\, \epsilon\, X\} \tag{3.1}$$

$$P(Y) = \{y' \in Y :\ \{y'' \in Y :\ y'' \succ y', y'' \neq y'\ \} = \emptyset\}. \tag{3.2}$$

Given a search scenario that requires a participant to find the cheapest flight with the shortest duration of travel that arrives at specified destination before 5 pm, a flight would be considered optimal if, and only if, it gets to the destination at the specified time and there are no cheaper or faster alternatives. For example, in the case of flights presented in Table 3.2, flight d is considered the optimal option because there is no other flight that arrives in Dublin before 5pm that is cheaper and faster.

Table 3.2: Available Flight Options. Bold indicates the optimal choice.

Flight ID #	Destination	Price	Duration of Travel	Arrival Time	Day of Travel
a	Dublin	£200	2.5h	6pm	1st Sep
b	Dublin	£150	3.5h	4pm	1st Sep
c	Dublin	£150	2h	6pm	1st Sep
d	**Dublin**	**£150**	**2.5h**	**4pm**	**2nd Sep**

In search scenarios provided to participants flight options were spread across several days (2 days for Study 1, and 3 days for Studies 2 and 3). Figure 3.6 presents the distribution of flight options in the search space. For all search scenarios featured in studies 2 and 3, prices and duration of flight options varied depending on the day of travel: day 1 consisted mostly of sub-optimal options (i.e. expensive flights with long duration of travel), day 2 contained mostly options that were short in duration, but expensive, and day 3 contained options that were mostly cheap but had longer duration of travel.

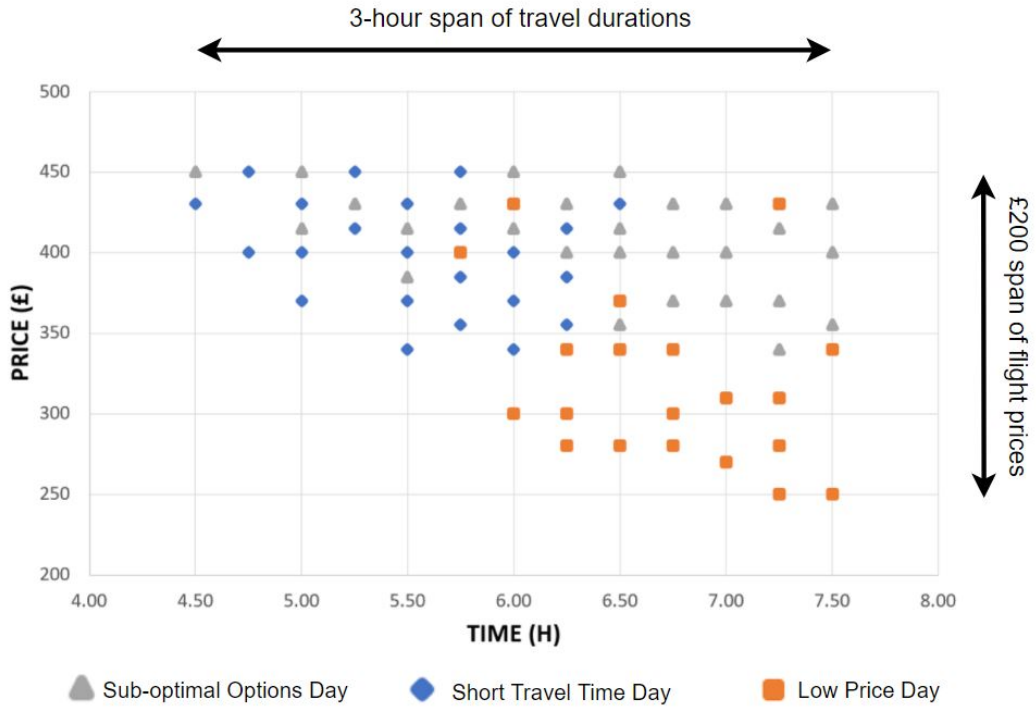


Figure 3.6: Comparison of available flight options across different travel days.

We decided to spread flight options across several days to check if participants were more prone to satisficing (i.e. the behaviour of selecting a minimally viable flight option without exploring available alternatives) [143]. Participants who checked all days were more likely to find an optimal flight option. In order to ensure comparability and consistency across all search scenarios, we ensured that for each scenario, the travel time of available flights spanned 3 hours, while their prices spanned £200

The concept of Pareto Optimality was used in all of the three studies to evaluate participants' performance. In Study 1 we considered it in terms of meeting the search criteria specified in search scenarios (primary performance indicator), while in studies 2 and 3 we also introduced additional metrics that reflected participants' expenditure of time and money (secondary indicators). Secondary indicators were calculated based on the difference between the optimal selection and the selection made by the participant. For instance, if a participant selected a

£200 flight with 3 hours travel time and the optimal flight cost £150 and took 2 hours, they wasted £50 and lost 1 hour of travel time.

3.4.2.2 Interaction Times

When evaluating user search experience we consider the overall task completion times and number of conversational turns that participants took during each interaction. These metrics have been used in previous research to indicate the performance of the dialogue system [119,167]. Since longer interaction can be associated with a higher cognitive cost [167], we assumed that the shorter task completion times would equate to a better performance of the agent.

3.4.3 Semi-Structured Interviews

In order to put the results obtained from subjective and objective measures into perspective, we conducted semi-structured interviews at the end of each study. The questions asked during the semi-structured interviews concerned participants' preferences regarding CAs, problems encountered during interaction and suggestions for future functionalities. The obtained feedback provided us with additional context and allowed us to triangulate our results.
Part of the novelty of our methodological approach lies in the triangulation of quantitative (standardised questionnaires and interaction metrics) and qualitative approaches (participants' feedback) - a combination that, to the best of our knowledge, has not been incorporated in previous work on the evaluation of conversational agents (a summary of relevant studies is provided in Table 2.3).

3.5 Chapter Summary

Chapter 3 provided information regarding the methodology used in the current thesis.
The chapter:

- Introduced the concept of interactive system evaluation from the perspective of Human-Computer Interaction and Information Retrieval. The evaluation methodology used in the current thesis draws from both these areas to account for both users' perceptions of conversational agents (subjective metrics) and their performance (objective metrics).
- Introduced the concept of a Wizard of Oz (WOZ) framework that allow for conversational agents to be evaluated via interactive simulation. A WOZ framework is used in all three studies featured in the current thesis.
- Provided a global overview of the experimental stages used in all of the studies featured in the current thesis.
- Discussed types of conversational agents featured in the studies based on the level of their conversational involvement.

- Provided a description of the interactive search tasks and the motivation for using them to evaluate conversational search agents.
- Introduced the concept of 'user search experience' and the metrics used for its evaluation and explained why the proposed combination of metrics is novel and meaningful.

In Part I of the thesis we have provided a theoretical underpinning for our investigation of conversational agents and discussed the methodology used to evaluate them. In Part II we will present 3 WOZ studies that will explore the role of agent's memory (Chapter 4) and the degree of conversational initiative when eliciting and presenting information - exploring the impact of active (Chapter 5), and proactive support (Chapter 6).

PART II:

Interactive Investigation of Conversational Agents

Chapter 4

Impact of Conversational Agent's Memory on User Search Experience

This chapter is an extension of the 'Investigating How Conversational Search Agents Affect User's Behaviour, Performance and Search Experience' research paper [52]. As explained in Chapter 2, interpreting human language is challenging as it requires interlocutors not only to understand the semantic but also the pragmatic meaning of any utterance (see Section 2.4). In the information seeking context, in order to communicate effectively, the information seeker and intermediary must keep track of information expressed in conversation and actively act on it. Since information is frequently expressed over many turns, interlocutors need to be able to keep track of what is being said and constantly update their understanding in order to take appropriate actions. The Theory of Conversational Search (discussed in Section 2.5.1) posits that memory and mixed-initiative are two crucial elements of a conversational agent. In the current chapter we will focus mostly on the first of these two elements - conversational memory. Mixed-initiative will be addressed in Chapters 5 and 6.

We address our first research question, **RQ1: ' How does a stateful conversational agent (with a memory component) differ from a stateless conversational agent with regards to: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?** For the purpose of comparison we introduce two agents: (1) (**CSA**), an agent with memory that preserves the context of dialogue and maintains state; and a voice search system (**VSS**) that does not. Based on [34, 95] we hypothesise that **CSA** will:

H1.1 : be less cognitively taxing - hypothesis linked to RQ1 (a).

H1.2 : lead to higher satisfaction - hypothesis linked to RQ1 (b).

H1.3 : lead to better performance - hypothesis linked to RQ1 (c).

H1.4 : lead to faster task completion - hypothesis linked to RQ1 (d).

To examine our hypotheses we conducted a Wizard of Oz (WOZ) study [38] in which we simulate **CSA** and **VSS**. **VSS** uses a slot-filling architecture, which requires users to provide information to fill in slots that are assigned with semantic class labels [116], simulating current state of the art approaches used in dialogue systems [79]. While the **CSA** has context awareness and remembers past interactions that allow user to provide information in a free-form natural language over several turns, the **VSS** requires user intent to be provided upfront in a pre-defined way (i.e. providing slots as prompted by the system). Participants were asked to complete simulated search tasks using both the **CSA** and the **VSS**. Performance, and perceived usability were measured using conversation logs as well as standard usability [24] and task load questionnaires [69]. Information on all metrics used to measure user search experience is provided in Section 3.4.

This chapter is structured as follows. We begin with background information about the architecture of systems that support voice search (Section 4.1). Then, in Section 4.2 we discuss related work focusing on the challenges of conversational search. In Section 4.3, we provide information on our methodology, including experimental procedure and information about our participants. We then present experimental results (Section 4.4) and discuss their implications (Section 4.5). Finally, we provide our conclusions (Section 4.6) and summarise the current chapter (Section 4.7).

4.1 Importance

Systems that support voice search (illustrated in Figure 4.1) are becoming increasingly more popular and widespread [1]. However, due to technical limitations, current voice based systems often lack the capacity to maintain a conversation. Examples of some of the most popular systems that support voice search include: Microsoft Cortana, Google Now and Apple Siri. These state of the art systems provide limited: (*1*) state tracking (i.e. estimating a user's goal in a conversation), (*2*) anaphora resolution (i.e. resolving references to earlier items in conversation), and (*3*) have problems with clarifying user intent. All of these issues currently make interaction through voice both an unnatural and tedious activity in many cases (cf. [36, 98, 121, 127]). However, Cook et al. [34] and Lison and Meena [95] suggest that voice based interaction could be efficiently used for a range of well defined information retrieval tasks, if conversation with a system is sufficiently natural.

[1] For Example see https://voicebot.ai/2019/04/15/smart-speaker-installed-base-to-surpass-200-million-in-2019-grow-to-500-million-in-2023-canalys/ (last accessed: 21st October 2020)

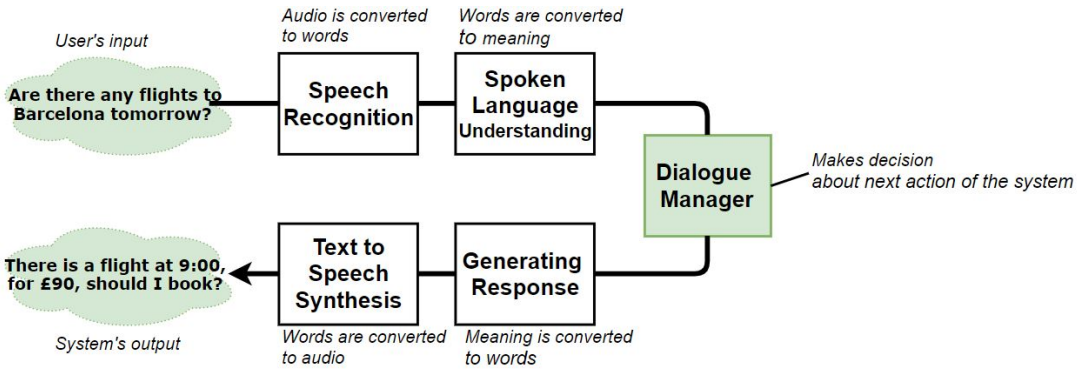


Figure 4.1: Architecture of a Voice Search System.

In the current chapter we will compare two simulated systems: (1) a voice search system that does not maintain state and requires the search query to be provided in one conversational turn in order to be processed (VSS) and (2) a stateful conversational agent (CSA). A more detailed explanation of both systems is provided in Section 4.3.4. Examples of participants' interactions with each system are provided in Section 4.3.6.

4.2 Context

Given the recent improvements in Automatic Speech Recognition (ASR) which is considered to have reached a 'human-like' performance [180] - and Text-to-Speech (TTS) which is almost indistinguishable from human speech [88, 170] - the focus of research in the field of artificial intelligence has switched to Natural Language Understanding (NLU) and Dialogue Management (DM) components of voice search systems. However, regardless of the technological improvements, voice search systems still struggle with tasks that require multiple conversational turns which leads to user dissatisfaction and infrequent use (cf. [36, 85, 86, 87, 98]). The challenges of voice interaction have recently captured the attention of the IR community and have led to discussion on the nature of human-like conversational search and issues that need to be addressed before robust conversational search agents can be developed cf. [76, 77, 160].

4.2.1 Relevant Research

4.2.1.1 Theory of Conversational Search

As explained in Chapter 2 (2.5.1) Radlinski and Craswell [126] defined a conversational search agent as a system for retrieving information, in which there is a mixed initiative between the user and the agent, and the agent's actions are selected in response to the user's needs within the context of the conversation (considering short and long term knowledge). Radlinski and Craswell (ibid.) further outlined five properties a conversational search agent should possess, highlighting that such a system needs to 'have memory to maintain state'. Given a lack of

memory of past interactions and an inability to maintain conversational state can be detrimental to user search experience (cf. [127]), in this chapter we consider the importance of these properties within a conversational agent.

Trippas et al. [155] proposed a search framework for a spoken conversational search system and highlighted the cognitive challenges that a user faces when dealing with such an agent. In more recent work, Trippas et al. (ibid.) expand their framework by analysing the behaviour of pairs of human participants engaging in collaborative search [124]. In contrast to that work, our study aims to analyse behaviour when interacting with a simulated search system. In our study, via use of interactive search experiment, we explore how the workload associated with using the different systems impacts on task completion rates and participants' perception of system usability.

4.2.1.2 Interactive Conversational Search Studies:

Another line of research on conversational search addresses the problem of designing an efficient and user-friendly search system. Lee et al. [91] focused on a system's personality and the discoverability of its affordances. In a Wizard of Oz study, Lee et al. (ibid.) evaluated an early prototype of a conversational search system. Lee's experiment involved two participants (one who simulated the search system and the other that conducted the search). The focus of the study was on understanding how to ascertain the participants' needs, so that they can be applied in the development of such systems. Similarly, Trippas et al. [157] carried out a study aimed at investigating mixed initiative conversational behaviour for an information search in an acoustic setting (e.g. voice only). The study involved 13 participants who completed a series of tasks with different levels of complexity. Trippas et al. found that an increase in task complexity was linked to more queries being issued by participants.

Vtyurina et al. [165] explored what a search experience would look like if a truly conversational system existed. In Vtyurina et al.'s study, 21 participants who used a text-based interface were asked to complete 3 different search scenarios; one scenario carried out with an existing commercial system, one with a human 'expert', and third with a 'wizard' simulating the system. The results indicated that participants did not mind using a voice based search system as long as it was usable.

In the research presented in this chapter, in a lab study, 22 participants completed a series of search tasks by interacting with two simulated search systems. To make the simulation more realistic, feedback is provided via synthetic voice (the details of the experimental setup are provided in Section 4.3.3).

4.2.1.3 Conversational System Design

Thomas et al. [149] explored the role of style in conversation seeking tasks. Their study, based on the analysis of recordings of strangers interacting with one another on assigned tasks,

shows that people tend to align their conversational style with intermediaries. In this study, we focus on the impact of different interaction styles i.e. free-form dialogue (offered by the conversational agent) and constrained interaction (offered by the slot-based voice system) on participants' behaviour.

Schulz et al. [139] proposed a state tracking model which enables a user to compare different results during a conversational search. The proposed model assigned dialogue acts of new user utterances to the frames created during the dialogue - with each frame corresponding to an individual goal. The attempt was described as the first step to creating a memory-enhanced search system that can understand when users refer to previous topics of conversation, and providing a user with accurate feedback by understanding the context of their request. Related to the above research is the 'Frames' corpus that was created to study the role of memory in voice search tasks [7]. Our study expands upon this prior work by employing a simulated conversational search agent, where we evaluate how natural dialogue (in which memory of the conversation is used and the state of the need is preserved) affects user search behaviour.

In this chapter, we evaluate a system that presents participants with results in a free-form natural-language dialogue and measure to what extent its performance differs from the slot-based voice system.

4.3 Method

In order to explore how user search experience changes when interacting with a conversational search agent (CSA) when compared to a voice based search system (VSS), we conducted a WOZ study (See section 3.2 for an explanation of a WOZ framework) where participants undertook four simulated search tasks. The current section provides information on the experimental setup, search agents, search scenarios and interaction design.

4.3.1 Study Design

Our study followed a 1x4 within-subjects design, where the independent variable was the type of the system (VSS vs. CSA) and dependent variables were four aspects of user search experience: (1) cognitive load (measured by NASA TLX [69], (2) satisfaction (measured by SUS questionnaire [24] and sentiment towards the system indicated by a sentiment analysis tool VADER [73]), (3) performance (evaluated in terms of the selected flight meeting the search criteria outlined in the task) and (4) interaction time for each of the search scenarios (measured in seconds).

4.3.2 Experimental Procedure

The experimental setup is diagrammatically presented in Figure 4.2.

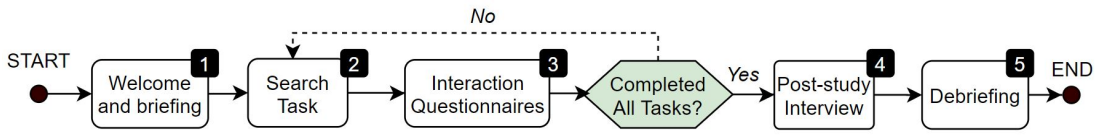


Figure 4.2: Illustration of Experimental Stages. Stage (2) consists of two search sessions (one with VSS and one with CSA); each session consisted of two search tasks and was followed by NASA TLX [69] and SUS [24] questionnaires.

Participants completed two search tasks per system. To reduce any priming effects, a Latin Square design [94] was applied to rotate the order in which search tasks and systems were presented to the participants. The experiment was conducted in a controlled lab environment (a quiet office, with a comfortable chair and desk). Each system was simulated using a mock-up setup, where the experimenter controlled the responses of the search systems.
Before the experiment, participants were briefed about the structure and purpose of the study, and provided informed consent to participate. Ethics approval for the experiment was granted by the Department of Computer and Information Sciences, University of Strathclyde (application no. 611). During the experiment, participants were instructed to complete four search tasks (described in more detail in Section 4.3.5). Participants were informed that neither the **VSS** nor the **CSA** would provide any visual feedback and that the interaction would be exclusively voice-based. The wizard (the experimenter in control of the systems) was using a computer to initiate the search tasks and switch between the systems once tasks were completed. Participants were informed that their interactions with the system would be audio recorded.

4.3.3 Wizard/Agent Setup

A Wizard of Oz (WOZ) [38] setup was used to compare and contrast the stateful conversational agent and the stateless system with no conversational memory. Similar setups have been used in several related studies (e.g. [41, 178]), and are often used when evaluating natural language based interfaces (cf. [103]). Using a WOZ setup allowed us to test our four research hypotheses in a controlled lab environment.

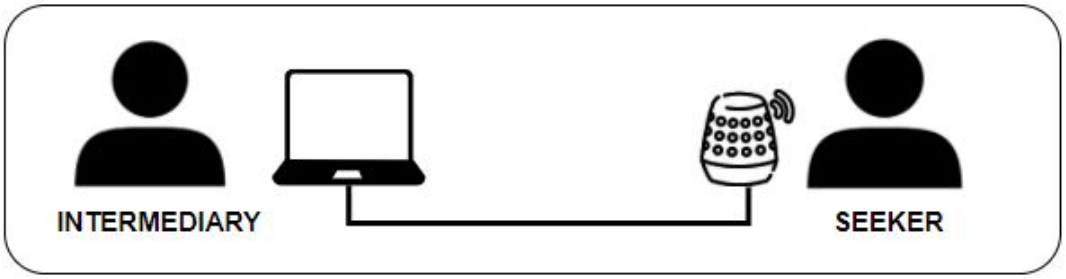


Figure 4.3: The Wizard of Oz framework: A wizard (intermediary) searches for a flight on behalf of a user (seeker) and provides them with results. The result is presented in synthetic speech through a stand-alone speaker.

The experimental setup is presented in Figure 4.3. The experiment took part in an office room at the University of Strathclyde. The layout of the office room is illustrated in Figure 4.4. In the study participants were acting as information seekers while the wizard (the lead researcher) was the intermediary. The wizard waited for a participant's query, looked-up a pre-defined pool of flights (presented in Table 4.1) and provided results back to the participant using a GUI console (illustrated in Figure 4.5). The console included: (1) a Java-script application with pre-recorded prompts and (2) a live-speech synthesis tool provided by Cereproc Ltd. [10] to deal with any unexpected participant responses. The prompts used to simulate both the **VSS** and the **CSA**, were prepared in accordance with the guidelines outlined in [31].

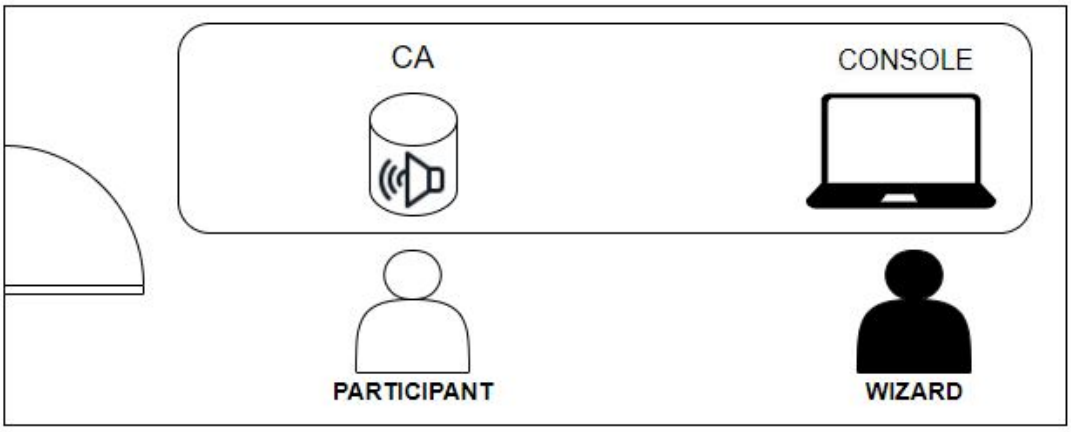


Figure 4.4: Layout of the office room (not to scale): A wizard was using a prompt console that played audio prompts via a stand-alone speaker (CA). Note: No picture was taken when the study was conducted and access to the office is no longer available due to the Covid pandemic.

Table 4.1: Overview of available flight options for each scenario. Bold indicates the optimal choice that meets the specified search criteria.

Search Scenario	Date of Travel	Departure Time, Cost
VSS1	11th November	11am, £110; 1pm, £90
	12th November	**9am, £95**; 10am, £110
VSS2	6th December	7am, £120; 2pm, £90
	7th December	**6am, £80**; 8am, £90
CSA1	4th November	6am, 120; **8am, £80**
	5th November	7am, £95; 9am, £90
CSA2	2nd December	6am, £100; **8am, £80**
	3rd December	5am, £85; 7am, £100

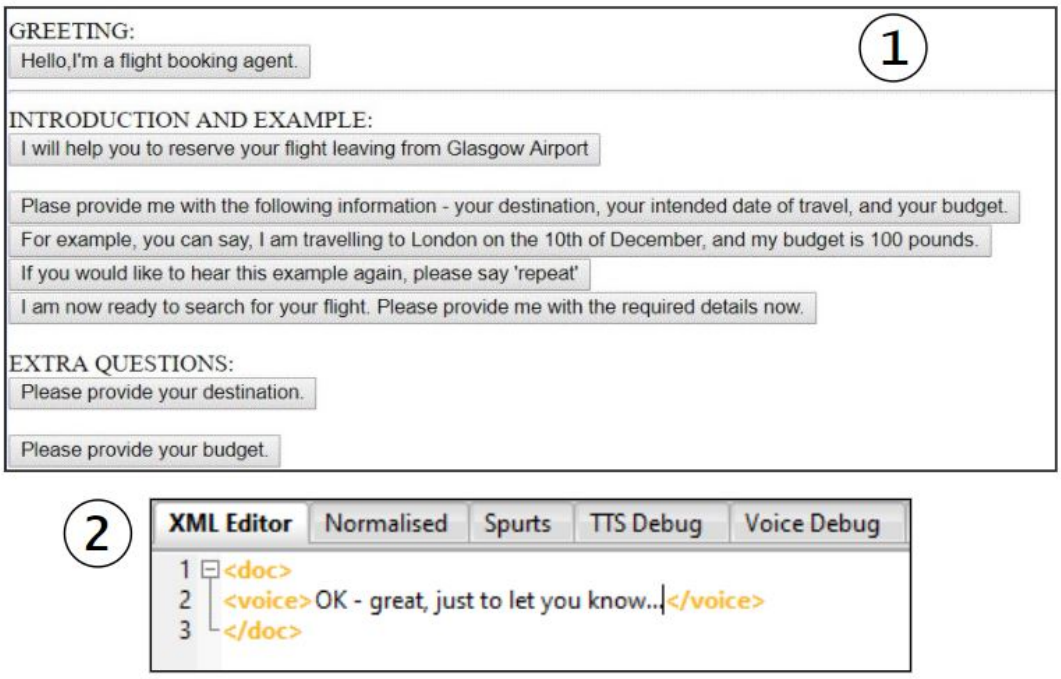


Figure 4.5: Wizard Console comprised of: (1) Java-script application with pre-recorded prompts and (2) Live-speech synthesis tool provided by Cereproc.

The prompts were made to resemble a natural spoken discourse by use of appropriate cohesive devices (pronouns and discourse markers), adhering to the principles of information structure (providing new information at the end of the utterance), and by applying Grice's 'Cooperative Principle' [67] (making assumptions about inferences that users will draw from the prompts).

4.3.4 Search Systems

Two search systems were used in the study. The voice search system (**VSS**) was based on a 'slot-filling' architecture and design recommendations outlined in [105]. At the beginning of each interaction, the **VSS** provided participants with a welcome message and presented them with available functionalities. Participants were asked to provide their search criteria, namely

'destination airport', 'date of travel', and 'available budget'. (Note: the departure airport was always fixed). The **VSS** also provided an example to help participants formulate their query: 'For example, you can say I'm travelling to London on the 2nd of December and my budget is 100 pounds'. The conversational search agent **CSA** started with a brief greeting: 'Hello, how can I help you?' after which participant was supposed to provide their search query. It should be noted that for comparative purposes, interaction time was measured from the moment that a participant started to speak after they were greeted by the system. Interaction with the **CSA** was not constrained in any way and participants were free to provide information in any order, whereupon the system would ask them follow up questions to clarify their intent. The **VSS**, however, could only process a query once all of the requested information was provided.

4.3.5 Simulated Search Tasks

Participants were asked to complete four search search scenarios - two for the **VSS** and another two for the **CSA**. The simulated search tasks were based on the evaluation model proposed by Borlund [22]. The model relies on using plausible, real-life, search scenarios that participants can easily relate to. In our study, the task for each of the search scenarios was based on booking a one-way flight to a travel destination in Europe. The place of departure was always fixed. In order to provide participants with a search intent, participants were given each search scenario which contained four pieces of key information, namely: (*1*) place of departure, (*2*) date of departure, (*3*) budget and (*4*) preferences. An example search scenario is shown in Box 4.1

You are planning to visit your friend who lives in **Bristol**. You will be flying from Glasgow Airport. Your total budget is **100 pounds**. You can travel either on the **11th or 12th of November**. **Indicative Request:** You want to find the cheapest possible deal but your flight needs to leave **on, or before 11am**. **Note:** Please wait for the agent to finish before you start to speak.

Search Scenario 4.1: Travel destination, budget, possible days of travel, and preferred time of arrival are provided in bold for participants' convenience.

In each scenario, participants were asked to find the cheapest flight with a specified time of arrival (indicative request). Providing such specific search criteria allowed us to assess participants' performance based on the flight that they booked and the time that they took to complete the task. Participants were instructed not to interrupt the system, and speak only when the system finished its turn. Every search task began with a pre-recorded prompt that welcomed participants and asked them to submit their search query. The task was considered completed once a participant had found and reserved a flight.

4.3.6 Interaction Design

Both **VSS** and **CSA** systems were operated by a Wizard (the lead researcher). The behaviour of the **VSS** system was designed to resemble a system that is based on a slot-filling architecture, which requires participants to provide the required information in a structured way so that their query can be processed. Unlike the **VSS**, the **CSA** was designed to imitate an unscripted, natural human-human conversation. The characteristics of natural human-human conversation used to simulate the **CSA** were based on features outlined in [18] i.e. 'ability to parse and non-fully-sentential grammar of spoken language', 'resolution of anaphora and ellipsis', 'keeping state of dialogue', 'ability to resolve hesitations, false starts and repairs' and 'ability to infer information from conversational context'. We also implemented the use of relative questions to clarify participants' intent (e.g. 'Do you have a budget in mind?') This is similar to the approach adopted in [27].

An example interaction between a participant and the **VSS** is provided below. The system begins by introducing itself to the participant and informs them how to use it.

VSS: [***INTRODUCTION***] Hello, I am a flight-booking agent, I will help you to reserve your flight leaving from Glasgow Airport. Please provide me with the following information: your destination, your intended date of travel and your budget. For example, you can say: I am travelling to London on the tenth of December and my budget is one hundred pounds. If you would like to hear this information again, please say repeat [PAUSE]. Please provide me with the required details now.

USER: I am travelling to Manchester on the seventh of December, my total budget is one hundred and twenty pounds.

VSS: There are two flights on the seventh of December that match your search criteria...

(Interaction with VSS)

Unlike the **VSS**, the **CSA** does not follow any strict interaction script, instead, it initiates the conversation with a question.

CSA: Hello, how can I help you?

USER: I would like to reserve a flight.

CSA: Ok, where would you like to fly to?

USER: Santiago

CSA: Sure, when would you like to travel?

USER: Either on the second or the third of December.

CSA: Do you have a budget in mind? (Interaction with CSA)

4.3.7 Logs

Participants' interactions with each of the systems (i.e. **VSS** and **CSA**) were recorded, and then analysed to extract a series of objective measures including: task completion time, task completion success (i.e. whether a flight was booked meeting all the specified criteria) and vocabulary used by participants (to evaluate participants' sentiment towards the system).
Table 4.2 provides an overview of descriptive statistics for participants' conversational behaviour when interacting with each of the systems.

Table 4.2: Dialogue statistics for both systems.

[Metrics] \ *[Agent Type]*			*Voice Search System No Memory*	*Conversational Search Agent With Memory*
Turns	First Task	Total	84	95
	Second Task		71	85
	Overall		**155**	**180**
Words per Turn	First Task	Mean	19.71	11.23
	Second Task		17.87	9.64
	Overall		**18.87**	**10.48**
Words per User	First Task		78.85	50.8
	Second Task		60.43	39
	Overall		**139.29**	**89.86**

We observed that participants interactions with the **VSS** were less dynamic than with the **CSA**. In total, we recorded 155 turns (i.e. pairs of conversational exchanges between the system and participant) for the **VSS** and 180 turns for the **CSA**. On average, participants spoke less when interacting with the **CSA** than with the **VSS**; we recorded 18.87 words per turn for the former and 10.48 per turn for the latter.

4.3.8 Sentiment Analysis

The VADER [73] sentiment analysis module of the Natural Language Tool Kit (NLTK) python library [97] was used to analyse participants' utterances. VADER uses a sentiment lexicon that contains a list of 7500 words, rated by human annotators on the scale ranging from -4 (extremely negative) to 4 (extremely positive). All words with the score above zero are considered positive, while all words with the score below zero are considered negative. For example, the word 'fantastic' has a score of 2.6, whereas 'disaster' is scored at -3.1. Words that are not included in the lexicon are considered neutral.
Through lexical analysis, VADER assigns positive, neutral and negative scores as ratios of text that fall into each sentiment category. For instance, the sentence 'He is smart, handsome and funny' is classified as 75% positive and 25% neutral. The sentence contains three positive words: smart, handsome and funny each of which has an individual score above zero.
In our analysis, we use positive, neutral and negative scores to provide multidimensional measures of sentiment for participants' utterances. Table 4.3 provides examples of words that fall

into 'positive' and 'negative' sentiment categories. For instance, examples ID 1 and ID 3 contain the words 'thank you' and 'thank you very much' that fall into the positive category. Examples ID 2 and 4, on the other hand, contain words 'f**k off' and 's**t' that fall into the negative category. A full breakdown of sentiment scores registered during the experiment is provided in Table 4.5.

4.3.9 Participants

Participants were recruited via social media, university mailing lists and flyers posted at a campus of two major UK universities (i.e. University of Strathclyde and University of Glasgow). The study inclusion criteria were that participants were at least 18 years old, native English speakers and had no hearing impairment. We decided to focus on native English speakers to limit the potential impact of language competence (native/non-native) on participants' ability to understand the synthetic voice with a local accent [1] that was used in the study. In total, 22 participants took part in the study. There were 9 females and 13 males. The age of participants ranged from 18 to 65 years (M = 28.67, SD = 9.9, Med = 28). The majority of participants (N = 15/22) had used a voice search system at least once before. On average participants who reported using voice search had been doing so for 2 years (SD= 1.5). All participants were given a 10 GBP shopping voucher for taking part in the experiment. The average completion time of the experiment was 35 minutes.

[1] The synthetic voice Heather was provided by Cereproc Ltd. http://cereproc.com

Table 4.3: Sentiment Categories. Examples of words that fall into 'positive' or 'negative sentiment categories are provided in bold.

ID	Agent Type	Sentiment Category	Example Interaction
1.	CSA	Positive	Participant: My total budget is one hundred pounds, but I want to find the best possible deal. CSA: One moment please, the best option on the second December is at eight am, it costs eighty pounds. Should I book? Participant: Yes, **thank you**.
2.	VSS	Negative	Participant: I want to travel to Manchester on either the sixth or the seventh of December. It must be before three pm and it cannot cost more than one hundred and twenty pounds. Please find me flights. VSS: Sorry, I can only search one day at a time. Participant: Ah, **f**k off...** Ok, Manchester, on the sixth December. It must be before three pm and cost one hundred twenty pounds or less.
3.	CSA	Positive	CSA: Hello, how can I help you? Participant: Oh, hello, I would like to fly to Cardiff on a one way ticket, travelling either on the second or the third of December. My flight must leave before one pm, and I cannot spend more than a hundred pounds. Could you please tell me the best option I can get? CSA: The best option is on the second of December. It is at eight am and costs eighty pounds. Should I book? Participant: Yes, that sounds fine, **thank you very much.**
4.	VSS	Negative	Participant: I would like to travel to Bristol on the eleventh November before eleven a m and my budget is a hundred and ten pounds. VSS: There are two options to Bristol that match your search criteria. One at eleven am for one hundred pounds and another one at one pm for ninety pounds. To make a reservation, please specify the option that you would like to book. To search again, please provide your new search criteria. If you would like to hear this information again, say repeat now. Participant: I am travelling to Bristol and on the twelfth of November and before eleven am, and flight, **oh s**t**, never mind. I am travelling to Bristol on the twelfth of November before eleven am, and my budget is hundred and ten pounds.

4.4 Results

When reporting results, unless stated otherwise, we used the Wilcoxon Signed-Rank test as most of our data was not normally distributed. All of the conducted statistical tests were two-tailed. The object of our comparison are the two systems **VSS** and **CSA**. Where appropriate we also report some differences between tasks within the system.

4.4.1 Subjective Metrics - Cognitive Workload and User Satisfaction

To examine whether both systems differed in terms of perceived workload and satisfaction, we compared the results of the NASA TLX and SUS questionnaires (summarised in Table 4.4) and the sentiment analysis of participant interaction logs (summarised in Table 4.5). The results are discussed in turn in the following subsections.

Table 4.4: Subjective Measures. Note: For NASA TLX the lower the score, the less cognitively taxing the system is perceived to be, for SUS the higher the score the more usable the system is perceived to be.'*' indicates $p < .05$, '**' indicates $p < .01$.

[Metrics] \ *[Agent Type]*			*Voice Search System (VSS) No Memory*	*Conversational Search Agent (CSA) With Memory*
NASA TLX Overall(0-100), Per Item(0-20)	Mental Demand	Med (IQR)	27.5 (32.5)	10 (21.25)**
	Effort		30 (37.5)	12.5 (15)**
	Performance		15 (25)	5 (10)*
	Frustration		22.5 (26.25)	10 (13.75)*
	Temporal Demand		17.5 (11.25)	12.5 (21.25)
	Overall		**29.5 (18)**	**14 (12.5)****
SUS (0-100)	**Overall**	Med (IQR)	**81.25 (27)**	**92.5 (17)****

4.4.1.1 Cognitive Workload

The results show that the **CSA** led to lower cognitive workload than the **VSS** for all five aspects under investigation (i.e. 'mental demand', 'temporal demand', 'performance', 'effort' and 'frustration'). Overall, every participant found the **CSA** system less taxing to use than the **VSS** system. The median NASA TLX score for the baseline system, **VSS**, was 29.5 as compared with 14 for the conversational agent **CSA**. Note, the lower the score, the less demanding in terms of workload the system is. There was a statistically significant difference ($p < .001$) in the overall perception of both systems. On the level of individual dimensions, significant differences were found for 'mental demand' ($p = .0002$), 'performance' ($p = .025$), the 'effort' required ($p = .00032$), and 'frustration' ($p = .037$).

It appears that participants using the **VSS** had to more closely monitor the interaction and keep track of their state and the state of the system. This made it increasingly difficult for participants to make direct comparisons between search results i.e. comparing search results by changing a given search aspect (e.g. 'Show me flights on the next day', 'Show me some cheaper flights' etc.). On the other hand, **CSA** allowed participants to use search commands such as 'find the cheapest option' or 'I want a flight like this but it needs to leave earlier'. This 'memory' feature of the **CSA** significantly reduced the number of items that participants had to memorise, resulting, in turn, in lower mental demand and effort.

4.4.1.2 User Satisfaction

System Usability: Table 4.4 reports the SUS scores for each system, where for the **VSS** the median was 81.25 whereas for the **CSA** the median was 92.5. This difference was statistically significant ($p = .003$). Note, the higher the score the more usable the system. The score achieved by the CSA corresponds to approximately the top 5th percentile of SUS scores [1], whereas the VSS falls into the top 30th - 25th percentile bracket.

Sentiment Towards the System: Table 4.5 presents information regarding participants' attitudes towards the systems **VSS** and **CSA**. Sentiment analysis conducted on transcripts of participants' interactions with the systems, indicates that, on average, participants displayed more positive sentiment towards the **CSA**. There is a statistically significant difference in positive sentiment between each of the systems both on the level of individual tasks, i.e. VSS1 vs. CSA1, ($p = .001$) VSS2 and CSA2 ($p = .0001$) as well as overall ($p < .0001$). For the above comparisons, the Bonferroni adjusted *alpha* was .0083.

[1] based on score interpretation guidelines provided by Sauro [135].

Table 4.5: Participants' Sentiment Towards the System. Note: Sentiment scores are ratios for proportions of participant utterances that fall into a particular sentiment category, i.e. 'positive','negative' or 'neutral'. All of the categories sum up to 1. '**' indicates p < .01.

[Metrics] \ *[Agent Type]*			*Voice Search System No Memory*	*Conversational Search Agent With Memory*
Expressed Negative Sentiment	First Task	M (Med)	1.08(%) (0%)	6.34% (0%)
	Second Task		1.12% (0%)	3.28% (0%)
	Overall		**1.1% (0%)**	**4.77% (0%)**
Expressed Neutral Sentiment	First Task		89.21% (89.73%)	67.16% (76.1%)
	Second Task		90.2% (91.15%)	65.53% (67.55%)
	Overall		**89.71% (89.13%)**	**66.33% (76.1%)**
Expressed Positive Sentiment	First Task		9.7% (10.27%)	26.5% (23.9%)**
	Second Task		8.67% (8.85%)	31.19% (32.45%)**
	Overall		**9.19% (10.27%)**	**28.9% (23.89%)****

4.4.2 Objective Metrics - Participants' Performance and Task Completion Time

Objective metrics used to evaluate both search systems were: participants' performance considered in terms of selected flight options, and time taken to finish the task. Both metrics are described in turn in this section.

4.4.2.1 Task Performance

Table 4.6 provides information on optimal flights chosen by the participants. As can be seen, while just half of participants interacting with the **VSS** managed to find the optimal flight, the figure was significantly higher for the CSA ($p < 0.001$).

Table 4.6: Finding the Optimal Flight. '***' signifies $p < 0.001$.

[Metrics] \ *[Agent Type]*			*Voice Search System No Memory*	*Conversational Search Agent With Memory*
Chose Optimal Flight	First Task	#	13/21 (48%)	20/21 (95%)
	Second Task		9/21 (43%)	18/21 (86%)
	Overall		**22/42 (52%)**	**38/42 (90%)****

4.4.2.2 Task Completion Time

Task Completion Times are presented in Table 4.7. For the **VSS** the task completion time was 215 seconds on average, while for the **CSA** the task completion time was approximately 117 seconds. The average task completion time was found to be significantly different for both systems (p = .00012). Regarding the change in completion times between the first (T1) and the second task (T2) - for the **VSS**, there was a significant drop in the time taken to complete T2 from 136 to 117 seconds on average (p = .012), while for the **CSA** the time to complete fell from 64 to 52 seconds (p=.046).

Table 4.7: Task Completion Times. '***' signifies p < 0.001 Note: Figures are rounded up to the nearest second.

[Metrics] \ [*Agent Type*]			*Voice Search System No Memory*	*Conversational Search Agent With Memory*
Interaction Time in sec.	First Task	Mean (SD)	136 (26)	64 (30)
	Second Task		116 (28)	52 (27)
	Overall		**215 (50)**	**117 (50)*** **

As presented in Figure 4.6, participants took longer to familiarise themselves with the baseline system, **VSS**, and to explore its functions. The conversational system, **CSA**, however, was more straightforward to use and did not curtail participants' interaction. The **CSA** also outperformed the **VSS** in terms of average task accuracy.

4.4.3 Post-study Interview

After completing the simulated search tasks, participants were asked to comment on their experience and indicate which system they preferred. The majority of participants (18/22) indicated the **CSA** as their preferred system. Justifying their choice of the **CSA**, participants mentioned that: it was natural to use, helped them to accomplish their tasks easily, and to clarify their intent easily. For example, **P4** said: *'It required less listening and it did not speak that much so there was not that much information for me to remember.',* while **P10** commented: *'[It was] much more natural. As a result of that it just feels less stressful. You can achieve what you need to do in less time compared to the second system [**VSS**].'* , and **P15** pointed out that: *'It [**CSA**] was much more easy to use, I found it much more intuitive. I did not have to remember any details to get the best result'.*

Participants who chose the **VSS** said that they preferred the 'command-and-control' style of interaction offered by that system and found it more predicable to to use. For instance, **P12** commented : *'I liked how it [**VSS**] was predictable. When I learned how to ask a question. I knew exactly what to do. Whereas, with the second one [**CSA**] I was worried that it would not understand what I was asking.'*

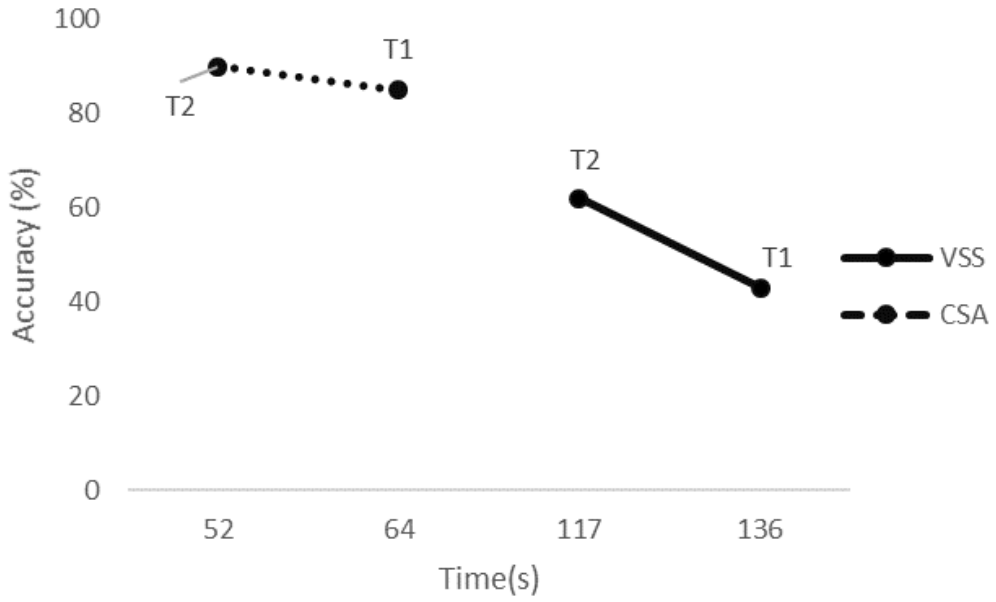


Figure 4.6: Task completion times and booking accuracy for VSS and CSA for first (T1) and second attempt (T2).

4.5 Discussion

Based on the experimental results, we can make the following observations. Firstly, with respect to our subjective metrics: we observe (as discussed in Section 4.4.1) that the CSA was less cognitively taxing than the VSS (**support for H1.1**) and led to higher satisfaction and more positive sentiment (**support for H1.2**). Secondly, as for our objective metrics (presented in Section 4.4.2), we have seen that participants performed better when using the CSA and selected a higher proportion of optimal flights that met the search criteria (**support for H1.3**). They also managed to complete their search tasks quicker than when using the VSS (**support for H1.4**). The positive impact of the CSA is evidenced in all applied metrics: i.e. questionnaires, semi-structured interviews and sentiment analysis. Overall, with regards to our **RQ1: 'How does a stateful conversational agent (with a memory component) differ from a stateless conversational agent with regards to: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?'** - we can see that using the CSA leads to a better user experience across all the four aspects. Using a CSA is less taxing, leads to higher satisfaction and allows participants to complete their tasks quicker while choosing better flight options.

We observed that participants could improve their performance with the VSS with more practice (see Figure 4.6). However, adaptability came at the price of accuracy - participants' performance using the VSS is inferior (slower and less accurate) than when interacting with a conversational

agent which allows them to perform search tasks quicker and more accurately. To examine these trade-offs and differences in behaviour, longitudinal studies focusing on prolonged exposure to both the VSS and the CSA styles of interaction are required.

Another observation is that while interacting with the CSA our participants used more positive language and displayed a more courteous attitude towards the agent then when speaking to the baseline system. We observed that while interacting with the CSA, participants thanked the agent more frequently and used more polite language (see Section 4.4.1.2 for the results of our sentiment analysis). This may indicate that a more natural conversational style (free-form language) encouraged our participants to approach the agent in a more personalised, human-like way. On the other hand, the rigid interaction style of VSS, made participants frustrated as they had to adjust their speaking style and phrase their requests according to a pre-defined template.

In our experiment we focused mainly on analysing participants' language at the level of syntax (types of words used). However, it would be interesting to see if the aforementioned positive attitude towards the agent is also reflected in changes in voice quality, i.e. in tone and pace. Such cues could be used as an implicit means of measuring performance and satisfaction with the system.

4.5.1 Reflections on Findings

Based on the experience of observed conversations, we would like to offer several reflections on the potential reasons for our experimental findings.

Firstly, we observed that several participants found the requirement to formulate their queries to match the query format required by VSS, irritating. As illustrated in Table 4.3, participants were frustrated by breakdowns in communication that occurred during interactions with the VSS (example ID 2) and the need to reformulate their search requests (example ID 4). Consequently, these challenges increased the time of the interaction, imposed an additional cognitive load on participants (reflected by higher NASA TLX scores) and, effectively, prevented participants from engaging in a deeper exploration of the search space (reflected by the lower accuracy). On the other hand, we observed that interactions with the CSA were more dynamic (more frequent and shorter turns) and led to better accuracy and performance of participants.

Secondly, while the majority of participants were satisfied with the natural and unconstrained style of interaction offered by the CSA, others found it unpredictable (see P12's comments in Section 4.4.3). This concern can be linked to the notion of discoverability, i.e. 'the ability of users to find and execute features via user interface' [182]. Discoverability is a common problem for screen-less devices (including smart speakers) that affects new users who need to learn the capabilities of the device through trial-and-error [84]. In the case of our experiment, some participants were unsure which queries were supported by the CSA, and therefore preferred the guided approach to interaction offered by the VSS. As indicated by previous research, a

user's perceptions of a CA's capabilities may be affected by their previous interactions with voice interfaces [36, 98] and assumptions of its human-like reasoning abilities [44]. One way of setting participants' expectations regarding the CSA would be for the agent to state explicitly at the beginning of interaction that it supports natural language interaction.

Thirdly, we observed that the CSA received high satisfaction scores (cf. Table 4.4) and enabled participants to complete the task quickly (cf. Table 4.7). However, it should be noted that the search space of available flight options for each scenario was relatively small and thus unlikely to reflect the real-life flight booking process. While this design choice was taken to make the interaction with CAs more dynamic and to reduce the number of synthetic speech prompts required, in a real-life situation participants would likely require more options and engage in longer interactions to make an informed choice. With this premise in mind, in the follow-up study (described in Chapter 5) we decided to expand the search space to make the task more realistic and reflective of real-life scenarios.

4.6 Conclusions

Our findings provide empirical evidence that suggests that, overall, conversational search systems are more user friendly and efficient than the current state of the art-systems based on slot-filling architecture. Our proposed conversational agent (CSA) offers an interaction experience that resembles human-human conversations: it is less constrained to use, leads to lower cognitive workload, encourages use of natural language, and incites positive sentiment. All these merits suggest, that, in the future, the development of conversational agents should focus on making them more responsive and less reliant on a fixed interaction protocol to ensure the best user experience when searching for information.

4.7 Chapter Summary

This chapter provided insights into the impact of memory on user search experience by conducting a comparison of state-full and stateless conversational agents. We have shown that an agent's ability to preserve conversational state is crucial to (1) reduce cognitive load, (2) increase satisfaction, (3) improve performance and (4) reduce the time required to complete the task. In the next chapter we will explore the impact of an agent's conversational initiative based on its involvement in eliciting user requirements and providing search results.

Chapter 5

Impact of Conversational Agent's Elicitation and Revealment Strategies on User Search Experience

This chapter is an extension of 'Conversational strategies: impact on search performance in a goal-oriented task' research paper [50].

In Chapter 4, we have demonstrated that an agent's memory (i.e. the ability to maintain conversational state) is fundamental to a positive user search experience. In the current chapter we will focus on 'mixed-initiative' which is the second of two crucial elements of a conversational search system (as discussed in Section 2.5.1). In particular, we will focus on how the degree of involvement of a conversational agent in eliciting and revealing information in the search process impacts on user search experience. Prior research posits that due to the non-persistent nature of speech, conversational agents (CAs) should support users in their search task by: (1) actively suggesting query reformulations [154], and (2) providing summaries of the available options [119]. Currently, however, the majority of CAs are passive (i.e. lack interaction initiative) and respond by providing lists of results – consequently putting more cognitive strain on users (cf. [35, 87, 98]).

In the current chapter, we address our second research question, **RQ2: 'How do different information elicitation (passive vs. active) and revealment (summary vs. listing) techniques impact on user search search experience in a goal-oriented task with regards to: (a) Cognitive load, (b) Satisfaction with the Agent, and (c) Task performance?'** For the purposes of comparison we introduce four agents that vary in their elicitation (active vs. passive) and revealment (summary vs. listing) techniques.

Based on the anticipated benefits of active search support (advocated, among others, by Radlinski and Craswell [126] and Trippas [154]) and a concise presentation of information (cf. Polifroni and Walter [119]), we hypothesise that compared to passive and listing agents, active and summary agents will:

H2.1 : be less cognitively taxing - hypothesis linked to RQ2 (a),

H2.2 : lead to higher satisfaction - hypothesis linked to RQ2 (b),

H2.3 : lead to better performance - hypothesis linked to RQ2 (c) and

H2.4 : lead to faster task completion - hypothesis linked to RQ2 (d).

To examine the above hypotheses and investigate the potential benefit of active search support, we conducted a second Wizard of Oz (WOZ) [38] study. Twenty-four participants undertook four goal-oriented search tasks interacting with four conversational agents (described in Section 5.3.4) that differed in the degree of their conversational involvement. Participants' user search experience was evaluated based on metrics outlined in Section 3.4.
This chapter is structured as follows. We begin by providing background information on the importance of role of the mixed-initiative in conversational systems (Section 5.1). Next, in Section 5.2 we discuss related work focusing on the challenges of conversational search, present the current state of the art for commercial applications of CAs, and propose our conversational strategies. In Section 5.3, we provide information on our methodology, including the experimental procedure and information about our participants. We then present our results (Section 5.4) and the findings of the semi-structured interviews (Section 5.5), and discuss their implications (Section 5.6). Finally, in Section 5.7 we provide conclusions before summarising the chapter in Section 5.8.

5.1 Importance

Conversational Agents (CAs) are systems that enable natural language interaction which is not constrained by menus, command prompts and key-words [115, p. xiv]. In theory, CAs should give a user a feeling of collaborating with a virtual companion rather than using a system [111]. In practice, however, due to the ambiguous nature of human language and the necessity for incremental interpretation of utterances in context, interaction with CAs can be a rather cumbersome and frustrating experience [52, 98]. Currently, most commercially available CAs (e.g. Google Home, Amazon Echo etc.) lack the capacity to ask meaningful follow-up questions to refine the user's intent and, consequently, struggle with tasks that involve interpretation, judgment or opinion. Due to these limitations, CAs have been mostly used for simple goal-oriented tasks that require structured conversation characterised by predictable user input and a small number of dialogue turns (cf. [5, 115, 173]). As such, most CAs tend

to be based on an inflexible, passive interaction strategy which may leave users uncertain about possible options and functionalities [113]. However, for a CA to be more usable, it needs to be able to correctly interpret a user's query, as well as present the information in ways that help the user achieve their goals [41]. With this premise in mind, rather than passively eliciting user information needs, and simply listing search results, this chapter focuses on evaluating alternative interaction strategies. Specifically, we aim to explore the influence of: (1) two elicitation strategies (Passive vs Active), (2) two revealment strategies (Listing vs Summarising) and (3) the combination of elicitation and revealment strategies on the user's search performance and their search experience.

5.2 Context

Over the past few years there has been a growing interest and resurgence in the design and development of conversational agents due to the maturation of speech and natural language processing technologies that facilitated their development. For example, with the recent advent of deep learning, we have seen a number of studies where neural ranking models are used to support interactive, multi-turn conversational search [2, 61, 123, 183]. In this chapter, however, rather than exploring the underlying mechanisms and automatic evaluation of conversational systems, we focus on the design of CAs and how the strategies that they employ impact how people interact with CAs and how well people perform using them. With current CAs being rather passive in nature, emulating the query-response search paradigm of search engines, various information retrieval researchers [27, 126, 156] have been calling for a paradigm shift to transform search engines from "passive query matchers" into "active search partners". Mixed-Conversational initiative (switching of initiative between the user and the system) has been identified as a crucial prerequisite for making such a transition from a passive to an active agent (cf. 2.5.1). However, the transition poses several challenges relating to the design of conversational interfaces. Studies on the design of CAs and their reception by users have recently been attracting increasing attention [28, 29, 113, 148] and have led to the development of guidelines and recommendations regarding the level of conversational initiative required by the agent. The guidelines provide generic suggestions on the design of CAs, such as: the agent should inform the user about its capacities [113, p.20] or the agent should use command and control for simple functional interactions" [106] etc. However, these suggestions fall short of providing details on how conversations with the agent should be conducted (i.e. "what to say?", "when to say it?" and "how to say it?") or on how to explore its impact on user search experience.

In terms of interaction with the user, two of the major challenges for developing conversational agents are:

1. Choosing the correct sequence of actions so as to help resolve a user's information need [11, 171].

2. Presenting search results effectively: i.e. "not overwhelming the users with information nor leaving them uncertain whether what they heard covered the information space" [158, p.14].

The above points are especially challenging in an audio channel (since users need to remember the information presented and reason about it simultaneously cf. [178]). When information is presented verbally, lack of persistence makes speech easy to miss and forget since - "Almost everyone is quicker to absorb written text than speech" [182, p.7], and "Despite being easy to produce, speech is much more difficult to analyze" [138, p.1].

Performing a goal-oriented dialogue can also be considered in terms of a cost-benefit trade-off, where the usefulness of a CA is determined by its ability to resolve a user's information need in a quick and comprehensive manner [11]. A problem with the current generation of CAs is that they provide information in a verbose way which puts strain on the user as they need to retain alternative options in their memory (an example of a conversation using Alexa's Kayak skill is presented in Appendix D for reference). For example, if a user asks about restaurants in the area, the CA will list a sequence of possible options, which the user will need to commit to memory (in order to compare or consider them later on). Consequently, due to cognitive overload, users tend to accept the first minimally acceptable option (satisficing behaviour) rather than continuing to absorb the cost of interaction in order to find a better option (maximising behaviour) [90]. On the other hand, users are unlikely to accept the CA's best suggested option without exploring alternative options [90]. Thus, a balance needs to be found between presenting enough options so that the user is satisfied and confident with their selection, and communicating the results in a manner which minimises the user's cognitive load.

Recently, several theoretical frameworks [126, 156, 162] have emerged to address the challenges of CA design. Trippas et al. [156] suggested that in Spoken Conversational Search (SCS) interaction, responsibilities should be shared between the user (who submits their information need) and the system (that actively decides which results to present back to the user via audio channel). As postulated by Trippas et al. [ibid], system initiative is crucial in the auditory setting, where the effective transfer of information depends on an active exchange between the user and the system. Otherwise, with a passive system, there is a risk of overloading users with information. The role of the active system is thus to provide incremental responses and create a common ground for collaboration via interaction with the user. Similarly, Feldman [57], highlighted that a conversational system needs to actively support a user by: (1) efficiently

navigating through the search space, (2) offering hypotheses based on knowledge bases and ontologies and (3) tackling the complexities and ambiguities of human language. More recently, Vakulenko et al. [162] proposed the QRFA (Query, Request, Feedback Answer) model. The model divides the conversational process into actions taken by the user (who submits queries and provides feedback) and the agent (answers queries or requests additional information form user.)

More closely related to our research is a theoretical framework proposed by Radlinski and Craswell [126] who put forward five properties that a search agent needs to have in order to be conversational. The properties are:

- User Revealment: where the user discloses their information needs to the agent
- Agent Revealment: where the agent reveals what the agent understands, what actions it can perform, and what options are available to the user
- Mixed Initiative: where both the agent and the user can share initiative and direct the conversation,
- Memory: where the agent tracks and manages the state of the conversation, the user's information need, etc., and,
- Set Retrieval: where the agent needs to be able to work with, manipulate and explain the sets of options/objects which are retrieved given the conversational context.

The above properties are required so that the CA can facilitate the search by helping the user to formulate their information need, and build expectations regarding the agent's capabilities. During the search process, a CA may take the initiative and use its memory to retain information relevant to the query. Before presenting results back to the user, the agent needs to reason about the utility of retrieved information and decide what and how to present the results to the user. Radlinski and Craswell posit that CAs should offer active elicitation support.

Although the above-mentioned theoretical frameworks [57,126,156,162] acknowledge the importance of interactivity between the user and the system in the search task, they do not provide detailed information on how the system should elicit and provide information during the search task, nor offer recommendations on CA design; which is the aim of this work.

5.2.1 Conversational Strategies

A more detailed conceptual theoretical model that builds upon Radlinski and Craswell's framework is proposed by Azzopardi et al. [11], where the actions and responses of the user and CA are enumerated and demonstrated based on examples. For instance, agent revealment actions are to: list the results, summarize the results, and compare the results, while agent revealment actions are to; extract specific details (i.e. slot-fill), clarify details, or not ask for further details

and passively await requests from the user. The different actions (and their combinations) that the CA can take, and how it implements these actions form the CAs strategy. Taken together, a CA can be described with respect to the approach that it takes when revealing and eliciting information. In terms of elicitation, these can be broadly categorised into three types:

- Passive: CA does not ask any questions - it leaves query refinement to the user.
- Active: CA asks questions to help a user make their query more specific.
- Pro-Active: CA suggests query reformulations that go beyond the scope of the original query and can thus change/expand the current information need.

In terms of revealment, meanwhile, three different approaches can also be taken:

- No Revealment: CA does not provide any results (either because no results are available, or the agents may decide not to disclose any results, in lieu of another action, i.e. to elicit more details about the information need).
- List: CA provides a detailed list with k elements: where k is the information retention threshold that varies between users.
- Summary: CA aggregates different sets of options and presents them to users in ranges.

In this work, we consider a CA's strategy as the combination of the type of elicitation approach and the type of revealment approach taken i.e. Passive/Active and List/Summary combinations. We will leave Pro-Active, No Revealment and Mixed approaches where CAs adopt a mixture of approaches for future work, and thus study the impact of taking a purely Passive vs. Active or Listing vs. Summary approach.
Current CAs tend to focus on well-defined, goal-oriented tasks for recurrent information needs (e.g. someone who already knows which flight/hotel they would like to book and are not concerned about exploring any alternatives etc.). Given the above classification, current CAs tend to be Passive (only asking a pre-defined set of questions to fill in slots) before Listing the results available. For example, on the Amazon Echo, KAYAK [80] and SkyScanner Fly Search [145] provide simple CAs. The KAYAK flight finder asks users a series of questions (slot filling), before providing a list of the cheapest available flights meeting those criteria, while the SkyScanner agent acts in a similar manner but reveals only the cheapest flight. Neither provide detailed information such as departure times, etc. nor provide any summary information such as range of prices, etc. Clearly, such agents lack the support for exploratory search tasks, where users explore and compare options (a phenomenon known as Comparative Shopping Notion [20]). In the current work, we will consider the context of flight booking where the user can explore a number of different options in order to select the best flight possible.

5.3 Method

To investigate the effect of the different elicitation and revealment strategies (see Section 5.2.1 for details), we conducted a controlled laboratory study using a 'Wizard of Oz' methodology [38] (more details in Sections 5.3.2 and 5.3.3). A flight booking context was selected for our experiment as it provides an exploratory, yet goal-oriented search task (i.e. a search activity whose purpose is to find and to book a flight). This type of task requires searching and exploration to find an 'optimal' flight given preferences and constraints, while providing a controlled interaction environment.

5.3.1 Study Design

Our study followed a 2x4 within-subjects design, where the independent variables were elicitation strategy (Passive and Active) and revealment strategy (Listing and Summarising) and the dependent variables were four aspects of user search experience: (1) cognitive load (measured by NASA TLX [69], (2) satisfaction (measured by SUS questionnaire [24] and SSES questionnaire [176]), (3) performance (evaluated in terms of the selected flight meeting the search criteria outlined in the task) and (4) interaction time for each of the search scenarios (measured in seconds) and number of conversational turns.
Since Study 2 involved a larger search space than Study 1, we decided to implement a Search Space Exploration Satisfaction (SSES) questionnaire to measure participants' satisfaction with the options provided by the agent (see Section 3.4.1.2 for a description of SSES questionnaire). We also added a Pareto Criterion when assessing participants' performance. Both of the new metrics will be discussed in more detail in Section 5.3.5.

5.3.2 Experimental Procedure

The experimental setup is diagrammatically presented in Figure 5.1.

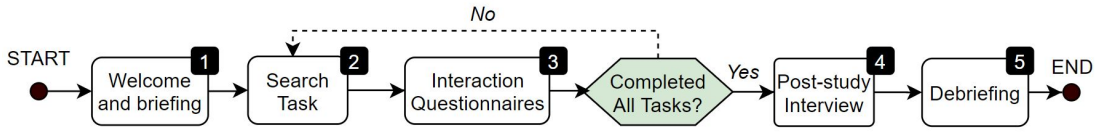


Figure 5.1: Illustration of Experimental Stages. Stage (2) consists of four search sessions (one with each CA); each session was followed by completing NASA TLX [69], SUS [24] and SSES [176] questionnaires.

Ethics approval for the study was granted by the Department of Computer and Information Sciences, University of Strathclyde (application no. 791). At the beginning of the study participants were welcomed, provided with an overview of the experiment and given the opportunity to ask questions about the experimental procedure. Participants were asked to fill in a demographic questionnaire, where they provide background details and comment on their experience

with goal-oriented CAs. The core part of the experiment involved four interactive search tasks (one per CA). Participants engaged with a CA simulated by the wizard to complete each flight search task (outlined in Section 5.3.5). Each participant completed four search tasks by interacting with four different CAs who represented our four interaction strategies (presented in Section 5.3.4). To control for the impact of the agent on search performance, a 4x4 Latin Square Design [23] was used to determine the sequence in which tasks and search agents were allocated to participants.

In order to ensure consistency and coherence of our experimental procedure, we informed the participants that:

1. They would be interacting with four different flight booking agents.
2. The experiment was not a memory test and that they were free to ask the CA for repetition whenever required.
3. Notes were not to be taken during interaction: to ensure that all participants were exposed to the same experimental conditions.

We did not provide any strict requirements as to how participants should interact with the agent other than not to interrupt the agent to mimic the interaction style of the current voice-based assistants (cf. [120]).

After each search scenario, participants were asked to complete three questionnaires i.e. NASA TLX [68], SUS [24] and SSES [176], see Figures B.1, B.2 and B.3 in Appendix B for details.

Having completed all the search tasks, the participants were debriefed and took part in a semi-structured interview where they were asked follow-up questions about their interaction with each of the CAs. We explicitly avoided asking participants to recall their interaction with a specific agent as this could have confused participants resulting in inaccurate accounts (cf. [108] on Interviewing Methods). Instead, following Flanagan's Critical Incident Technique [59], we asked participants to comment on a positive aspect(s) of their interaction, an aspect(s) that they found challenging, and to make suggestions on how their user search experience could be improved.

5.3.3 Wizard / Agent Setup

The study was conducted using a WOZ set-up [38] illustrated in Figure 5.2). Unlike the Study 1, the search results were provided by the wizard in his own voice rather than by means of synthetic speech. This decision was motivated by the fact that the search space in Study 2 was considerably larger than in Study 1 (75 flight options spread across 3 days vs. 8 flight options spread across 2 days) and would require a large number of synthetic prompts as well as adjusting them to different information presentation strategies.

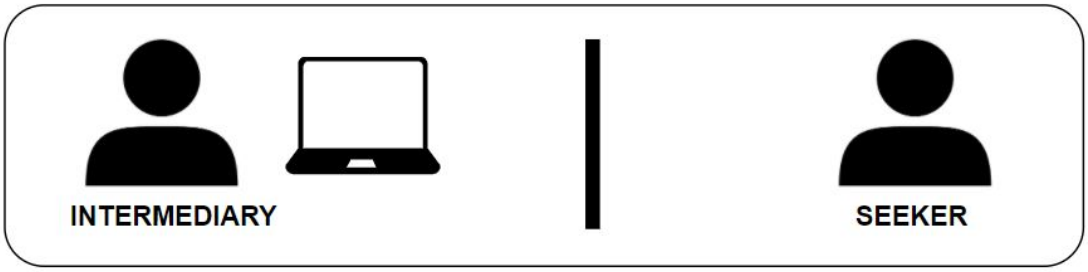


Figure 5.2: The Wizard of Oz framework: A wizard searches a flight database on behalf of a user and provides them with results.

Figure 5.3: Experimental Setup with the researcher's area marked with the dotted frame on the left and the participant's area marked with the white rectangle. During the experiment, participants were facing a partition screen that separated them from the wizard.

The experiment took place in the Usability Lab at the University of Strathclyde. The layout of the lab is illustrated in Figure 5.3. During the experiment, a researcher was present to provide task instructions and to record flight options selected by the participants during the interaction with CAs. To ensure consistency, we recruited a wizard with extensive call centre experience who was used to conducting structured goal-oriented conversations. None of the principal investigators acted as the wizard to reduce any potential unconscious experimental bias i.e. the possibility of carrying out the task in the way that supports the research hypotheses. During search tasks, in response to participants' queries the wizard searched a flight database

and provided results back to the participant. All of the flight options were stored in an excel spreadsheet (illustrated in Figure 5.4) and could be filtered based on the following attributes: airline type, departure time, arrival time, ticket price and duration of travel.

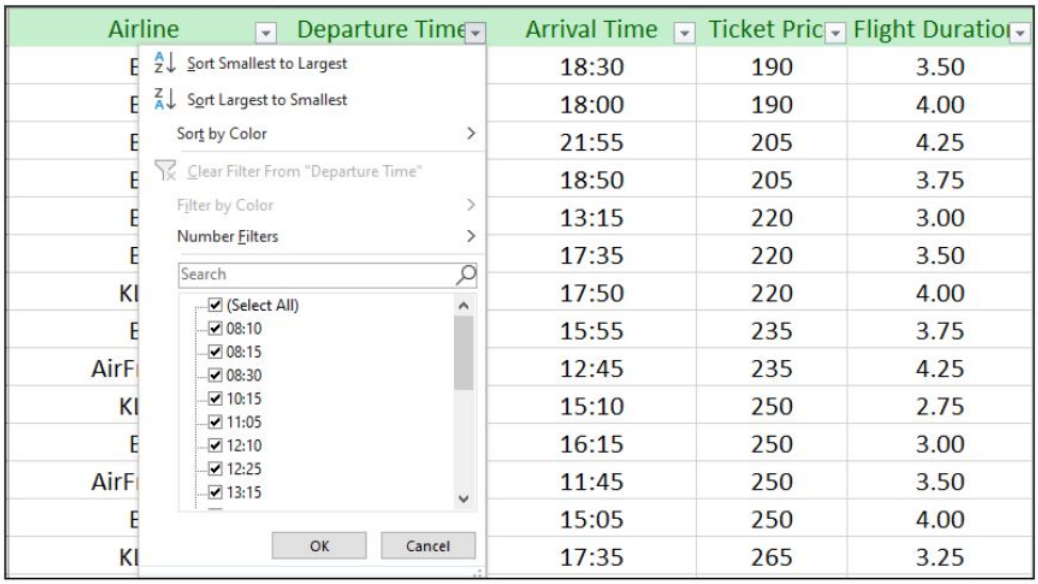


Figure 5.4: View of excel spreadsheet used by the wizard.

In order to make the simulation more realistic, and to exclude the impact of body language on communication, the wizard and participant were separated by a barrier and were not visible to each other.

5.3.4 Agent Conversational Strategies

Conversational strategies featured in this study are based on the spectrum of conversational strategies, introduced in Section 3.3.1, which specifies different methods of eliciting and revealing information. In total, we evaluated four conversational strategies (illustrated in Figure 5.5). We assigned a name to each of the conversational agents to make them easily identifiable to the participants. The names of the agents were:

1. Agent Angus: Passive Summarising (PS)
2. Agent Blair: Passive Listing (PL)
3. Agent Calum: Active Summarising (AS)
4. Agent David: Active Listing (AL)

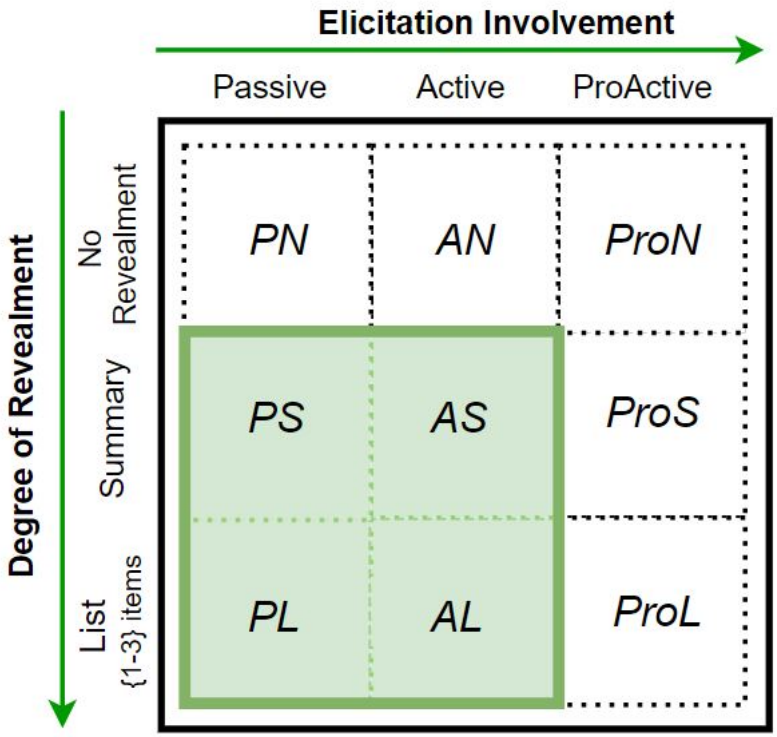


Figure 5.5: Conversational Agents Used in Study 2.

When designing the agents, we considered that the suitability of different information presentation methods might vary between different participants based on their ability to retain information. Commarford et al. [33] showed that people with short working memory spans prefer longer lists of options since providing fewer options within a conversational turn leads to a larger number of conversational turns. When using lists the number of presented elements needs to be capped to prevent overloading a participant's memory. In their study, Demberg et al. [41] limited the number of provided options to 3 at a time. Demberg et al. refrained from presenting complex information in a single conversational turn. Instead, each time that there were more than three flights in the results cluster, only attributes that distinguished the flights were presented to the user (e.g. "The three direct flights are on Continental, Lufthansa, and Delta. They arrive at 9:55 am, 10:15 am, and 11:15 am"). In our study, in order to avoid overloading the working memory of our participants, we have taken a slightly more conservative approach than Demberg et al. [41] and limited the number of presented options to a maximum of two flights at a time. This decision was taken to facilitate the retention of information for the broader spectrum of participants.

5.3.5 Simulated Search Tasks

The search tasks used in the current study are a modified version of the tasks used in Study 1 (see Section 4.3.5 for more details). In Study 2, the tasks were modified to make them more

relatable and increase participants' motivation to engage in a more thorough exploration of the search space. This section provides a brief overview of the tasks used in the current study.
Booking a flight is a prototypical goal-oriented task that allowed us to measure the performance of participants and evaluate the adequacy and impact of the different conversational strategies. In the context of flight booking, numerous factors such as the cost of the flight, its duration, departure/arrival time, the airline, fare class etc. impact upon which flight is selected. For our tasks, we focus on the most salient variables: flight price, flight duration and arrival time (as identified by the IATA Global Passenger Survey [122]). In our experiment, participants were asked to imagine that they were a traveller looking for a one-way flight from Glasgow Airport. There were four search scenarios in total (one per conversational strategy).
An example search scenario is presented in Figure 5.1. Participants were instructed to explore available flight options over three days and to find the shortest and cheapest flight possible. The rationale was that when searching for flights most people prefer to get to their destination as quickly as possible for the least amount of money. Participants were also asked to find flights which would preferably get them to their destination by a certain time of day. Travel time and price were considered hard constraints (e.g. find the cheapest and shortest flight), while preferred time of arrival was considered a soft constraint (e.g. "you would like to arrive around 10am" etc.) We also provided a more extensive background story to make the scenario more relatable and encourage greater participant involvement in completing it.

You will be attending a student conference in **Stockholm**. You will be travelling there on either **Monday the 5th, Tuesday the 6th, or Wednesday the 7th of November**. Your university advised you that you will be allocated money from your conference fund that you will use to fund other events till the end of your academic course. To be able to attend more events in the future, you want to save money while not spending too long getting there. The student dorms where you will be staying charge extra for late check-in, so you will be aiming to arrive at around 7pm to be able to check in to your accommodation on time.
Indicative Request: Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time).
Note: Please wait for the agent to finish before you start to speak.

Search Scenario 5.1: Travel destination and possible days of travel are provided in bold for participants' convenience.

Contrary to Study 1, possible flight options in Study 2 for each task were carefully controlled to create a tension between the shortest flights and the cheapest flights available, such that the cheaper flights tended to be longer, while the more expensive flights tended to be shorter, i.e. there was not one 'best flight'. Instead, as is typically the case when booking flights, there was a trade-off involved. As explained in Section 3.4.2.1, given two flights which have same duration and arrive at the same time, the optimal choice would be the cheaper flight. Conversely, given two flights which cost the same price and arrive at the same time, the optimal choice would be the shorter flight. Stated more formally, flights which are on the Pareto frontier are considered

Pareto Optimal, providing us with a means of evaluating the merits of a participant's flight choice (see Figure 3.5 for the illustration of Pareto frontier).

Participants were asked to find a flight within a three-day window i.e. they were given flexibility to decide when they arrived, and consequently had to search over different days to explore all the possible options. To check/ensure that participants explored the different options, different days presented different subsets of results, where flights on that day were all sub-optimal (both longer and more expensive), mainly cheaper (but longer), or mainly faster (but more expensive). By designing the choices in this manner we wanted to ensure that participants had to explore flights over a number of days, and that the trade-off meant they would have to make compromises (so as to create a more realistic search scenario), while still being able to evaluate the quality of their choices (in terms of whether they selected a Pareto Optimal flight or not, and if not, did participants waste time and/or money).

5.3.6 Interaction Design

A diagrammatic illustration of interactions with CAs is presented in Figure 5.6 (passive agents) and Figure 5.7 (active agents).

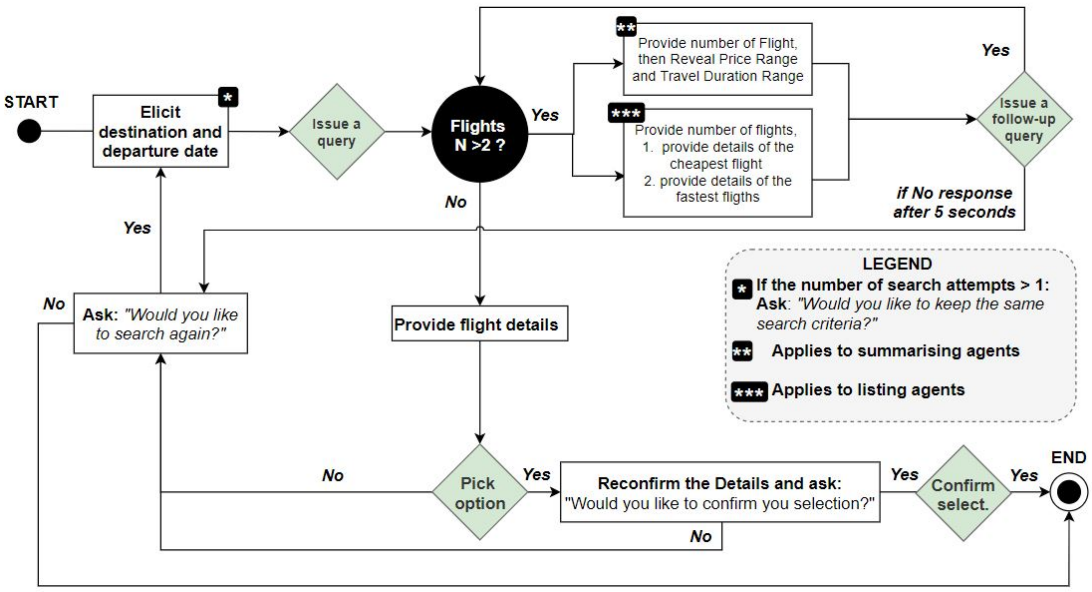


Figure 5.6: Passive Agents Interaction Flow: White squares denote the agent's actions and green diamonds denote the participant's actions.

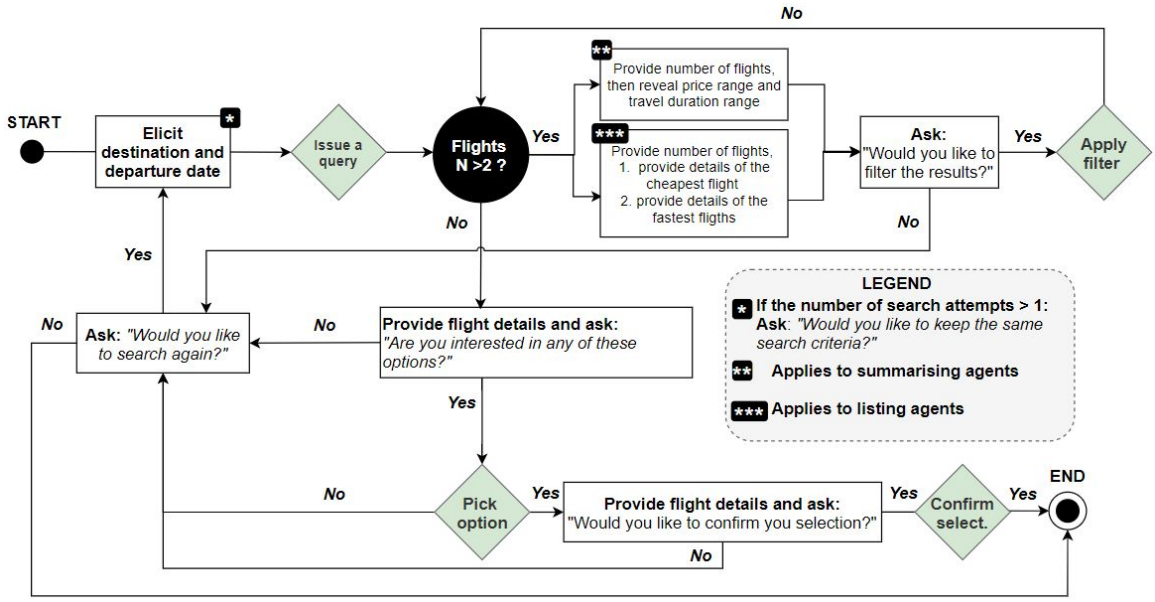


Figure 5.7: Active Agents Interaction Flow: White squares denote agent's actions and green diamonds denote participant's actions.

To ensure consistency of interaction, each agent was based on an algorithm that specified how to elicit and reveal information from and to the participant. The 4 algorithms used in Study 2 are

provided below. Every interaction began with the agent prompting the participant to provide them with their destination and day of travel. If the number of flight options corresponding to a participants' query was larger than 2, the Summarising agents (Angus and Calum - Algorithms 1 and 3) provided the participant with a range of flights followed by their price range and range of travel duration, while the Listing agents (Blair and David - Algorithms 2 and 4) provided details of the cheapest and the fastest flights. If the number of results was 2 or fewer, all agents provided detailed information about the flight. During the search sessions the Passive agents (Angus and Blair) did not make any suggestions with regards to results filtering while the Active agents (Calum and David) asked the participant questions to progressively narrow down the list of flight options.

ALGORITHM 1: Passive Summarising Strategy - **Agent Angus**

```
1  while Flight has not been selected do
2      wait for a query
3      for a query do
4          if n of flights > 2, then
5              provide n of flights
6              provide flights range (price)
7              provide flights range (travel duration)
8          else
9              provide detailed flight result(s) (airline, departure and arrival times, price
                and duration)
10         end
11     end
12 end
```

ALGORITHM 2: Passive Listing Strategy - **Agent Blair**

```
1  while Flight has not been selected do
2      wait for a query
3      for a query do
4          if n of flights > 2, then
5              provide n flights
6              provide details of cheapest flight (price,duration)
7              provide details of fastest flight (price, duration )
8          else
9              provide detailed flight result(s) (airline, departure and arrival times, price
                and duration)
10         end
11     end
12 end
```

The questions (See Algorithm 3 and Algorithm 4, lines 8-10) were always presented in chronological order, unless a participant made a specific request (e.g. 'Tell me about the earliest flight on Monday.') If a query returned no results (Ø), the agent would suggest changing filtering criteria. For instance, if there were no flights within the specified price range, the agent would

ALGORITHM 3: Active Summarising Strategy - **Agent Calum**

```
1  while Flight has not been selected do
2      wait for a query
3      for a query do
4          if n of flights > 2, then
5              provide n of flights
6              provide flights range (price)
7              provide flights range (travel duration)
8              Case 1. say: 'Would you like to filter by price?'
9              Case 2. say: 'Would you like to filter by duration?'
10             Case 3. say: 'Would you like to filter by departure time?'
11         else
12             provide detailed flight result(s) (airline, departure and arrival times, price
                and duration)
13             say: 'Would you like to select this flight/any of these flights?'
14         end
15     end
16 end
```

ALGORITHM 4: Active Listing Strategy - **Agent David**

```
1  while Flight has not been selected do
2      wait for a query
3      for a query do
4          if n of flights > 2, then
5              provide n flights
6              provide details of cheapest flight (price,duration)
7              provide details of fastest flight (price, duration )
8              Case 1. say: 'Would you like to filter by price?'
9              Case 2. say: 'Would you like to filter by duration?'
10             Case 3. say: 'Would you like to filter by departure time?'
11         else
12             provide detailed flight result(s) (airline, departure and arrival times, price
                and duration)
13             say: 'Would you like to select this flight/any of these flights?'
14         end
15     end
16 end
```

suggest relaxing the search criteria (e.g. 'There are no flights for less than £200 that leave before 2 pm, try to increase your price, change travel time' etc.). If the query issued by the participant was outwith the scope of the agent, it would inform the user that the given functionality was not supported. The agent could search only one day at the time so that participants had to perform a number of searches to explore the space (which is consistent with flight booking systems).

In order to ensure consistency across the strategies we created a list of exceptions that particular agents could not handle. The exceptions were:

- Searching multiple days at the same time
- Comparing results

If a participant issued any of the above queries, the CA replied: 'Sorry, this functionality is not currently supported.'

Participants were however allowed to ask agents to re-apply the same search criteria between different days. When switching between the days, agents asks the participant if they would like to use the same filters from their most recent search. By using algorithms to dictate the agent's interactions we managed to ensure consistency across participants. Examples of participant-agent interactions are provided for reference in Table 5.1 (Summarising Agents) and Table 5.2 (Listing Agents).

Table 5.1: Interactions with Summarising Agents. Note: Questions that distinguish the Active Summary Agent from its Passive equivalent are underlined for clarity.

Passive Summary Agent (PS)	**Active Summary Agent (AS)**
	AS: Hello my name is Calum, I am a flight search agent and I will help you to find your flight, to get started, please provide me with your destination and day of departure.
	Participant: Hi Calum, the destination is London and the date of departure is between Wednesday the twelfth and Friday the fourteenth of December.
PS: Hello, my name is Angus. I am a flight search agent and I will help you to find your flight. To get started, please provide me with your destination and day of departure.	**AS:** Ok, which day would you like to check first?
Participant: Emm, I am going to Rome on a Friday in September. But it doesn't specify any specific Friday yet.	**Participant:** Wednesday.
PS: One moment, which date would you like to check first?	**AS:** One moment please. There are twenty-nine flights available. The prices range from seventy pounds to one hundred and forty and the durations range from one hour thirty to four hours fifteen. Would you like to filter by price?
Participant: The Friday fourteenth.	**Participant:** Yes.
PS: One moment please. There are thirty flights available. The prices range from one hundred and forty pounds to two hundred and fifteen pounds and flight durations range from three hours to six hours fifteen.	**AS:** Ok, how would you like to filter?
Participant: Can you check Friday the twenty first instead?	Participant: Cheapest first, please.
PS: One moment please. There are twenty-five flights available. The prices range from one hundred pounds to one hundred and seventy pounds and the flight durations range from four hours to six hours fifteen [. . .]	**AS:** One moment please. Do you have a maximum price in mind?
	Participant: Ninety.
	AS: There are six flights available. The prices range from seventy pounds to ninety pounds and the durations range from three hours to four hours. Would you like to filter by duration?
	Participant: No, can I filter by arrival time?
	AS: Ok, how would you like to filter?
	Participant: Arrive at or before ten a.m.
	AS: There are no flights that arrive at or before ten am. You may want to increase the arrival time. [. . .]

Table 5.2: Interactions with Listing Agents. Note: Questions that distinguish the Active Listing Agent from its Passive equivalent are underlined for clarity.

Passive Listing Agent (PL)	**Active Listing Agent (AL)**
PL: Hello, my name is Blair. I am a flight search agent and I will help you to find your flight. To get started, please provide me with your destination and day of departure.	
Participant: I would like to get Stockholm on Monday, the fifth of November.	
PL: One moment please. There are twenty-nine flights available. The cheapest flight costs one hundred and seventy pounds and takes four hours forty-five and the fastest flight costs two hundred and twenty pounds and takes two hours thirty minutes.	**AL:** Hello my name is David, I am a flight search agent and I will help you to find your flight, to get started, please provide me with your destination and day of departure.
Participant: What is the time of arrival for the cheapest flight?	**Participant:** Stockholm, Monday the fifth.
PL: The cheapest flight arrives at fifteen zero five.	**AL:** One moment please. There are twenty-nine flights available. The cheapest flight costs one hundred and seventy pounds, and takes four hours forty-five and the fastest flight costs two hundred and twenty pounds and takes two hours thirty. Would you like to filter by price?
Participant: Ok, how about Tuesday, the sixth of November?	**Participant:** Yes, please.
PL: One moment please. There are twenty-five flight available. The prices range from one hundred pounds to one hundred and seventy pounds and the flight durations range from four hours to six hours fifteen.	**AL:** Ok, how would you like to filter?
Participant: What's the arrival time for the cheapest flight?	**Participant:** Lowest to highest, or between.... Well, between two... Sorry can you repeat the cheapest price again, please?
PL: The cheapest flight arrives at eighteen zero zero.	**AL:** It is one hundred and seventy-five pounds.
Participant: And how about Wednesday, the seventh of November?	**Participant:** Could you filter between two hundred and three hundred pounds?
PL: One moment, please. There are twenty-eight flights available. The cheapest flight costs one hundred and thirty pounds and takes four hours forty-five minutes and the fastest flight costs one hundred and eighty pounds. and takes three hours	**AL:** Ok, one moment please. There are twenty-three flights available. The cheapest flight costs two hundred and ten pounds and takes three hours forty-five, and the fastest flight costs two hundred twenty pounds and takes two hours thirty. Would you like to filter by duration?
Participant: What is the arrival time for the cheapest flight?	**Participant:** Emm, no.
PL: The arrival time for the cheapest flight is fourteen forty-five.	**AL:** Ok, would you like to filter by departure time?
Participant: And what is the price, again?	**Participant:** Yes, please.
PL: One hundred and thirty pounds.	**AL:** Ok, how would you like to filter?
Participant: And what was the arrival for the one on Tuesday the sixth of November that is the cheapest?	**Participant:** Emm, departing before three pm.
PL: That was eighteen zero zero.	[...]
Participant: And the price?	
PL: One hundred and seventy pounds.	
Participant: Ok, and is there a cheaper flight... No, sorry. What is the cheapest option on the seventh again?	
PL: The cheapest flight on the seventh is one hundred and thirty pounds. It is an Air France flight, it departs at ten zero zero and arrives at fourteen forty-five, it has flight duration of four hours forty-five minutes. [...]	

5.3.7 Participants

Participants were recruited by posting notices on the campus of Strathclyde University. A snowball sampling method [19] was also utilised where word of mouth advertising was encouraged. All participants were compensated for their time with a £10 shopping voucher for taking part in the study. A demographic questionnaire was used to gather information about participants and their familiarity with travel search engines and conversational search technology. 24 participants completed the study. There were 12 males and 12 females. The youngest participant was 18 years old and the oldest was 40. The average age was 26 (SD 6). 13 of the participants were native English speakers. 11 participants who were non-native English speakers were all university students with a minimum of 6.5 IELTS score. 80% of participants (n = 19) reported having used flight booking websites such as SkyScanner or Expedia before. All but 2 participants reported to having booked or searched for a flight within the last year, with 30% saying that they had done it more than ten times in that period. When asked to rank the most important criteria considered when booking a flight, 82% mentioned price as the most important factor, followed by number of stops 8%, travel time 5%, and airline preference 5%. As for the 2nd most important criterion: 33% of participants chose travel time, followed by departure time (25%) and price (17%).

5.4 Results

Our experimental results are presented in Table 5.3 (Subjective Measures), Tables 5.4 and 5.5 (Performance Measures) and Table 5.6 (Interaction Metrics). Since most of our data was not normally distributed, for pair-wise comparisons, unless otherwise stated, we use the Kruskal Wallis H Test. All of the conducted statistical tests are two-tailed.

5.4.1 Subjective Measures: Cognitive Load and Satisfaction

A summary of the findings for subjective measures (NASA TLX [68], SUS [24] and SSES [176] questionnaires) can be found in Table 5.3. Pairwise comparisons between Passive and Active agents, and Summarising and Listing agents revealed no statistically significant differences. The most notable differences were observed between Active and Passive agents for: Performance (Z =-1.327, p = 0.184), Results Presentation Speed (Z = -1.212, p = 0.226), Overall NASA TLX (Z = -0.902, p = 0.367), Mental Demand (Z = -0.646, p = 0.518) and SUS (Z = -0.539, p = 0.590), The differences were less pronounced in the case of Summarising vs. Listing agents: Overall NASA TLX (Z = -0.718, p = 0.473), SUS (Z = -0.532, p = 0.595), Mental Demand (Z = -0.176, p = 0.860), Performance (Z = -1.118, p = 0.264) and Results Presentation Speed (Z = -0.063, p = 0.950).

We observe that, overall, Passive CAs put more strain on a participant's workload. As can be seen in Table 1, the average overall NASA TLX score for the Passive agents was 33, which

Table 5.3: Subjective Measures. The table aggregates data from questionnaires that reflect participants' perception of the Conversational Agents. Note: For NASA TLX the lower the score, the better. For SUS and SSES, the higher score the better.

[Metrics] \ [*Agent Type*]			*Passive*	*Active*	*Summ*	*Listing*	*Passive Summ*	*Passive Listing*	*Active Summ*	*Active Listing*
NASA TLX Overall(0-100), Per Item(0-20)	Mental	Med (IQR)	10 (9)	8 (9)	9 (10)	10 (9)	11 (9)	10 (8)	8 (10)	9 (9)
	Demand		6 (8)	7 (8)	5 (8)	7 (9)	4 (8)	7 (8)	7 (8)	7 (10)
	Effort		7 (7)	5 (6)	5 (5)	7 (7)	5 (7)	8 (7)	5 (5)	5 (8)
	Performance		5 (11)	5 (9)	3 (10)	5 (10)	3 (12)	6 (10)	3 (9)	5 (10)
	Frustration		5 (6)	5 (5)	6 (6)	4 (4)	5 (8)	5 (5)	5 (5)	4 (4)
	Temporal Demand		**33.17 (21.33)**	**28.35 (20.27)**	**29.69 (20.68)**	**37.78 (21.15)**	**31.75 (4.98)**	**34.6 (4.65)**	**28.55 (4.30)**	**28.15 (4.86)**
	Overall									
SUS (0-100)	**Overall**	Med (IQR)	**76 (29)**	**80 (22)**	**79 (23)**	**78 (27)**	**79 (26)**	**71 (25)**	**79 (20)**	**81 (26)**
SSES Overall(0-20), Per Item(1-5)	Presentation	Med (IQR)	4 (1)	4 (1)	4 (1)	4 (1)	4 (1)	4 (1)	4 (1)	4 (1)
	Overview		4 (1)	4 (1)	4 (1)	4 (1)	4 (1)	4 (1)	3 (1)	4 (2)
	Confidence		3.5 (2)	3.5 (2)	3 (1)	4 (2)	3.5 (2)	3 (2)	3 (1.75)	4 (2)
	Speed		4.5 (1)	4 (1)	4 (1)	4 (1)	4 (1)	4.5 (2)	4 (1.75)	4 (1)
	Overall		**16 (4)**	**16 (4)**	**16 (4)**	**16 (6)**	**16.5 (4)**	**16 (6)**	**15.5 (4)**	**16 (4)**

is seven points more than for the Active agents (26). Notably, among all the agents, Passive Listing performed the worst (NASA score of 37). An analogical trend can be observed for Mental Demand, where the Passive agents performed worse than the Active ones (10 vs. 8 - respectively). In terms of Performance, the top strategies were Active and Summarising, while Passive and Listing performed noticeably worse (5 for each in the former pair compared to 7 for each in the latter). We observed a similar pattern for the SUS scores, with Active and Summarising agents, again, outperforming Passive and Listing agents. In terms of individual strategies, again, Passive Listing was the worst (SUS score of 71) while Active Listing came out on top (SUS score of 81).

5.4.2 Objective Metrics - Task Performance and Interaction Metrics

The objective metrics used to evaluate all CAs were: (1) participants' performance considered in terms of selected flight options (primary indicators) and expenditure of money and time (secondary indicators); and (2) interaction metrics - task completion time and number of conversational turns. Both metrics are described in turn in this section.

5.4.2.1 Task Performance

As a reminder, we measured task performance in terms of the relationship of the selected flight to the Pareto frontier (Discussed in Chapter 3, Figure 3.5) - hard constraint, and in terms of preference (desired travel time specified in the search scenario) - soft constraint. The flight was considered Pareto Optimal if, and only if, it met both of the constraints. In terms of the task outcome; for each conversational strategy, we also report if participants lost money and/or wasted time by selecting sub-optimal flight options.

Task performance measures are presented in Table 5.4 and Table 5.5. We consider Task Performance in terms of Primary Indicators ('Meeting Time Preference' and 'Hitting Pareto Optimal'), and Secondary Indicators ('Losing Money' and 'Wasting Time'). Primary Indicators concern meeting the flight arrival requirements specified in each search scenario and indicate if the selected flight was Pareto Optimal (i.e. it offered the best combination of price and duration while meeting the arrival requirement). Secondary Indicators concern the impact of a participant's flight selections on their resources, i.e. time and money (i.e. if a participant selected a more expensive flight when a cheaper one of the same duration was available - they lost money; if they selected a flight with a longer duration when a shorter alternative was available at the same price – they lost time). Since all the indicators are considered in binary categories, i.e. a selected flight is either on the Pareto frontier or not, the selected flight either meets the time preference or not etc., Cochran's Q Test was used to compare different conversational strategies. The Bonferroni adjusted alpha-level (.008) was used for all post-hoc analyses.

Table 5.4: Performance Measures (Primary Indicators) '***' signifies $p < .001$.

[Metrics] \ [*Agent Type*]		*Passive*	*Active*	*Summ.*	*Listing*	*Passive Summ.*	*Passive Listing*	*Active Summ.*	Active Listing
PRIMARY INDICATORS	Met Time Preference	28/48 (58%)	**40/48 (83%)*** **	33/48 (69%)	35/48 (72%)	13/24 (54%)	15/24 (63%)	20/24 (83%)	20/24 (83%)
	Pareto Optimal	5/48 (10%)	13/48 (27%)	11/48 (23%)	7/48 (15%)	3/24 (13%)	2/24 (8%)	8/24 (33%)	5/24 (21%)

Meeting Time Preference: There is a statistically significant difference between the conversational strategies (Cochran Q =23.368, $p < .001$). For pair-wise comparisons, a McNemar post-hoc test indicated a statistically significant difference between Active and Passive agents ($p < .001$) but not between Summary and Listing agents (p =.79).

Hitting the Pareto Optimal: There is no statistically significant difference between strategies (Cochran Q = 6.667, p =.83). There is, however, a noticeable difference between the Active and Passive agents for booking flights on the Pareto Optimal, 5 and 13 respectively (p =.077). The difference is less pronounced between Summary and Listing strategies, 11 to 7 respectively (p =.388).

Table 5.5: Performance Measures (Secondary Indicators). '**' signifies p < .01.

[Metrics]	[*Agent Type*]	*Passive*	*Active*	Summ	*Listing*	*Passive Summ.*	*Passive Listing*	*Active Summ*	*Active Listing*
SECONDARY INDICATORS	Money Lost	11/48 (23%)	9/48 (19%)	8/48 (17%)	12/48 (25%)	6/24 (25%)	5/24 (21%)	2/24 (8%)	7/24 (29%)
	Time Wasted	27/48 (56%)	**18/48** (**38%**)	21/48 (44%)	24/48 (50%)	13/24 (54%)	14/24 (58%)	8/24 (33%)	10/24 (41%)

Time Wasted: We observe a statistically significant difference between the conversational strategies (Cochran Q = 13, p =.005). A post-hoc test indicated a statistically significant difference between Active and Passive agents (p =.004) but not between Summary and Listing agents (p =.219).

Money Lost: No statistically significant differences were observed between the conversational strategies (Cochran Q = 4.286, p =.232). Pair-wise comparisons indicate that there is little difference between Active and Passive agents (p =.625) or between Summary and Listing agents (p =.219).

Overall, in terms of performance, the Active conversational strategy consistently outperforms the Passive conversational strategy for all performance aspects under consideration. An analogical trend can be observed for the Summarising strategy, which outperforms the Listing strategy for all aspects but 'Meeting Time Preference'. At the level of the individual CAs, an Active Summary yields the best performance for both Primary and Secondary performance indicators. The most notable differences are observed when it comes to 'Hitting the Pareto Optimal' and 'Money Lost'.

5.4.2.2 Interaction Metrics

Objective measures are provided in Table 5.6. Objective measures provide information about the impact of the conversational strategy on the duration of conversation and the number of conversational exchanges between the agent and the participant. We considered conversation time and the number of turns as indicators of the agent's efficiency. We did not observe any statistically significant differences for any of the objective measures between the Passive and Active agents with regards to time (Z = -953 p = 0.341) and number of turns (Z = -0.737 p = 0.461) or for Summarising vs. Listing comparison: (Z = -0.590, p = 0.555) and (Z = -1.299, p = 0.194) respectively. However, the Active Summarising strategy stands out, with the highest number of conversational turns (Med = 20) and the longest time of interaction (Med = 271s).

Table 5.6: Interaction Metrics. Note: Interaction Time is rounded up to the nearest second.

[Metrics]	[*Agent Type*]	*Passive*	*Active*	Summ	*Listing*	*Passive Summ.*	*Passive Listing*	*Active Summ.*	*Active Listing*
Interaction Time in sec.	Med (IQR)	232 (122)	252 (132)	246 (136)	226 (108)	243 (113)	210 (180)	271 (161)	229 (88)
Conversational Turns		18 (10)	19 (8)	19.5 (10)	17 (7)	18 (9)	17.5 (13)	20 (13)	17 (6)

Section 5.4 presented quantitative results from the analysis of the experimental data (questionnaires, task outcomes, and interaction metrics). In the following section, in order to provide additional context we will present a qualitative analysis of the semi-structured interviews.

5.5 Semi-structured Interviews

All of the study participants (n=24) took part in semi-structured interviews after all the interactive tasks were completed. The interviews provided a number of insights on the participant's perception of conversation strategies (represented by four CAs - described in Section 5.3.4).

5.5.1 Procedure

All interviews were audio recorded. The length of the interview ranged from 88 seconds to 640 seconds (M = 268s, SD = 152s). Following the Critical Incident Technique [59], participants were asked three questions:

1. Was there anything in the agent's conversational behaviour that you found useful during interaction? (Positive Aspects of Interaction)
2. Was there anything that you found challenging? (Interaction Challenges), and
3. How would you improve the agent(s)? (Suggestions for New Functionalities).

5.5.2 Analysis

Interviews were transcribed and analysed using a hybrid thematic approach that involved deductive and inductive coding [58]. Transcripts were split into utterances and marked with the corresponding participant IDs ([P1], [P2], [P3] etc.) In the first, deductive stage, based on the interview questions, participants' utterances were grouped into three categories: (1) Positive Features of Interaction, (2) Interaction Challenges and (3) Suggestions for New Functionalities. At the second, inductive stage, two members of the research team worked independently to prepare an initial code book based on an analysis of participants' utterances in each group. Any differences in the code book were resolved by discussing theme adequacy (ensuring that themes are relevant to conversational strategies) and uniqueness (ensuring that themes are not duplicated). Following the discussion, the two researchers agreed on the final version of the code book which contained 15 codes (see Table 5.7 for details). The transcripts were then re-coded by the primary author and an external researcher (a person who did not take part in the study). The coding comparison indicated a substantial inter-coder agreement, with Cohen's Kappa coefficient (K) = 0.78 (for more information on interpreting K ranges see [163]).

Table 5.7: Codes Identified in Semi-structured Interviews.

Codes within Interview Categories					
Positive Features	**# mentions**	**Challenges**	**# mentions**	**New Functionalities**	**# mentions**
Active Filtering Support	13	Remembering and Reasoning about Results	16	Result Synthesis	17
Agent's Responsiveness	10	Lack of Visual Feedback	10	Controlling the Scope of Flight Details	15
Learning Effect	7	Finding optimal Flight	5	Displaying Results on the Screen	12
Agent's Memory	6			Audio Bookmark	11
Results Presentation	5			Additional Travel Services	5
User in Control of Search	3			Agent Embodiment	2

5.5.3 Findings

In the current sub-section, we summarise the findings of our thematic analysis with regards to three categories , namely (1) Positive Features of Interaction, (2) Interaction Challenges and (3) Suggestions for New Functionalities. We refer to interview participants by ID numbers (IDs: P1-P24). The quotes featured below have been selected based on the number of instances identified within each code category. Note that short quotations (of less than 5 words) are not featured in the analysis, as they were not in-depth enough to provide substantial insights. The code categories used in our analysis are provided in bold for ease of reading.

5.5.3.1 Positive Features of Interaction

In terms of positive features of interaction, several participants (n=13) referred to the **Active Filtering Support** employed by the Active agents as helpful. Flight filtering suggestions offered by the Active agents made it easier for participants to reformulate their search queries and, consequently, made their search process more focused, as illustrated by the following quotes:

> *I was suggested things to search rather than just having to think straight away. I was like, I can get my head around what's going on rather than just 'Uh, what do I ask first, what's going on?' So that was quite good* (P3)

> *I think the second one [the Active Listing agent] that gave you some hints that you might want to do this. Both of them had this 'would you like to filter by price?' They gave you some kind of ideas on how to interact with it.* (P9)

On the other hand, a small number of participants (n=3) praised the Passive agents because they gave them a feeling of **Being in Control of the Search Process**. Participants P21 and P24 considered the suggestions made by the Active agents as disruptive to their search, as illustrated by the quotes below.

> *I have enjoyed the first two [the Passive Listing and Passive Summarising agents] most probably. Mainly, because I had the impression that I controlled the agent and*

not the agent tried to persuade me to do stuff. So, I took most of the decisions myself, basically. OK, so the best thing about these two systems was that they were quite autonomous. That it was up to you to decide what you want to choose. (P21)

I liked the two [the Passive Listing and Passive Summarising agents] that did not follow up giving you [suggestions]... Like trying to almost force you to filter in a certain way. Because, then I felt guilty when I did not want to do that. (P24)

Some participants (n=7) commented that the **Learning Effect** played a role in their interactive search experience. As participants were becoming more familiar with the scope and requirements of the task, their confidence grew. This process of gaining search experience, is demonstrated by the following quotes:

Because by the time I got Angus [the Passive Summarising agent], I already knew what to ask so like, it was a bit easier. (P3)

I think the last two [the Active Summarising and Active Listing agents] because I got used to using them properly. (P5)

I also think that after each one I kind of gained confidence in using it. (P23)

It should be noted that, in the experiment, we used a Latin Square Design [23] to reduce the impact of the sequence in which agents were introduced on participants' performance. Participants also made generic comments (concerning all agents) about the CAs' **Responsiveness** (n=10): their prompt responses; **Memory** (n=6): the fact that the CA remembered user queries over multiple turns; and **Results Presentation** (n=5): results being presented in an easy to follow way. Since most of the comments did not provide much detail on conversational strategies and/or were too brief (less than five words), they are not featured in the analysis.

5.5.3.2 Interaction Challenges

The interaction challenge most frequently mentioned during the interviews was **Remembering and Reasoning about the Search Results** (n=16). Three participants commented about the need to write information down during the interaction. P5 stated:

So, you need kind of to make mental notes and say right, that is the cheapest. Not all information is in front of you. I think that unless you have a pen and paper in front of you to scribble down notes you will struggle to remember because there is quite a lot to remember. Different flight times, duration, costs. All those things.

P14 noted that the level of details that they had to process to find a good flight that met their provided search criteria was challenging.

Each flight has many details on them, and you are searching, and you are under pressure searching for flights with certain criteria and you need to remember flights details on each day and keep them in your head. (P14)

Another challenge related to information overload was **Lack of Visual Feedback**. The lack of a results display screen was identified as a problem (n=10). P5 commented:

The hardest thing that I can remember, it's obviously that you don't have it in front of you.

One of the participants noted that the lack of **Agent Embodiment** made the search process challenging.

It's always easier to have a person in front of you whom you can see and dealing with all your doubts and questions rather than having a virtual system. (P2)

Some participants (n=5) also briefly commented on the difficulty of **Finding an Optimal Flight** as specified in the search scenarios as they had to take multiple flight criteria (price, travel time, arrival time) into account before making their final choice.

5.5.3.3 Suggestions for New Functionalities

Participants made several suggestions on how to improve the functionality of the agents. The functionality that was most frequently requested (n=17) was a **Results Synthesis** that would involve a CA carrying out search over multiple days and then providing a summary of its findings back to the user. The characteristics of the synthesis feature are reflected by the quotes below.

If I could have asked what is the cheapest flight to London, between Wednesday the twelfth, Thursday thirteenth and Friday fourteenth then I would save a lot of time. So just comparing between days, would be my most requested feature. (P8)

So, I would prefer to be able to look up all the flights over multiple dates together so rather than having to remember the flights for individual days just being able to compare all flights. (P23)

A **Results Synthesis** feature was considered useful, as it would allow for a quicker exploration of the search space and facilitate comparisons between different options. Participants (n=15) requested the ability to **Control the Scope of Flight Details** that are provided by the agent to prevent information overload. Participants made comments about adjusting the level of detail and type of information that they are provided with.

I am saying that I don't need any more information. So, for example, time of the flight. Because I think this is kind of disturbing. (P20)

> *The thing is that, I did not want to filter them [results] according to prices and things like that because, I just wanted to be able to tell just straight away what my requirements are.* (P4)

A participant suggested that the agent should be able to facilitate a search by starting with a single flight criterion and then progressively narrowing the scope of the results down by different attributes:

> *Ok, so just limiting the criteria to one aspect of it and then narrowing it down. Maybe you keep filtering by price starting by going here is the prices, and then do you wish to know that other information and giving the option to say yes or no to it.* (P 24)

Another frequently requested function was an **Audio Bookmark** – the ability to store the flight result in an agent's memory for future reference. Participants (n=11) noted that being able to mark/save a flight for later could have improved their search experience. Below are some of the quotes that illustrate this point:

> *Maybe being able to tag certain flights if I heard the agent say this flight leaves at this time and has this price and say: 'Can you tag this flight?', and then after talking to it for a bit check my tagged flights, maybe. I think that would be helpful. Cause then I don't feel the pressure of 'Ah, I don't remember that'. But instead you can tag it and then come back to it.* (P16)

> *Probably to remember one flight, keep that in mind and go to others and compare to that one that is already in the memory. And once I have found a better option than just change it until you are happy with your last choice.* (P21)

An **'Audio Bookmark'** feature was deemed useful by participants, as it could help users to reduce cognitive workload and make more informative choices. Some participants (n=12) commented that Displaying Results on a Screen would make the search process easier and more accurate than in a voice-only condition. These observations were in line with views on the Interaction Challenges (expressed in subsection 5.5.3.2) where participants complained about the lack of visual feedback. Some participants (n=5) mentioned that a CA should be able to provide **Additional Services** beyond a flight search, such as 'booking accommodation' and 'information on public transport'. We do not analyse these options as they are beyond the scope of our investigation which focuses solely on the flight booking functionality of a goal-oriented CA.

5.6 Discussion

In this section, we discuss the findings with regards to our research hypotheses concerning the impact of elicitation and revealment strategies. We hypothesised that compared to Passive and Listing agents, Active and Summarising agents would: be less cognitively taxing **(H2.1),** lead to higher satisfaction **(H2.2)**, lead to better performance **(H2.3)** and allow participants to complete their tasks quicker **(H2.4)**. Subsequently, we provide an answer to our **RQ2** i.e. **'How do different information elicitation (passive vs. active) and revealment (summary vs. listing) techniques impact user search experience in a goal-oriented task with regards to: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?** Finally, we consider the limitations of our study, offer directions for future work, and make CA design recommendations.

5.6.1 Impact of Conversational Strategy on Work Load

Although we found no significant differences in terms of an agent's impact on cognitive work load of participants (**lack of support for H2.1**) , there are some interesting themes emerging. We observe that (Section 5.4.1), overall Active CAs put less strain on a participant's cognitive workload (cf. Table 5.4.1, NASA TLX) . In terms of individual strategies, Passive Listing is worst in terms of Cognitive Load and perceived Usability compared to Active Listing which performs best. It can be argued that a Passive Listing strategy is the most strenuous agent, as participants need to decide on how to refine their search criteria without being provided with any hints from the CA. In the Active Listing condition, however, the algorithm supports the participant by offering refinement suggestions that reduce cognitive load. Our observations are in line with hypothesis put forward by Radlinski and Craswell [126] who posited that CAs should offer active elicitation support.

5.6.2 Impact of Conversation Strategy on Satisfaction

There were no significant differences in terms of satisfaction (**lack of support for H2.2**). However, overall, Active CAs were perceived as more usable (cf. Table 5.4.1, SUS). When it comes to satisfaction with the agent's support in exploring the search space, differences are marginal (see Table 5.3, SSES). As demonstrated in Section 5.4.1, Active agents match Passive agents across the board, except for the speed of results presentation, where Passive agents score higher (4.5/5 vs. 4/5).

5.6.3 Impact of Conversational Strategy on Performance

Despite no significant differences in data obtained from subjective measures, there is a noticeable difference in performance measures. For elicitation strategies, we observe that while using Active agents, participants were significantly more likely to select a flight that met the required travel

time and wasted significantly less money in the process (**partial support for H2.3**) . We also observe that participants who used Active agents selected more Pareto Optimal flights and lost less money in the process (see Table 5.4). This finding supports our H2.3 with regards to search performance and provides empirical evidence that validates the assertion (advocated in previous research [27,57,126]) that the active involvement of the CA can boost search performance. With regards to revealment strategies, Summarising CAs outperformed Listing CAs with all aspects but Meeting Time Preference. While using Summarising CAs, participants selected more Pareto Optimal flights and wasted less money. However, the differences were not statistically significant and therefore we cannot support our H2.3 with regards to revealment strategies.

5.6.4 Impact of Conversation Strategy on Interaction Metrics

Although there are no statistically significant differences regarding interaction metrics (**lack of support for H2.4**), Listing agents stand out in terms of the number of conversational turns e.g. the Active Listing agent has on average 17 turns, compared with 20 turns for the Active Summarising agent (see Table 5.6). This discrepancy can be explained by the fact that in the Summarising condition, CAs provide information in ranges, which urges participants to apply more filters to successively refine the search space. Listing agents, on the other hand, provide the cheapest and fastest options which, based on our observations, have led participants to continue their exploration in that direction. While using Listing agents, exploration was focused on clusters of the cheapest, or fastest flight options.

5.6.5 Perceptions of Conversational Strategies

Participants were generally in favour of results filtering suggestions, and praised active support and agent responsiveness. However, several participants (n=15) expressed the need to have more control over the scope of the presented results (limiting the number of presented flight attributes). Some participants (n=3) also emphasised the importance of being in control of the search process; stating that the active strategy 'imposed' the direction of the search by offering filtering cues (see Section 5.5.3 for more details). One of the most requested functionalities was the ability of the agent to synthesize the search results over multiple days (n=17), and remember flight information for future comparisons (n=11). The comments of the participants highlight the positive impact of an Active Elicitation strategy on user search experience. As for revealment, it could be argued that a modified version of a Summarising Revealment strategy (where options are aggregated over multiple dates) would be preferable.

5.6.6 Reflections on Findings

Based on the experience of observed conversations, we would like to offer several reflections on the potential reasons for our experimental findings.

Firstly, as indicated by the semi-structured interviews, many participants found the lack of visual feedback challenging (see Section 5.5.3.2 for details). This challenge was reflected in participants' search behaviour. We observed that a high cognitive workload, associated with the need to memorise and reason about multiple flight options, prevented participants from a detailed exploration of the search space across all available travel days, and effectively led to missing Pareto Optimal options. Some participants requested that CAs compare and contrast flight options across several days (see Section 5.5.3.3 for more details). However, we decided not to support this functionality in order to require participants to engage in longer interactions with the CAs and increase their exposure to different result presentation techniques. In future research, it would be interesting to explore if splitting the experiment across several booking sections with long breaks in between could help to improve participants' performance.

Secondly, we noticed that after competing their first search task several participants developed a preferred interaction strategy which they then tried to apply to all of the CAs. To some extent, this 'carry-over' effect prevented participants from trying different exploration techniques and made all agents appear more similar. One way to address this problem would be to change the experiment design to a between-subjects model, where each participant interacts with one agent only, or to split the experiment across several sessions. However, the decision to follow a within-subjects design was taken to ensure that each participant was exposed to all conversation strategies.

Finally, during the semi-structured interviews some participants indicated that they felt 'compelled' to follow the filtering suggestions offered by the Active CAs, and that the Passive CAs provided them with a higher sense of search autonomy (see comments of P21 and P24 in Section 5.5.3.1). While in general the active support of conversational agents was welcomed, during the interactions we observed that some participants found filtering suggestions provided by the Active CAs to be intrusive. This behaviour may be related to their unwillingness to offload search decisions [159], and because trust in competence of the automated interlocutor may vary between individuals [36].

5.7 Conclusions

We set out to examine the impact of an agent's conversational strategy on user search experience in a goal-oriented task. To that end, we conducted a user evaluation study where participants completed a series of search tasks in a flight domain. Our results indicate that participants' subjective perceptions of different CAs do not differ much in terms of workload, perceived usability, and satisfaction with exploration of search space. There is, however, a significant

difference in performance when it comes to choosing the best flight options and wasting less time when using Active agents.

Our findings indicate that future generations of goal-oriented CAs should ask users questions to actively support narrowing down the search space and offer a possibility to dynamically manipulate the scope of the query as well as switch the focus between different aspects of the options presented.

Coming back to our research question: **'How do different information elicitation (passive vs. active) and revealment (summary vs. listing) techniques impact user search experience in a goal-oriented task with regards to: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?'** We demonstrated that even though the difference between CA strategies was not reported as subjectively different by participants, there were clear differences in search outcomes. While we have only examined the impact of conversational strategies within a specific context, we have shown that the strategy does impact upon performance, and we believe that it is likely to be similar in other contexts. While we only consider pure conversational strategies (rather than their combinations), our work motivates the need for agents to vary their strategies to adapt to the conversational context and user preferences.

Our study is one of the first empirical investigations of different elicitation and revealment strategies for voice based search. We provide insights on how users would interact with different audio-only CAs in a multi-turn, goal-oriented task, and elicit their expectations regarding such agents. Our findings on how to elicit and present information can help pave the way for more active search support and the development of more usable CAs in the future.

5.8 Chapter Summary

This chapter investigated the impact of different revealment and elicitation strategies on user search experience. We have shown that active elicitation and revealment can significantly improve user performance in a goal oriented task. In the next chapter we will explore if switching the conversational initiative further towards the agent (proactive involvement) can further benefit user search experience.

Chapter 6

Impact of Conversational Agent's Proactivity and Recommendations on User Search Experience

In Chapter 5, we have demonstrated the benefits of active search support on user search experience. Specifically, we have shown that a conversational agent showing increased initiative led to better performance in a goal-oriented task. In the current chapter, we focus on the impact of an agent's pro-activity by investigating if further increasing an agent's conversational initiative can yield even more benefits. While theoretical frameworks postulate that a higher conversational involvement of an agent is desirable (cf. [11, 154]), it comes at the expense of the user's agency as more control over the search process is delegated to the agent [16].

In the current chapter, we address our third research question, **RQ3: 'How do agents that proactively elicit search criteria (proactive elicitation) and proactively recommend search results that are outside of the original scope of the user's query (proactive recommendation) vary in terms of their impact on: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?'** For the purposes of comparison, we introduce four agents that vary in their elicitation (active and proactive) and revealment (no recommendation vs. recommendation) strategies. Proactive agents aim to reduce cognitive workload by helping users to elicit their search criteria (proactive elicitation) and recommend results that go beyond the user's original query (proactive revealment). Active agents provide less support with regards to elicitation (requiring the participant to apply their search criteria step-wise) and revealment (no additional results that go beyond the specific scope of a participant's query are recommended). A detailed presentation of all the agents featured in the current study is provided in Section 6.3.4.

Based on the anticipated benefits of proactive search support (advocated, among others, by [11, 154]), we hypothesise that compared to active and non-recommending agents, proactive and recommending agents will:

H3.1 : be less cognitively taxing - hypothesis linked to RQ3 (a),

H3.2 : lead to higher satisfaction - hypothesis linked to RQ3 (b),

H3.3 : lead to better performance - hypothesis linked to RQ3 (c) and

H3.4 : lead to faster task completion - hypothesis linked to RQ3 (d).

To examine the above hypotheses and investigate the potential benefit of proactive search support and recommending search results we conducted a Wizard of Oz (WOZ) [38]. Twenty-four participants undertook four goal-oriented search tasks interacting with four conversational agents (described in Section 6.3.4) that differed in the degree of their conversational involvement (active vs. proactive). Since the Active Summarising agent led to the best performance overall in Study 2, we decided to use it as a baseline in the current study. Participants' user search experience was evaluated based on metrics outlined in Section 3.4.
This chapter is structured as follows. We begin by providing background information on why increased pro activity of a system can be beneficial to a user (Section 6.1). Next, in Section 6.2 we discuss related work focusing on proactive system support in information retrieval. In Section 6.3, we provide information on our methodology, including our experimental procedure and information about our participants. We then present the results (Section 6.4), semi-structured interviews (Section 6.5) and discuss their implications (Section 6.6). Finally, in Section 6.7 we provide our conclusions and summarise the chapter (Section 6.8).

6.1 Importance

In conversational search, due to the transitory nature of speech, it may be desirable to delegate control over certain aspects of the search to the CA [154]. It is thus important to determine when the user should be involved in directing the search and when the CA should take control. Gaining a deeper understanding of the degree to which an agent's involvement in elicitation and revealment is required by users, can support development of more usable conversational agents.
Trippas noted that due to the high cost of delivering search results via voice, intermediaries (CAs) need to constantly make an assessment of what information it is appropriate to share with the seeker at different points in the search process [156]. This means that a conversational agent, in order to be an effective intermediary, needs to 'adapt, accommodate, and support the user so that the user has to expend as little effort as possible' [154, p.141]. Trippas further emphasised that the proactive agent has to 'adapt their information presentation techniques

based on a 'cost-benefit' trade-off and considering the impact on the seeker's satisfaction [154].' Switching initiative towards conversational agents can be beneficial to the user but it requires willingness to give up some control over the search process. For instance, in the context of a product search, while the user may find an acceptable option sooner when using a proactive agent, they may miss out on other suitable alternatives as the agent has control over which options are revealed.

In this chapter, we will investigate the impact of a CA's pro-activity in query formulation (elicitation) and results recommendation (revealment) on user search experience. More specifically, through the current study, we seek to test if proactive search support can yield the theorised benefits postulated in previous research [11, 126, 154] to provide better search support to the user in a goal-oriented task.

6.2 Context

The model of Spoken Conversational Search (SCS) proposed by Trippas [154] emphasises the importance of adaptability and proactivness in a conversational system. Previous work has shown that in human-human information seeking scenarios, seekers explicitly requested intermediaries to make search decisions on their behalf which was motivated by trust and competence in another human's abilities [156]. However, to the best of our knowledge, the willingness to delegate control over conversational search to a machine has not yet been explored in the spoken domain.

In the textual domain, the topic of proactive search support has been investigated by White and Ruthven, who used implicit and explicit feedback techniques to investigate the process of indicating the relevance of search results [172]. White and Ruthven investigated 3 different search systems that varied in the degree of control they had in creating queries, indicating which results were relevant in making search decisions. The study [172] indicated that while participants were willing to delegate the task of recommending potential key words to the system, they still wanted to retain the control of adding the keywords themselves. This finding highlights that proactive agent support may be detrimental to user search agency if the agent has too much control over the search process.

The notion of search agency and the willingness of users to retain control over the search process can be illustrated by the following quote by Bates:

> *Most people have a strong desire for a sense of effectiveness in and mastery of their environment, particularly with respect to things that affect them in a close and personal way. Control of tools or powerful machinery can touch deep issues of personal power and freedom. As computers are experienced more and more as commonplace personal utilities, I think we can expect to see the same urge for control over computer systems, including information retrieval systems, that we see with*

cars. In seeking to provide the convenience of a wholly automatic information search to users, we in information science, may unwittingly be robbing people of the power and freedom of choice that they want to keep for themselves. [16, p.19]

The argument made by Bates is pertinent to conversational search agents and demands closer attention given the growing popularity of voice interfaces. While theoretical frameworks of conversational search [11,126] assume that pro-active agents can enhance the spoken conversational search experience and be more user-friendly than their passive equivalents, this assertion has not been verified empirically. The current study addresses this gap by evaluating user search experience in a goal-oriented task.

6.3 Method

To investigate the impact of the agent's proactive search support (see Section 6.3.4 for details), we conducted a controlled laboratory study, using a WOZ methodology. Again, as in the studies described in Chapters 4 and 5, the current study was conducted in a flight booking context.

6.3.1 Study Design

Our study followed a 2x4 within-subjects design, where the independent variables were elicitation strategy (Active and Proactive) and revealment strategy (No Recommendation and Recommendation) and dependent variables were four aspects of user search experience: (1) cognitive load (measured by NASA TLX [69]), (2) satisfaction (measured by SUS questionnaire [24] and SSES questionnaire [176]), (3) performance (evaluated in terms of the selected flight meeting the search criteria outlined in the task) and (4) interaction time for each of the search scenarios (measured in seconds) and number of conversational turns.

6.3.2 Experimental Procedure

The experimental stages of Study 3 are illustrated in the Figure 6.1

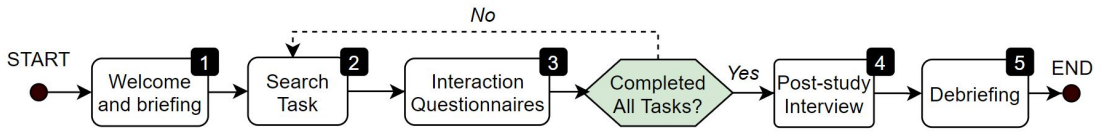


Figure 6.1: Illustration of Experimental Stages. Stage (2) consists of four search sessions (one with each CA); each session was followed by completing NASA TLX [69], SUS [24], and SSES [176] questionnaires.

Ethics approval for the study was granted by the Department of Computer and Information Sciences, University of Strathclyde (application no. 1155). At the beginning (Stage 1), participants were briefed about the study, asked to sign the participant form, fill in the demographics questionnaire and allowed to ask any study-related questions. Next, in Stage 2, participants

interacted with one of 4 Agents - the sequence of agents was manipulated using Latin Square design. All interactions were followed by questionnaires (Stage 3). The aim of the questionnaires was to assess: workload (NASA TLX [68]), perceived satisfaction ((SUS [24]) and (SSES [176])). In Stage 4, participants took part in a semi-structured interview where they were asked to retrospectively comment on their flight choices, evaluate their search experience and rank the agents. Participants were asked four open questions[1]: (1) 'What difference did you notice between the agents, if any', (2) 'Which agent(s) did you find most useful and why?' (Positive Features of Interaction), (3) 'Which agent(s) did you find least useful and why?' (Interaction Challenges) (4) 'How would you improve the agent/agents?' (Suggestions for New Functionalities). To facilitate retrospective reflection and provide performance feedback, at the end of each semi-structured interview, we provided participants with a visualisation of the search space with their flight choices marked for each scenario. Since the study was deployed online, the flight choices were highlighted by using Zoom's annotate function as illustrated in Figure 6.2.

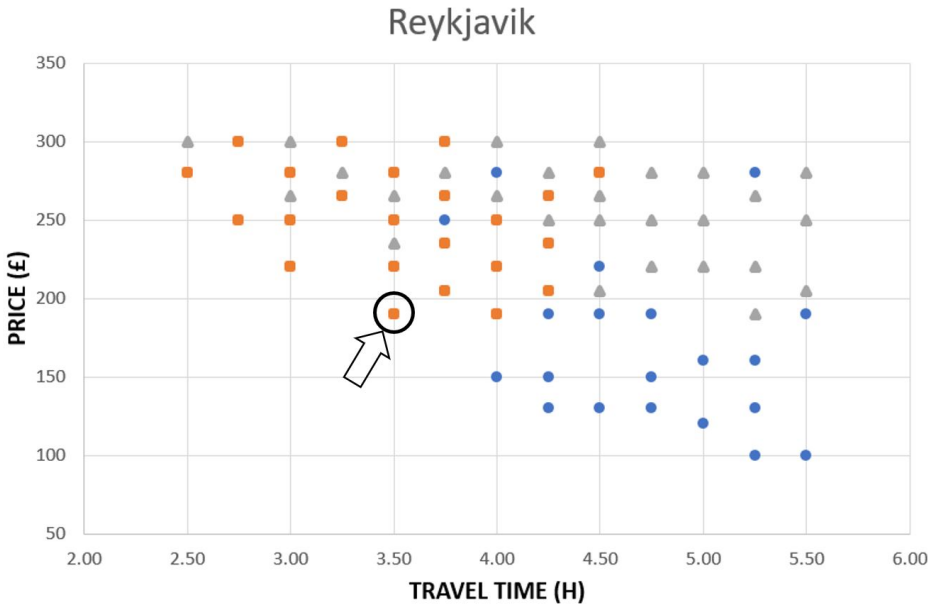


Figure 6.2: Visualisation of the search space presented to participants during Zoom call. Black circle indicates the flight selected by the participant. The arrow has been added for emphasis.

6.3.3 Wizard/ Agent Setup

Our user evaluation utilised a Wizard of Oz (WOZ) set-up [38] in which the CAs were simulated by a human (wizard) who interacted with a user and provided the user with results via voice (an illustration of this setup is presented in Figure 6.3). This setup is an online version of the setup used in Study 2 (see Section 5.3.3 for details) with the main investigator acting as a wizard. The modifications were necessary as the COVID-19 pandemic made face-to-face studies prohibitive

[1]The questions are modified versions of questions used in Chapter 5. The decisions to modify the questions was to focus discussion on merits and faults of different CAs.

(see 'University Ethics Committee Covid-19 Guidance' in Appendix D.2)[1]. The station used by the wizard during the experiment is presented in Figure 6.4.

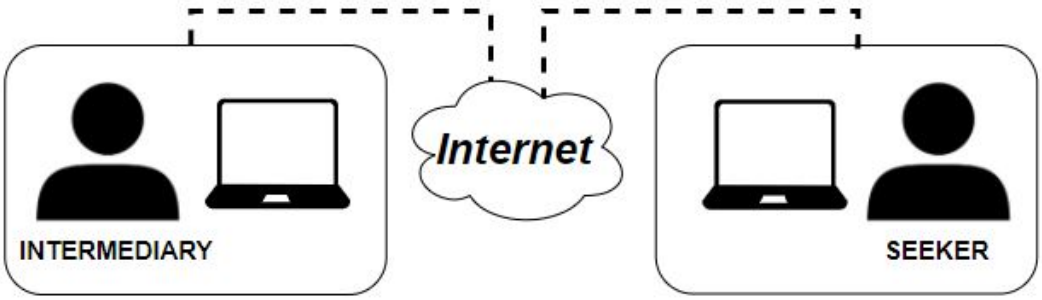


Figure 6.3: The Wizard of Oz framework: a wizard searches a flight database on behalf of a user and provides them with results.

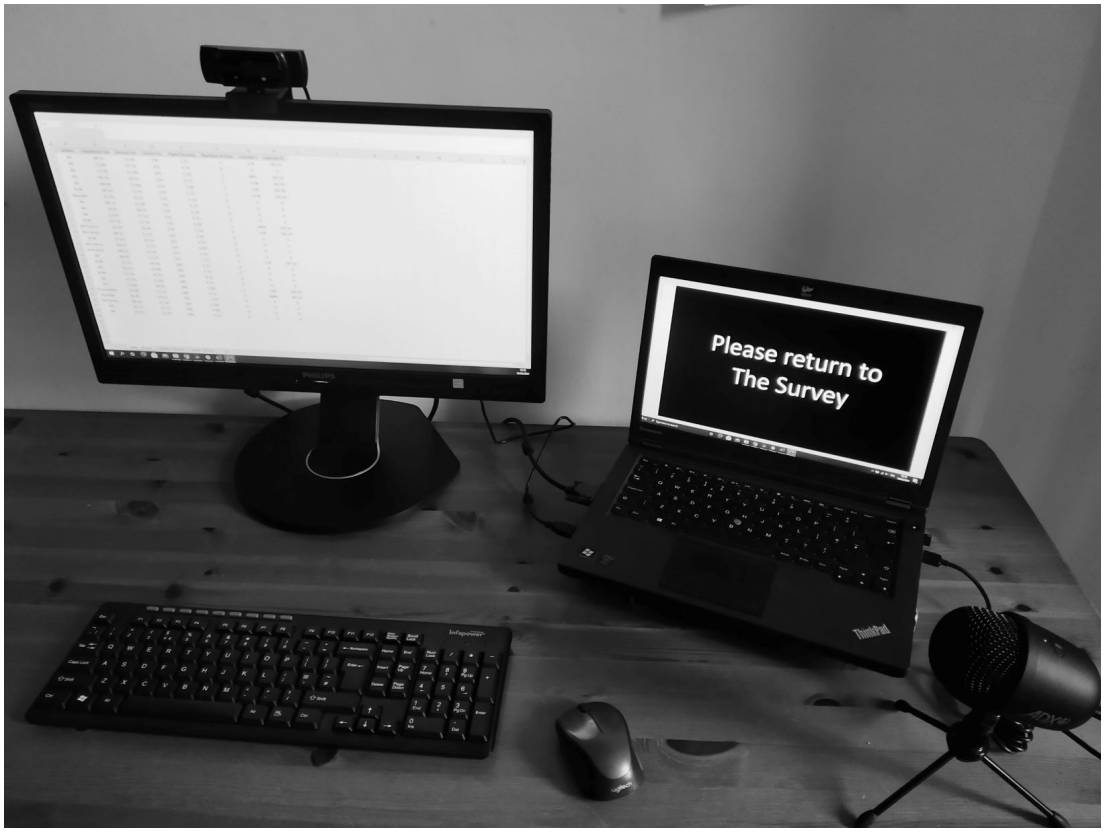


Figure 6.4: Wizard's station with two screens. During the experiment one screen was used for filtering the flights database (left) and the other for sharing instructions with participants through the Zoom 'screen share' function.

[1]Please note that the guidance is updated as the situation with Covid-19 develops. The regulations can be accessed via: https://www.strath.ac.uk/coronavirus/staff/universityethicscommittee/ (last accessed on 26th October 2020)

6.3.4 Agent Conversational Strategies

Four conversational agents featured in this study are presented in Figure 6.5. As in Study 2, we assigned a name to each of the conversational agents to make them easily identifiable to the participants, the names of the agents were:

1. **Agent Calum:** Active Summarising (AS)
2. **Agent Euan:** Active Recommendation (AR)
3. **Agent Frank:** Proactive Listing (ProL)
4. **Agent Graham:** Proactive Recommendation (ProR)

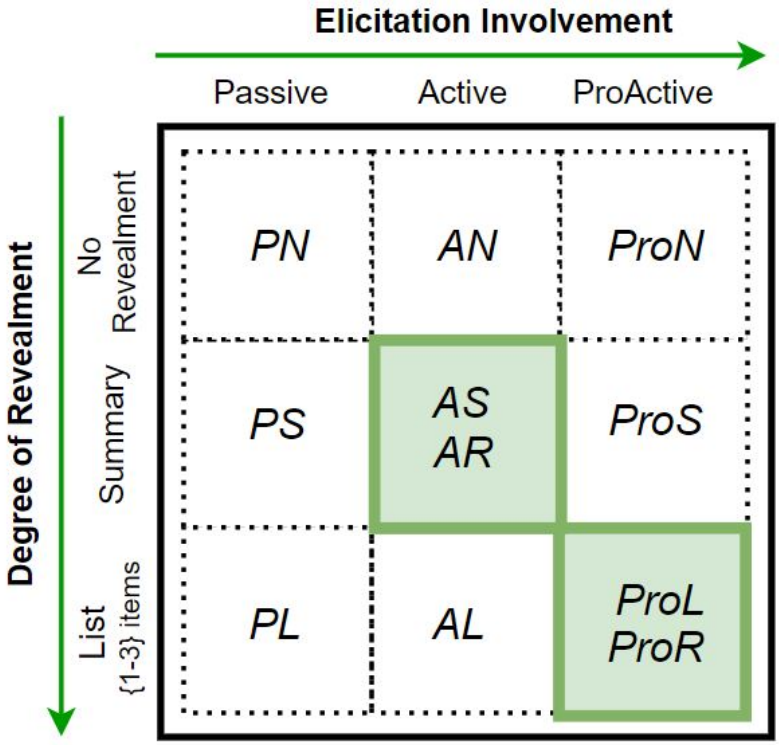


Figure 6.5: Conversational Agents Used in Study 3.

For the purposes of our study, a CA is considered 'pro-active' if, given the context of the initial query provided by the user at the beginning of interaction, the agent, before providing results, will (1) help to narrow down the query and determine search priorities (most salient flight attributes), and (2) make recommendations by providing additional results outside the original query in order to help the participant save money and/or time. The goal of the pro-active agent is to quickly reduce the size of the search space (considered in terms of flight options that are relevant to the user) at the elicitation stage and to offer alternative results during the revealment stage.

In the current study the level of agent pro-activity varied. An agent could be proactive both in terms of elicitation and revealment (ProR), elicitation only (ProL), revealment only (AR) or neither (AS). Conversational strategies are presented in Algorithm 3 (Active Summarising), Algorithm 5 (Active Recommendation), Algorithm 6 (Proactive Listing) and Algorithm 7 (ProActive Recommendation). It should be noted that the AS agent was used as a baseline as it led to the best performance in Study 2. We decided not to implement the summary option in the proactive revealment condition to make the agents more easily distinguishable for participants. Moreover, since the goal of proactive elicitation is to quickly identify the most salient flight options without overloading participants with too many details, implementing it as a listing strategy was more adequate. A more detailed illustration of how CAs elicited and revealed information is presented in Section 6.3.6.

6.3.5 Simulated Search Tasks

Search tasks used in the current study were based on the same structure as in Study 2 (see Section 5.3.5 for details). The only modification were different background stories used to motivate the search. The search scenarios used in the current study are provided for reference in Appendix E.3. During the search tasks the description of the search scenario was provided to participants via the Zoom screen-share feature, as illustrated in Figure 6.6.

Figure 6.6: Screen shared with participants during search task.

6.3.6 Interaction Design

The interaction strategy of the pro-active conversational agent is presented in Figure 6.7. For comparison, the conversation flow of an Active CA (baseline) is presented in Figure 6.8.

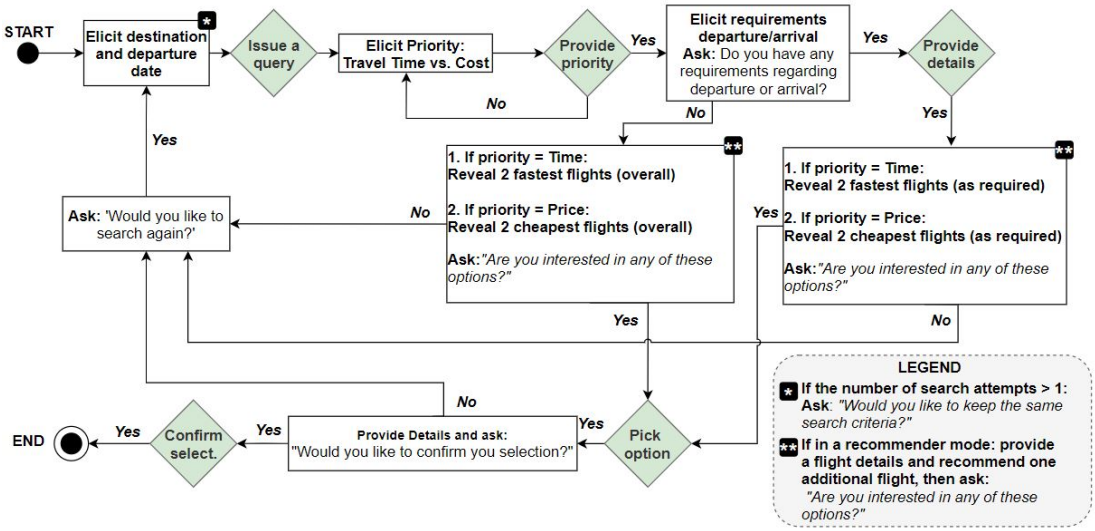


Figure 6.7: Overview of conversational flow for Pro-Active CA. White rectangles denote the CA's actions while green diamonds denote the user's actions. Note: in a recommender mode ('**') the CA provides one flight result and recommends one additional alternative.

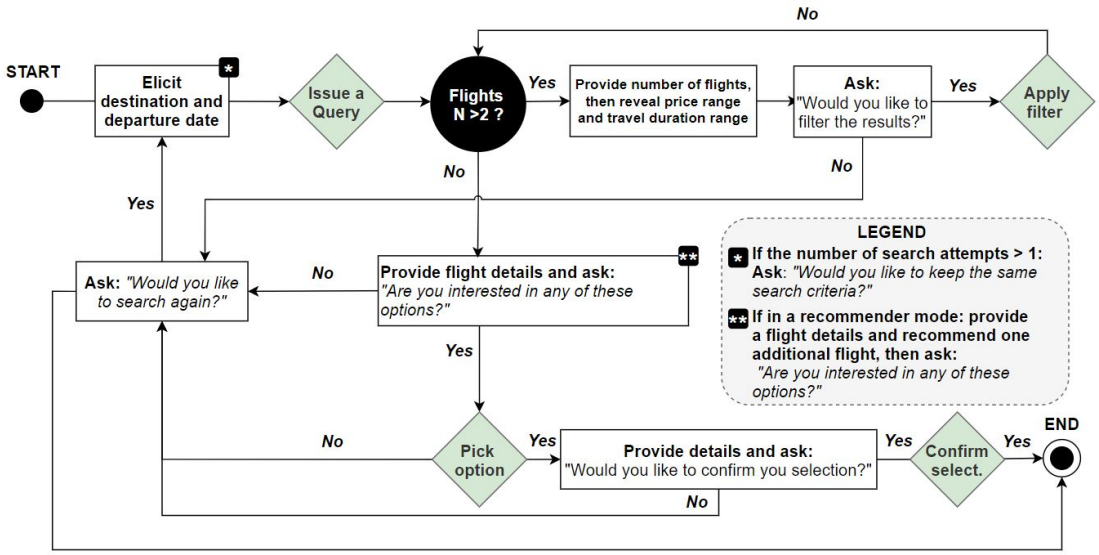


Figure 6.8: Overview of conversational flow for an Active CA. White rectangles denote the CA's actions while green diamonds denote the user's actions. Note: in a recommender mode ('**') the CA provides one flight result and recommends one additional alternative.

To ensure consistency of interaction, as in Study 2, each agent was based on an algorithm that specified how to elicit and reveal information from and to the participant. The four algorithms used in Study 3 are provided below. Every interaction began with the agent prompting the

participant to provide their destination and day of travel. With regards to elicitation, the main difference between active (AS and AR) and proactive (ProL and ProR) agents is that while the former provide the participant with an overview of the search space each time a search filter is applied and the number of flight options is larger than 2 (see Algorithm 3 and Algorithm 5, lines 5-7), the latter ask a series of questions to elicit the participant's priority before providing a detailed result (see Algorithm 6 and Algorithm 7, lines 6 & 7). Proactive agents required participants to provide their search priority (e.g. flight cost over duration of travel). If the participants did not indicate the priority the agent would inform them: 'You need to pick one priority in order to continue your search'.

ALGORITHM 5: Active Recommendation Strategy - **Agent Euan**

```
1  while Flight has not been selected do
2      wait for a query
3      for a query do
4          if n of flights > 1, then
5              provide n of flights
6              provide flights range (price)
7              provide flights range (travel duration)
8              Case 1. say: 'Would you like to filter by price?'
9              Case 2. say: 'Would you like to filter by duration?'
10             Case 3. say: 'Would you like to filter by departure time?'
11         else
12             provide detailed flight result (airline, departure and arrival times, price and
                duration)
13             provide one alternative flight (outside the scope of the original query)
14             say: 'Would you like to this flight/any of these flights?'
15         end
16     end
17 end
```

ALGORITHM 6: ProActive Listing Strategy - **Agent Frank**

```
1  while Flight has not been selected do
2      wait for a query
3      for a query do
4          if n of flights > 2, then
5              Elicit Priority
6              Step 1. say: 'Is your priority cost of the flight or duration of travel?'
7              Step 2. say: 'Do you have any requirement regarding departure/arrival
                time?'
8          else
9              provide detailed flight result (airline, departure and arrival times, price and
                duration)
10             say: 'Would you like to this flight/any of these flights?'
11         end
12     end
13 end
```

ALGORITHM 7: Active Recommendation Strategy - **Agent Graham**

```
1  while Flight has not been selected do
2      wait for a query
3      for a query do
4          if n of flights > 1, then
5              Elicit Priority
6              Step 1. say: 'Is your priority cost of the flight or duration of travel?'
7              Step 2. say: 'Do you have any requirement regarding departure/arrival
                time?'
8          else
9              provide detailed flight result (airline, departure and arrival times, price and
                duration)
10             provide one alternative flight (outside the scope of the original query)
11             say: 'Would you like to this flight/any of these flights?'
12         end
13     end
14 end
```

Active and Proactive strategies were implemented in either Recommendation or No Recommendation mode. While No Recommendation Agents provided two full flight results each time that the number of search results was less or equal to 2, Recommendation Agents provided only one result and offered participants an alternative that highlighted a certain trade-off, e.g. 'If you spend £30 more you can get to your destination an hour quicker.'

In order to ensure consistency across the strategies we created a list of exceptions that particular agents could not handle. The exceptions were:

- Searching multiple days at the same time
- Comparing results
- Asking No Recommendation agents for Recommendations

If a participant issued any of the above queries, the CA replied: 'Sorry, this functionality is not currently supported.'

Participants were however allowed to ask agents to re-apply the same search criteria between different days. When switching between the days, the agents asked the participant if they would like to use the same filters from their most recent search.

As in Study 2, for Active agents if a query returned no results (Ø), the agent would suggest changing the filtering criteria. For instance, if there were no flights within the specified price range, the agent would suggest relaxing the search criteria (e.g. 'There are no flights for less than £200 that leave before 2 pm, try to increase your price, change travel time' etc.). If the query issued by the participant was outwith the scope of the agent, it would inform the user that the given functionality was not supported. The agent could search only one day at the time so that participants had to perform a number of searches to explore the space (which

is consistent with current flight booking systems). Examples of participants' interactions with agents are provided for reference in Table 6.1 (Active Agents) and Table 6.2 (Proactive Agents).

Table 6.1: Interactions with Active Agents. Note: The passage that distinguishes a Recommendation from a Summary interaction strategy is underlined for clarity.

Active Summary Agent (AS)	**Active Recommendation Agent (AR)**
AS: Hello my name is Calum, I am a flight search agent and I will help you to find your flight, to get started, please provide me with your destination and day of departure.	
Participant: Hi Calum, the destination is London and the date of departure is between Wednesday the twelfth and Friday the fourteenth of December.	**AR:** Hello my name is Euan, I am a flight search agent and I will help you to find your flight, to get started, please provide me with your destination and day of departure.
AS: Ok, which day would you like to check first?	
Participant: Wednesday.	**Participant:** I want to go to Reykjavik and I am flying on the sixth of April.
AS: One moment please. There are twenty-nine flights available. The prices range from seventy pounds to one hundred and forty, and the durations range from one hour thirty to four hours fifteen. Would you like to filter by price?	**AR:** Sure, one moment. Okay, so we have twenty seven flights on that day. Prices range from one hundred ninety to three hundred pounds and flight durations range from two and a half hours to five and a half hours. Would you like to filter that by price?
Participant: Yes.	**Participant:** Can we filter by time? Please get rid of anything that arrives after three thirty pm.
AS: Ok, how would you like to filter?	**AR:** We have six flights in that range. Prices range from one hundred ninety to three hundred pounds and flight durations range from two and a half hours to five and a half hours. Would you like to filter that by price?
Participant: Cheapest first, please.	**Participant:** Yes, how long does the cheapest flight take?
AS: One moment please. Do you have a maximum price in mind?	**AR:** That flight takes five hours fifteen minutes and it costs one hundred ninety pounds. <u>However, if you spend thirty pounds more you can get there half an hour quicker.</u>
Participant: Ninety.	**Participant:** OK, could you do the same search for the next day? [...]
AS: There are six flights available. The prices range from seventy pounds to ninety pounds and the durations range from three hours to four hours. Would you like to filter by duration?	
Participant: No, can I filter by arrival time?	
AS: Ok, how would you like to filter?	
Participant: Arrive at or before ten am.	
AS: There are no flights that arrive at or before ten am. You may want to increase the arrival time.	
Participant: Are there any that arrive at or before eleven am?	
AS: No. you may want to increase your arrival time.	
Participant: Can I try Thursday? [. . .]	

Table 6.2: Interactions with Proactive Agents. Note: Passages that demonstrate Proactive elicitation are underlined for clarity.

Proactive Listing Agent (ProL)	**Proactive Recommendation Agent (ProR)**
ProL: Hello, my name is Frank. I am flight search agent and I will help you to find your flight. To get started, please provide me with your destination and day of departure.	
	ProR: Hello my name is Graham, I am a flight search agent and I will help you to find your flight, to get started, please provide me with your destination and day of departure.
Participant: I would like to book a flight for Las Palmas in the Canary Islands, Spain. I am hoping to fly on Wednesday, the eighth of July.	
	Participant: Hi, I am looking to book a flight to Reykjavik, Iceland on Friday the sixth of April, please.
ProL: Sure, are you more concerned about the cost of the flight or about the duration of travel?	
	ProR: Sure, are you more concerned about the cost of the flight or about the duration of travel?
Participant: The cost of the flight.	
	Participant: Duration of travel.
ProL: Ok, do you have any requirements regarding departure or arrival time?	
	ProR: Ok, do you have any requirements regarding departure or arrival time?
Participant: I would like to arrive before six pm.	
	Participant: It would be ideal to arrive before four pm, please.
ProL: Ok, one moment please.	
The two cheapest flights that arrive before six pm, are as follows. The first one is a KLM flight, departure six fifty am, arrival two zero five pm, it costs three hundred seventy pounds and takes seven hours and fifteen minutes. The second option is a BA flight, departure time ten am and arrival four-forty five pm, it costs three hundred seventy pounds and takes six hours forty-five minutes. Would you like to select any of these flights?	**ProR:** Okay, so the quickest flight on that day that arrives before four pm, is Air France departure time eight fifteen am, arrival eleven fifteen am it costs two hundred and sixty five pounds and it takes three hours. However, if you are willing to spend forty-five minutes longer travelling, you can save fifteen pounds . Would you be interested in any of these flights?[...]
Participant: Yes, I would like to go for the first option. [...]	

6.3.7 Participants

24 people took part in the study (10M, 14F). The average age of participants was 26 (SD=3.8). There were 20 native English speakers, 3 people described their English level as Proficient and 1 person as Advanced. 20 participants reported having used Expedia or Skyscanner at least one time in last year. 8 participants reported having used a smart speaker or other form of personal voice assistant such as Amazon Echo at least once. 20 participants indicated price as the most important factor that they consider when booking flights. The data of one participant was excluded from the analysis as they did not follow the instructions outlined in the search scenarios.

Since the experiment was conducted online, via Zoom, after each search task we evaluated the audio quality by asking participants the statement 'It was easy to understand what the agent was saying' on the 5-point Likert scale, where 1 indicated 'strongly disagree' and 5 indicated 'strongly agree'. The average audio comprehension score was 4.6 (SD = 0.6) - indicating that the audio quality was good and therefore did not have a detrimental effect on participants' flight selections.

6.4 Results

Our experimental results are presented in Table 6.3 (Subjective Measures), Tables 6.4 and 6.5 (Performance Measures) and Table 6.8 (Interaction Metrics). Since most of our data was not normally distributed, for pair-wise comparisons, unless otherwise stated, we use the Kruskal Wallis H Test. All of the conducted statistical tests were two-tailed.

6.4.1 Relatability

In order to ensure that the type of the scenario did not have an impact on the level of participants' involvement in each search task, after each task, we asked participants to answer the following question: 'How relatable did you find the task?'. The responses were measured on a 5-point Likert scale, where 1 signified 'completely unrelatable' and 5 signified 'completely relatable'. The average, overall score across the tasks was 4.3 (SD=0.7) and there was little variance between the tasks: 4.2 (SD=0.7) for Scenario 1, 4.4 (SD = 0.8) for Scenario 2, 4.3 (SD=0.7) for Scenario 3 and 4.4 (SD=0.7) for Scenario 4.

6.4.2 Subjective Measures

A summary of the findings for subjective measures (NASA TLX, SUS and SSES questionnaires) can be found in Table 6.3. Pairwise comparisons between Active and ProActive agents, and No Recommendation and Recommendation agents revealed no statistically significant differences. Active and ProActive agents performed similarly for: Overall NASA TLX (Z = -0.648, p = 0.517), SUS (Z = -0.635, p = 0.526), Mental Demand (Z = -0.498, p = 0.618), Performance (Z = -0.224, p = 0.823), Frustration (Z = -0.810, p = 0.418). There was also little difference between No Recommendation and Recommendation Agents: Overall NASA TLX (Z = -0.613, p = 0.540), SUS (Z = -0.457, p = 0.648), Mental Demand (Z = -0.534, p = 0.593), Performance (Z = -0.213, p = 0.831), Frustration (Z = -0.962, p = 0.336).

We observe that, overall, Active CAs put more strain on a participant's workload. As can be seen in Table 6.3, the average overall NASA TLX score for Active agents was 29, which is four points more than for the ProActive agents (25). Notably, among all the agents, Active Recommendation performs the worst (NASA score of 33). An analogical trend can be observed for Mental Demand and Performance. In terms of usability, Active agents score lower (SUS score of 80) than any other agent group (SUS score of 82.5 each).

For SESS, the scores are uniform across the board. The exception is Speed of interaction where ProActive and Recommendation Agents score higher than Active and No Recommendation Agents (4.5 vs. 4 for each group respectively). This difference in scores could be attributed to the fact that ProActive agents elicited preference information upfront which reduced interaction time. Recommendation Agents presented one result and then recommended one alternative result, compared to two results (with no alternative) presented by No Recommendation Agents

Table 6.3: Subjective Measures. The table aggregates data from questionnaires that reflect participants' perception of Conversational Agents. Note: For NASA TLX the lower the score, the better. For SUS and SSES, the higher score the better.

[Metrics] \ [*Agent Type*]			*Active*	*Pro Active*	*No Rec*	*Rec*	*Active Summ*	*Active Rec*	*Pro Summ*	*Pro Rec*
NASA TLX Overall(0-100), Per Item(0-20)	Mental Demand	Med (IQR)	7.5 (8)	7 (6)	7 (6)	5 (7)	6 (8)	8 (10)	7 (6)	7 (6)
	Effort		6 (5.5)	5 (6)	5 (5)	6 (5)	5 (7)	6 (12)	5 (6)	5 (7)
	Performance		5 (7)	5 (5)	5 (6)	5 (7)	5 (5)	5 (8)	5 (5)	5 (5)
	Frustration		3 (6)	2 (5)	2 (4)	3 (6)	2 (4)	3 (6)	1 (4)	2 (6)
	Temporal Demand		4 (4)	4 (3)	4 (6)	4 (4)	4 (7)	4 (4)	4 (5)	4 (3)
	Overall		**29 (27)**	**25 (9.5)**	**24.5 (20)**	**27 (17)**	**27 (31)**	**33 (16)**	**24 (9)**	**26 (18)**
SUS	**Overall**	Med (IQR)	**80 (21)**	**82.5 (9)**	**82.5 (19)**	**82.5 (14)**	**80 (22.5)**	**80 (20)**	**82.5 (17.5)**	**82.5 (5)**
SSES Overall(0-20), INDV(1-5)	Presentation	Med (IQR)	4 (1)	4 (1)	4 (1)	4 (1)	4 (1)	4 (1)	4 (1)	4 (1)
	Overview		4 (0.5)	4 (1)	4 (1)	4 (0.5)	4 (0.5)	4 (1)	4 (1)	4 (0.5)
	Confidence		3 (2)	3 (2)	3 (2)	3 (2)	3 (2)	3 (2)	3 (2)	3 (2)
	Speed		4 (1)	4.5 (1)	4 (1)	4.5 (1)	4 (1)	4 (1)	4 (1)	5 (1)
	Overall		**16 (3)**	**16 (2.5)**	**16 (3)**	**16 (2)**	**16 (3)**	**16 (2.5)**	**16 (3.25)**	**16 (2.25)**

which was perceived as a faster mode of delivery. In particular ProActive Recommendation Strategy scored higher (5 points) than any other strategy (4 points).

6.4.3 Objective Metrics - Task Performance and Interaction Metrics

Objective metrics used to evaluate all CAs were: (1) participants' performance considered in terms of selected flight options (primary indicators) and expenditure of money and time (secondary indicators); and (2) interaction metrics - task completion time and number of conversational turns. Both metrics are described in turn in this section.

6.4.3.1 Task Performance

We measured task performance in terms of distance of the selected flight from the Pareto frontier (Discussed in Chapter 3 , Figure 3.5) - hard constraint, and in terms of preference (desired travel time specified in search scenario) - soft constraint. The flight was considered Pareto Optimal if, and only if, it met both of the constraints. In terms of the task outcome; for each conversational strategy, we also report if participants lost money and/or wasted time by selecting sub-optimal flight options.

Task performance measures are presented in Table 6.4 and Table 6.5. We consider Task Performance in terms of Primary Indicators ('Meeting Time Preference' and 'Hitting the Pareto Optimal), and Secondary Indicators ('Losing Money' and 'Wasting Time'). Primary Indicators

concern meeting the flight arrival requirements specified in each search scenario and indicate if the selected flight was Pareto Optimal (i.e. it offered the best combination of price and duration while meeting the arrival requirement). Secondary Indicators concern the impact of participants' flight selections on their resources: time and money (i.e. if a participant selected a more expensive flight when a cheaper one of the same duration was available - they lost money; if they selected a flight with a longer duration when a shorter alternative was available at the same price – they lost time). Since all the indicators are considered in binary categories, i.e. a selected flight is either on the Pareto frontier or not, the selected flight either meets the time preference or not etc., Cochran's Q Test was used to compare different conversational strategies. The Bonferroni adjusted alpha-level (.008) was used for all post-hoc analyses.

Table 6.4: Performance Measures (Primary Indicators).

[Metrics] \ [*Agent Type*]		*Active*	*Pro Active*	*No*Rec	*Rec*	*Active Summ.*	*Active Rec*	*Pro Summ.*	Pro Rec
PRIMARY INDICATORS	Met Time Preference	42/46 (91%)	45/46 (98%)	42/46 (91%)	45/46 (98%)	20/23 (87%)	22/23 (96%)	22/23 (96%)	23/23 (100%)
	Pareto Optimal	28/46 (61%)	34/46 (74%)	27/46 (59%)	35/46 (76%)	12/23 (52%)	16/23 (70%)	15/23 (65%)	19/23 (83%)

Meeting Time Preference: There was no statistically significant difference between the conversational strategies (Cochran Q = 5.400, p = 0.145). However, participants obtained better results when using ProActive and Recommender agents than when using Active and No Recommender agents. Only participants who interacted with a ProActive Recommendation Agent managed to meet the time requirement at every attempt.

Hitting the Pareto Optimal: There is no statistically significant difference between strategies (Cochran Q = 7.317 , p = 0.062). Again, participants using a ProActive and Recommender Agents performed better than when using Active and No Recommender agents (74 % and 76 % for the former compared to 61 % and 59 % for the latter). Overall, the Proactive Recommendation Agent led to the best performance with participants booking 83% Pareto flights.

Table 6.5: Performance Measures (Secondary Indicators).

[Metrics] \ [*Agent Type*]		*Active*	*Pro Active*	*No*Rec	*Rec*	*Active Summ.*	*Active Rec*	*Pro Summ.*	*Pro Rec*
SECONDARY INDICATORS	Money Lost	16/46 (35%)	9/46 (20%)	16/46 (35%)	9/46 (20%)	10/23 (43%)	6/23 (26%)	6/23 (26%)	3/23 (13%)
	Time Wasted	13/46 (28%)	11/46 (24%)	15/46 (33%)	9/46 (20%)	8/23 (35%)	5/23 (22%)	7/23 (30%)	4/23 (17%)

Money Lost: No statistically significant differences were observed between the conversational strategies (Cochran Q = 7.171 , p = 0.067). However, ProActive and Recommender led to better results with only 20% of participants losing money for each of the strategies, compared to 35% for both Active and No Recommender strategies each.

Time Wasted: We did not observe a statistically significant differences between the conversational strategies (Cochran Q = 3.740, p = .291). Fewer participants wasted travel time when interacting with Proactive and Recommender Agents (24% and 20% respectively) as compared to Active and No Recommender agents (28% and 33% respectively). Overall, the Proactive Recommendation Agent led to the least amount of travel time wasted (17%).

Overall, in terms of performance, the ProActive conversational strategy consistently outperformed the Active conversational strategy for all performance aspects under consideration. An analogical trend can be observed for the Recommendation strategy which outperforms the No Recommendation strategy for all performance measure. At the level of the individual CAs, the Proactive Recommendation agent yields the best performance for both Primary and Secondary performance indicators. The most notable differences are observed when it comes to 'Money Lost' and 'Time Wasted'.

6.4.3.2 Recommendation - Performance comparison.

Tables 6.6 and 6.7 provide a detailed breakdown of participants' performance when using agents with a Recommendation strategy (Table 6.6), compared with participants who used agents with a No Recommendation (Table 6.7).

Table 6.6: Interaction Statistics for Agents with Recommendation.

Participants who	# Participants	# Got Pareto	Interaction Time in seconds (SD)	# Turns (SD)
Followed Recommendation	19/46 (41%)	13/19 (68%)	303 (104)	14 (3.8)
Did not Follow Recommendation	27/46 (59%)	22/27 (81%)	280 (107)	13 (5.8)
Participants who	# Participants	# Missed Pareto	Interaction Time in seconds (SD)	# Turns (SD)
Followed Recommendation	19/46 (41%)	6/19 (32%)	183 (70)	10 (3.8)
Did not Follow Recommendation	27/46 (59%)	5/27 (19%)	359 (132)	17 (6.4)

Table 6.7: Interaction Statistics for No Recommendation Agents

Type of Agent	# Participants who Got Pareto	Interaction time in seconds (SD)	# Turns
No Recommendation	27/46 (59%)	266 (106)	13 (6)
Type of Agent	# Participants Missed Pareto	Interaction time in seconds (SD)	# Turns
No Recommendation	19/46 (41%)	260 (154)	12 (9)

It is noteworthy that participants who followed the recommendation and missed the Pareto, made their flight selections in almost half the time that it took participants who followed

recommendation and got the Pareto option (see Table 6.6 for details). All the participants who followed the recommendation and missed Pareto explored only one day, effectively missing better options (see Figure 3.6 for comparison of flight options across different days). This was not the case for agents that did not offer recommendation - as participants interacting with them spent roughly the same time interacting, regardless if they missed the Pareto or not (see Table 6.7 for details).

6.4.3.3 Interaction Metrics

Interaction Metrics, summarised in Table 6.8, provide information about the impact of conversational strategy on the duration of a conversation and the number of conversational exchanges between the agent and the participant. We considered conversation time and the number of turns as indicators of the agent's efficiency.

Table 6.8: Interaction Metrics. Note: Interaction Time is rounded up to the nearest second. '***' signifies $p < .001$,'**' signifies $p < .01$.

[*Agent Type*] / [Metrics]		*Active*	*Pro Active*	*No Rec*	*Rec*	*Active Summ.*	*Active Rec*	*Pro Summ.*	*Pro Rec*
Interaction Time in sec.	Med (IQR)	273 (149)	**229 (135)*** **	249 (144)	271 (149)	255 (99)	304 (158)	239 (134)	217 (152)
Conversational Turns		13 (7)	**10 (5)** **	10.5 (6)	12 (7)	12 (7)	14 (6)	10 (6)	11 (5)

We found statistically significant differences for objective measures between Active and ProActive agents with regards to interaction time ($Z = 6.4$ $p = 0.0011$) and number of conversational turns ($Z = 8.6$ $p = 0.003$). We also found statistically significant differences for the No Recommendation vs. Recommendation comparison with regards to interaction time ($Z = -2.674$, $p = 0.008$) and conversational turns ($Z = -2.834$, $p = 0.005$). Among all the strategies, Active Recommendation stood out with the highest number of conversational turns (Med = 14) and the longest interaction time (Med = 304s).

6.5 Semi-structured Interviews

All but one participant (n=23) took part in semi-structured interviews after all the interactive tasks were completed. The interviews provided a number of insights on participants' perception of conversation strategies (represented by four CAs - described in Section 6.3.4).

6.5.1 Procedure

All interviews were audio recorded. We asked participants four questions:

1. What difference did you notice between the agents, if any?
2. Which agent(s) did you find most useful and why? (Positive Features of Interaction)
3. Which agent(s) did you find least useful and why? (Interaction Challenges)
4. How would you improve the agent(s)? (Suggestions for New Functionalities).

6.5.2 Findings

In the current sub-section, we summarise the findings of our thematic analysis with regards to three categories , namely (1) Positive Features of Interaction, (2) Interaction Challenges and (3) Suggestions for New Functionalities. We refer to the interview participants by ID numbers (IDs: P1-P23).

12 out of 23 participants were able to see the difference between the agents. Table 6.9 provides a summary of the agents perceived by participants as the best and worst. The following sections will provide more detail on participants' opinions regarding the agents.

Table 6.9: Participants' perceptions of agents.

Agent (Type)	Perceived as Best	Perceived as Worst
Calum (Active Summary)	2	5
Euan (Active Recommendation)	2	3
Frank (ProActive Listing)	6	3
Graham (ProActive Recommendation)	4	1

6.5.2.1 Positive Features of Interaction

In total 14 participants commented about positive features of their interaction with conversational agents. The majority of participants (10/14) preferred Agents that elicit search criteria upfront before revealing the results (ProActive Agents). The main reasons provided in support of this preference were convenience of access and less cognitive effort. The preference for ProActive Agents is illustrated by the quotes below:

> *The ones that gave you all information up front. It was the easiest one because it just told you all things.* (P9)

> *They just provided more detail without me having to ask.* (P18)

On the other hand, some participants (4/14) preferred the Active Agents as they provided a better overview of available options and provided broader context that allowed participants to make more informed decisions. The quotations below illustrate participants' preference towards Active Agents.

> *I liked agents who provided a range of prices. Before you start filtering down, it is also good to know the minimum duration of the flight and the maximum duration. The agents who did not do it said that the flight was seven hours, but I didn't know it that was quick or not. And so, I think it would be good if it gave minimums and*

maximums before you start filtering down so that the user knows whether it's a long time or not. (P3)

I preferred to have filters available as it gave a better overview of what was available in the search space. It made my search more specific. (P4)

6.5.2.2 Interaction Challenges

Twelve participants commented on challenges encountered during their interaction with conversational agents. Most participants (8/12) complained about Active agents due to their lack of support with preference elicitation - as illustrated by the following quotations.

I did not like Calum because it did not ask me for my required arrival time (P3)

Calum, it was the only one that I really would not choose to use for buying tickets, I felt suspicious of it. (P9)

I did not like Euan and Calum because they provided the least amount of extra information. So for example, they asked for my requirements and that was easier to provide upfront. Just right off the bat, I was able to find something at 4pm, I could give them all my details up front. Whereas in the other ones it was a step by step process, so it felt a little more arduous to have to say: 'I want this day', and then you provide time, and then some more questions. (P18)

Several participants (4) also mentioned that the Recommendation strategy placed an additional cognitive strain on them, making them reconsider the flight selections that they had already made. This is illustrated by the following quotation.

Recommending agents put more strain on your memory. I think for me, it was the one that was suggesting the second best option like slightly more expensive option with a better time. It just made me think about more things. I was lost because it was suggesting when I already have made my decision. I felt I am going to stick with this one and then he came and said, 'Oh but if you pay a bit more and then you arrive sooner'. (P10)

This was a bit confusing because it just made me think about more things and added more effort. Once I had made my choice, the agents (Recommending) suggested something and then I had to think about it again and memorise options. (P10)

6.5.2.3 Suggestions for New Functionalities

Participants made several suggestions for improving the conversational agents. The most recommended functionality was for the agent to be able to provide a summary of results across several days.

I always write down what flights I have found on what different days and then I compare the price and the duration. But if you were doing it exclusively through voice, you should be able to save the flight that they have given you so that you can listen back to it, and then you are able to compare easily amongst other options. (P3)

It would be good to have a running summary of results - a global comparison. After looking into different days, the agent would say: 'So, overall, the cheapest flight is on that day and the quickest flight is on that day.' That would help me to remember things without having to ask. (P5)

Participants also requested that the agent provide a comparison between different options that they selected and highlight the trade-offs involved for each choice.

The agent should be able to highlight trade-offs between different options, rather than just providing me with the cheapest one or the shortest one. (P14)

It would be good to have feedback on different options and how they compare. If I am looking on multiple days, I want to know the cheapest option on that day and want to be able to have some way of saying it when I am exploring other other days so I can then compare it. (P6)

One participant also mentioned that they would like the agent to sound more robotic. This would make them less compelled to make a quick decision and probably allow a more thorough exploration of the search space.

It may sound strange, but it's like if this sounds a bit more like a robot, I would feel better with it because this one just felt so human and so real. I felt like, is this person waiting for me to make the decision, should I make my mind up quicker? Whereas with a more robotic voice, if it was something that did not feel like a real person that would just make me feel more comfortable and I would take more time to make my choices (P10)

Another participant suggested that voice-only interaction with conversational agent should be an initial, browsing stage whose purpose would be collecting information on available flights. The follow-up would be for the agent to forward the summary of results to the user so they could compare the results using a graphical interface before making their final decision.

It would be good if the agent could send me a summary of flights that I have selected to my phone for comparison. So, if there were three days, then I could see what I have picked for each day and I knew what my options were. I think that seeing results is important before you book any of them. (P21)

6.6 Discussion

In this section, we discuss our findings with regards to our research hypotheses concerning the impact of Proactive Elicitation and Proactive Revealment (recommendation). We hypothesised that compared to Active and No Recommendation agents, Proactive and Recommending agents will: be less cognitively taxing **(H3.1),** lead to higher satisfaction **(H3.2)**, lead to better performance **(H3.3)** and allow participants to complete their tasks quicker **(H3.4)**. Subsequently, we provide an answer to our **RQ3** i.e. **'How do agents that proactively elicit search criteria (proactive elicitation) and proactively recommend search results that are outside the original scope of user's query (proactive recommendation) vary in terms of their impact on: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?**

6.6.1 Impact of Conversational Strategy on Cognitive Workload

Although we found no significant differences in terms of an agent's impact on participants' cognitive workload (**lack of support for H3.1**), we noticed some differences between the agents. We observe that, overall, ProActive CAs place less strain on a participant's cognitive workload. (cf. Table 6.3, NASA TLX). These results confirm our expectations with regards to the impact of an elicitation strategy on a users' cognitive workload. In terms of individual strategies, Active Recommendation is the worst in terms of cognitive load as compared to ProActive Listing which performs best. It can be argued that an Active Recommendation strategy is the most strenuous agent, as participants are exposed to more information at the beginning of conversation (i.e. they are not able to provide their preferences upfront) and have to employ additional cognitive resources to reason about flight alternatives. In the ProActive Listing condition, however, the algorithm supports the participant by offering refinement suggestions to reduce cognitive load.

6.6.2 Impact of Conversation Strategy on Satisfaction

There were no significant differences in terms of satisfaction (**lack of support for H3.2**). However, overall, ProActive CAs were perceived as the more usable (cf. Table 6.3, SUS). When it comes to satisfaction with the agent's support in exploring the search space, differences are marginal (see Table 6.3, SSES). As demonstrated in Section 6.4.2, Active and No Recommendation agents match ProActive and Recommendation agents across the board, except for the speed of results presentation, where ProActive and Recommendation agents score higher (4.5/5 vs. 4/5). Overall, a Proactive Recommendation Strategy is the only one that scores 5/5 for results presentation speed.

6.6.3 Impact of Conversational Strategy on Performance

Although we did not observe statistically significant differences between the Agents in terms of Performance (**lack of support for H3.3**), we have seen that ProActive and Recommendation agents led to better performance, with more participants selecting Pareto Optimal Flights than when using Active and No Recommendation Agents (see Table 6.4, Pareto Optimal). A similar trend was observed for meeting the time preference specified in the scenarios (see Table 6.4, Meeting Time Preference). ProActive and Recommendation Agents also led to less money being lost and less time being wasted (see Table 6.5). Interestingly, we have seen that while participants who interacted with Recommendation Agents got more Pareto Optimal flights, sometimes following an agent's recommendation led to missing a Pareto flight (see Table 6.6 for details). This could be explained by the fact that participants settled on the first recommendation without exploring alternatives available on different days, effectively missing optimal options.

6.6.4 Impact of Conversational Strategy on Interaction Metrics

Despite no significant differences in the data obtained from Subjective measures, there is a noticeable difference in Objective measures with regards to interaction time and number of conversational turns (see Table 6.8 for details) **(support for H3.4)**. We observed that participants managed to complete their search task significantly faster and in a smaller number of conversational turns when using Proactive Agents than other agent types.

6.6.5 Perceptions of Conversational Strategies

Participants were generally in favour of being exposed to the broader scope of flights - preference for Active Elicitation (e.g. P3 and P4); however, the opinion was split, and other participants were more in favor of being provided their results quicker (e.g. P9 and P18). Opinions were also split in terms of recommendations. While some people found them useful, they caused others uncertainty (see Table 6.9 for details). In terms of challenges encountered while using the agents (summarised in Section 6.5.2.2), lack of visual feedback proved to be the main concern as participants were not ready to make their decisions based only on what they heard as they would like to have had the options written down for comparison.

6.6.6 Reflections on Findings

Based on the experience of observed conversations, we would like to offer several reflections on the potential reasons for our experimental findings.

Firstly, we noticed that the conversational strategy of the CAs impacted on participants' trust and affected how they interacted with the agents. For instance, P3, P9 and P18 expressed their suspicion about some CAs because they did not elicit their preferred search criteria at

the beginning of the interaction (please see comments in Section 6.5.2.2 for more details). This 'suspicion' could have been aroused by certain CAs violating participants' expectations by failing to fulfil their 'functional role' [29] i.e. not offering the adequate level of search support that was required to facilitate completion of the task. We believe that any future interactive evaluation experiments should incorporate trust-related metrics in order to provide an additional context to interpret the findings. Recent research in the field of human-agent collaboration shows that a behavioural measure of trust can be used to predict task outcome [71]. Secondly, the fact that all CAs 'sounded very natural and human-like' (see P10 comments in Section 6.5.2.3) could have made participants more concerned about politeness norms and made them regard the agent as a social actor with emotions [107] rather than merely a device to complete the task. P10 remarked that they made their flight selection quickly, because they did not want the agent to wait for too long. Possibly, concerns regarding the CA's patience could have lowered the number of flight options that participants were willing to explore during the interaction. We admit that using a more robotic voice (e.g. an outdated HMM speech synthesiser [151] characterised by unnatural prosody and pitch) would have eliminated such politeness concerns, however it could have been somewhat 'artificial', as currently naturalness of the state-of the art synthetic speech makes it almost indistinguishable from a human voice [88]. Finally, as in Study 2, the lack of visual feedback was challenging for participants (please see comments in Section 2.4 and Section 6.5.2.2 for more details). We observed that some participants explored only one out of three available travel days thoroughly and then, due to the high cognitive workload imposed by the exploratory nature of the task, settled on the first option that they considered acceptable. While participants' feedback and behaviour puts the suitability of voice-only flight search into question in its current form, there are some possible adjustments that could make the task more accessible. For instance, as suggested by P21 (see comments on page 119), the task could be split into two stages where first at (the voice only, pre-selection stage) a participant explores the search space in order to shortlist potential flight options which are then displayed on the PC/mobile screen for comparison (the selection stage). This would reduce cognitive workload and allow participants to make more informed decisions.

6.7 Conclusions

We conducted a user WOZ study to examine the impact of an agent's proactive search support and provide an answer to our **RQ3: 'How do agents that proactively elicit search criteria (proactive elicitation) and proactively recommend search results that are outside the original scope of user's query (proactive recommendation) vary in terms of their impact on: (a) Cognitive load, (b) Satisfaction with the Agent, (c) Task performance, and (d) Interaction time?'** The results indicate that participants' subjective perceptions of different CAs do not differ much in terms of perceived workload, satisfaction

and performance. However, proactive agents were shown to allow participants to complete the allocated tasks faster and in fewer conversational turns. Our findings indicate that proactive agent support could lead to tasks being executed faster, while preserving performance and a satisfactory overview of the search space.

6.8 Chapter Summary

This chapter investigated the impact of an agent's proactive involvement on user search experience. While we have not seen significant improvements in subjective measures or participants' performance, proactive elicitation was found to help participants complete search scenarios significantly faster and with significantly fewer conversational turns. In the next chapter we will revisit all studies presented in Chapters 4-6 to compare the results based on the differences in experimental setups and differences in the design of simulated search tasks.

Chapter 7

Comparison of Three Wizard of Oz Studies

The current chapter provides a comparison of three Wizard of Oz studies (presented in Chapter 4 (Study 1), Chapter 5 (Study 2) and Chapter 6 (Study 3) that contribute to this PhD. The aim of the comparison is to contextualise the results obtained in each of the studies and reflect on changes in experimental design that were implemented in order to: (1) address the increased complexity of variables under investigation (changes from Study 1 to Study 2) and adapt the study to online delivery (change from Study 2 to Study 3 - dictated by the COVID-19 pandemic). We first discuss the difference in experimental setups between all of the studies (Section 7.1) and compare the ways in which results were presented by the intermediary to the seeker. Next, in Section 7.2, we present the differences in search task design used in the studies. The impact of using different experimental setups and different search task designs is then presented by comparing the results for objective and subjective measures that were applied to evaluate user search experience across the studies (Section 7.3). Then we provide reflections on our results, focusing on: (1) experimental setup (Section 7.4) and (2) search task design (Section 7.5). In Section 7.6 we discuss the contributions of the three WOZ studies and concede their limitations. Finally, in Section 7.7 we draw conclusions derived from the comparison of the studies. This chapter is partly based on our research paper 'Interactive Evaluation of Conversational Agents: Reflections on the Impact of Search Task Design' [53]. [2]

[2]The research paper focuses on a comparison between Study 1 and Study 2. The current chapter provides additional insights by also comparing Study 2 and Study 3.

7.1 Experimental Setup Comparison

Figure 7.1 illustrates differences in experimental setups across the three studies.

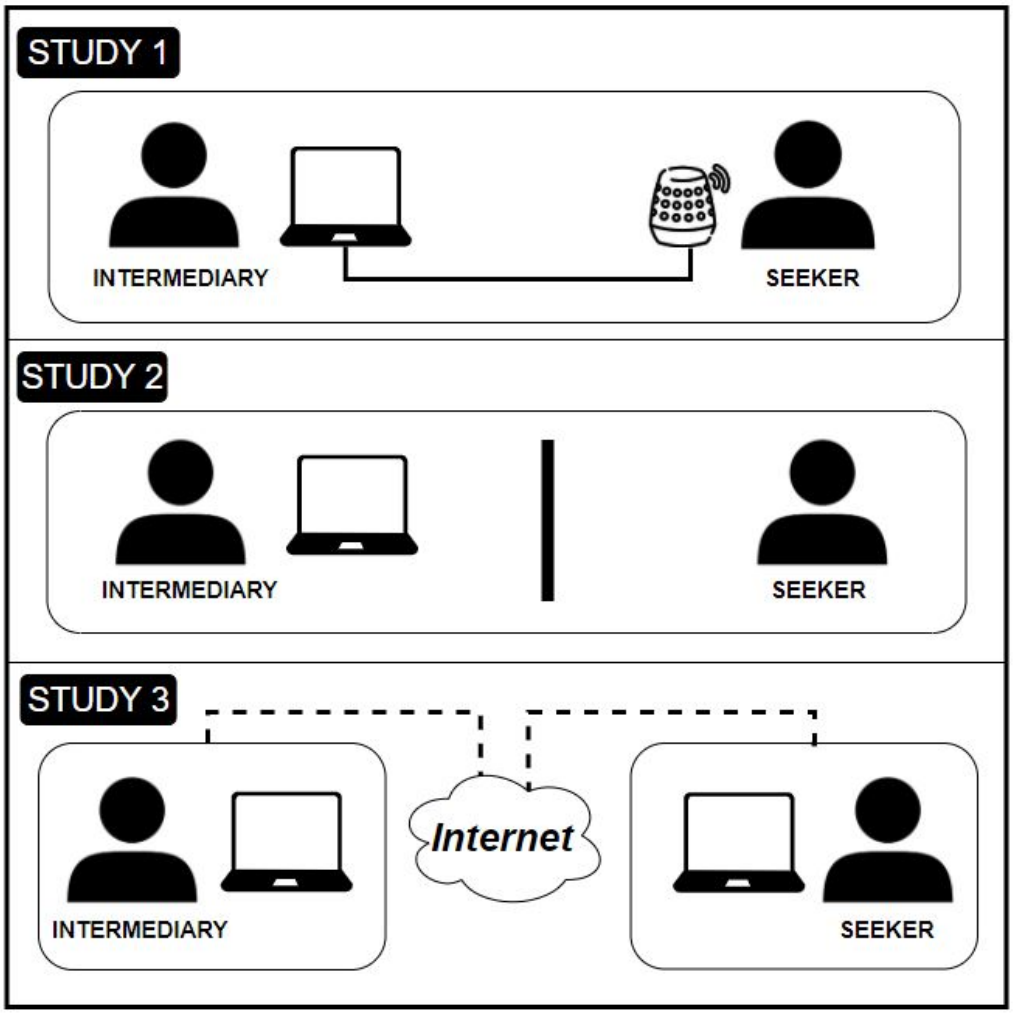


Figure 7.1: Comparison of different setups used for Studies 1-3. The roles of conversational partners are fulfilled by an intermediary (wizard) and a seeker (participant).

In Study 1 (described in Chapter 4), which focused on the impact of conversational agent memory, the search results were provided to participants via synthetic speech that was played though a smart speaker. The wizard was using a console that contained a list of prerecorded prompts and had a live speech synthesis tool to synthesise speech ad-hoc. In Study 2 (described in Chapter 5), which explored the impact of eliciting and presenting information to the seeker, due to the higher number of utterances required for synthesis, in order to ensure consistency and timely feedback, we decided to provide the search results back in a human voice. The wizard recruited in Study 2 was separate to the experimenter. The final study (described in Chapter 6), which explored the proactivity and recommendation features of a conversational agent, was an online adaptation of Study 2 that was conducted via Zoom. In Study 3 results were also provided via human voice, however, with the main experimenter acting in the capacity

of the wizard. The decision not to recruit another wizard was taken due to time constrains, as there was not sufficient available time for training. Instead, experimental instructions were provided in a synthetic voice to distinguish between the experimenter and the wizard[1]. Moving the final experiment online was necessitated by the outbreak of the COVID-19 pandemic which prohibited in-lab, face-to-face studies. Please see the document in Appendix D.2 for guidance on conducting user studies during the COVID-19 pandemic.

7.2 Comparison of Search Tasks

The current section provides a comparison of simulated search tasks between Study 1 and Studies 2 and 3. In Study 1, the goal of participants was to find the cheapest flight with a specific departure time. An example search scenario used in Study 1 is presented below.

> You are planning to visit your friend who lives in **Bristol**. You will be flying from Glasgow Airport. Your total budget is **100 pounds**. You can travel either on the **11th or 12th of November**.
> **Indicative Request:** You want to find the cheapest possible deal but your flight needs to leave on, or before 11am.
> **Note:** Please wait for the agent to finish before you start to speak.

In Studies 2 and 3, participants were instructed to explore the available options to find the shortest and cheapest flight *that met a specific arrival preference* (e.g. 'You want to reach your destination around 1pm to avoid traffic.'). Contrary to Study 1, we did not provide a strict budget but encouraged participants to explore the trade-off between flight cost and travel time. In order to motivate participants to explore the search space by using a background story (e.g. 'You are a student who is attending a conference in Stockholm, try to save money from your travel fund while making sure that you reach your accommodation on time'). We also highlighted the implications of not meeting the provided search criteria (e.g. having to pay for late check-in). An example search scenario, used in Studies 2 and 3 is presented below.

> You will be attending a student conference in **Stockholm**. You will be travelling there on either **Monday the 5th, Tuesday the 6th, or Wednesday the 7th of November**. Your university advised you that you will be allocated money from your conference fund that you will use to fund other events till the end of your academic course. To be able to attend more events in the future, you want to save money while not spending too long getting there. The student dorms where you will be staying charge extra for late check-in, so you will be aiming to arrive at around 7pm to be able to check in to your accommodation on time.
> **Indicative Request:** Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
> **Note:** Please wait for the agent to finish before you start to speak.

All search scenarios are provided in the Appendix for reference (see E.1 for Study 1, E.2 for Study 2 and E.3 for Study 3)

[1]Participants' instructions for Study 3 can be accessed via the following link: https://strathcloud.sharefile.eu/d-sdcbd7c8ba5243a8a (last accessed on the 1st November 2020).

7.3 Impact on User Search Experience

This section provides a comparison of user search experience scores obtained from both subjective and objective measures (the detailed discussion of all the metrics is provided in Section 3.4) registered across studies 1-3. It should be noted that we refer to the VSS system as a passive agent and the CSA system as an active agent. VSS is considered passive because, unlike CSA and other active agents featured in Studies 2 and 3, it does not offer any support with the search process. The decision to modify the names of the systems from Study 1 was taken to facilitate a comparison between all agent types and make the nomenclature uniform across the studies. The comparison is presented chronologically by focusing on pairs of studies (i.e. Study 1 and Study 2, and Study 2 and Study 3). A comparison of Studies 1 and 2 provides insights on how the impact of search task design differed between the studies, while comparison of studies 2 and 3 focuses on the mode of delivery (i.e. in-lab for the former and online for the latter).

7.3.1 Comparison of Impact on Cognitive Load

7.3.1.1 Study 1 vs. Study 2

Figure 7.2 presents a comparison of NASA TLX scores for Study 1 and Study 2. We can see that as task complexity increases the difference in perceived cognitive workload decreases. While scores for the Passive agents remain similar between the studies - 29.5 Med (18 IQR) in Study 1 and 33 Med (21 IQR) in Study 2, there is more variance between the Active Agents – 14 Med (12.5 IQR) in Study 1 and 28.35 Med (20 IQR) in Study 2. While in Study 1 there was a statistically significant difference between Active and Passive agents ($p < .001$), in Study 2 we did not observe any statistically significant difference between the Active and Passive agents ($p = 0.517$).

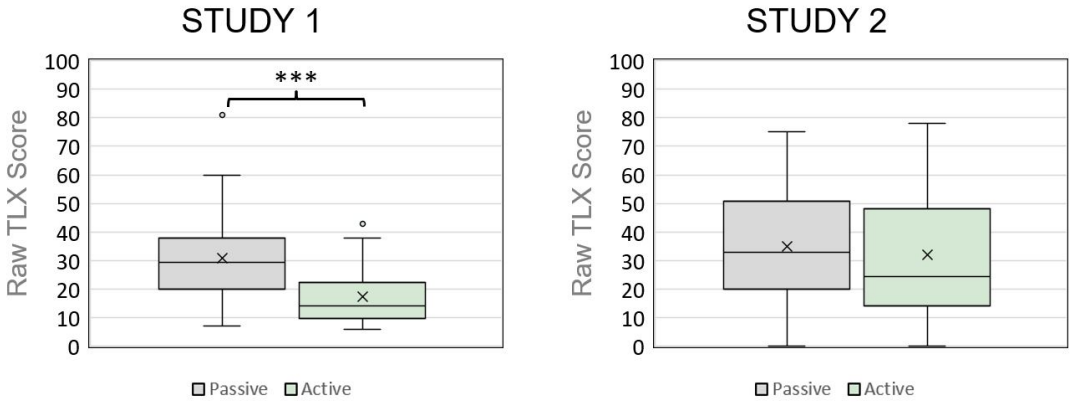


Figure 7.2: Comparison of Raw NASA TLX scores between Study 1 and Study 2.

7.3.1.2 Study 2 vs. Study 3

Figure 7.3 presents a comparison of NASA TLX scores for Study 2 and Study 3. We can observe that the scores for Active Agents are very similar – 28 Med (IQR = 20.27) in Study 2 and 29 Med (IQR = 27) in Study 3. In Study 3, we can observe that ProActive agents achieved the lowest NASA TLX scores - 25 Med (IQR = 9.5). We can also see in the bottom part of the figure that Recommending agents led to a higher cognitive workload, which is understandable as participants needed to consider alternative options while interacting with the agents. It should be noted that despite the difference in the mode of deployment (lab-based study vs. online experiment) the ranges of scores for the agents are similar.

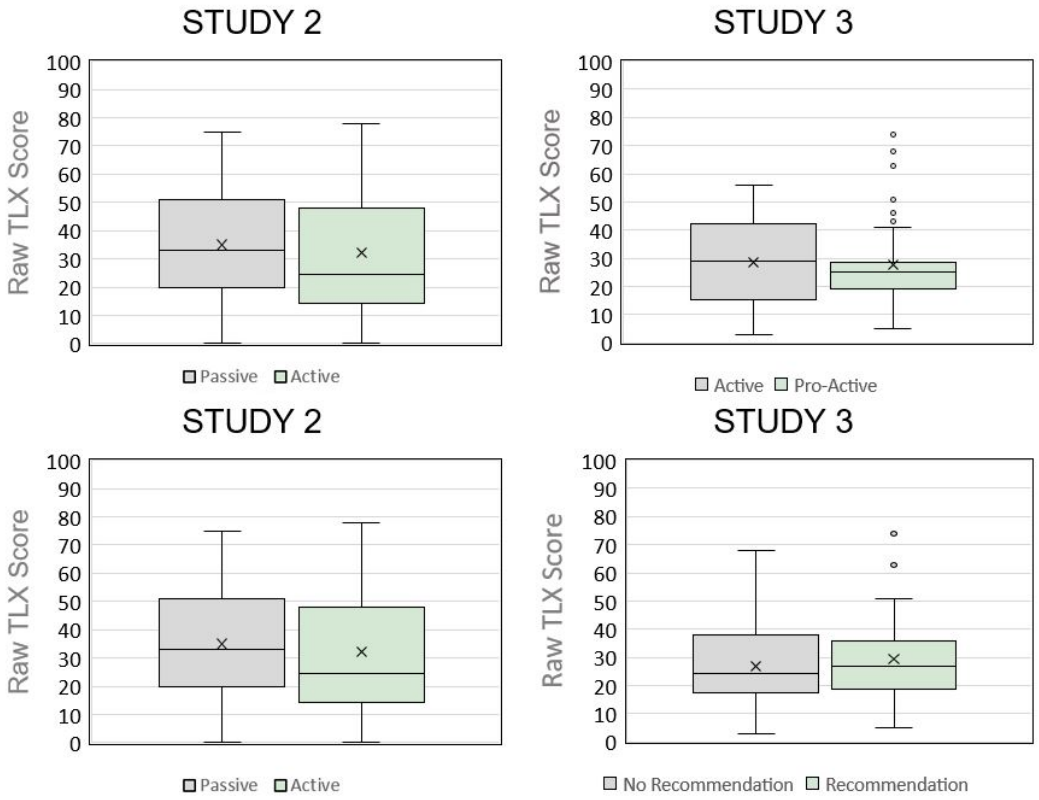


Figure 7.3: Comparison of Raw NASA TLX scores between Study 2 and Study 3.

7.3.2 Comparison of Impact on Satisfaction

7.3.2.1 Study 1 vs. Study 2

Figure 7.4 presents a comparison of SUS scores for Study 1 and Study 2. We can observe that the level of search task complexity translated to different satisfaction scores. While in Study 1 there was a significant difference between Active and Passive agents ($p = 0.003$), no such difference was observed in Study 2 ($p = 0.590$). It is important to note that in Study 1 Passive Agents did not have conversational memory which led to more frustration and, in turn, lower

SUS scores. In Study 2, where both agents were stateful, participants did not have to repeat their queries and thus the benefits obtained from the Active support became less explicit, a fact reflected in the SUS scores, which differ less between each agent. SUS scores for both Passive and Active Agents are higher in Study 1 (81.25 and 92.5 respectively) than in Study 2 (76 and 80 respectively) which can be attributed to an increased level of difficulty resulting from the trade-off imposed by search scenarios in Study 2.

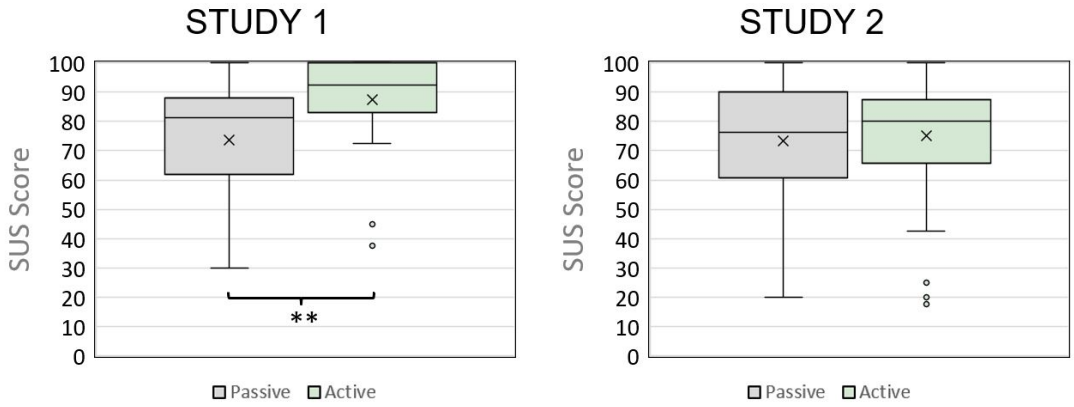


Figure 7.4: Comparison of SUS scores between Study 1 and Study 2.

7.3.2.2 Study 2 vs. Study 3

Figure 7.5 presents a comparison of SUS scores for Study 2 and Study 3. Overall, the results are very similar for all agent families i.e. Passive (81.25), Active (80 - both Study 2 & 3) Proactive (82.5) and Recommendation (82.5). No statistically significant differences were observed within the studies.

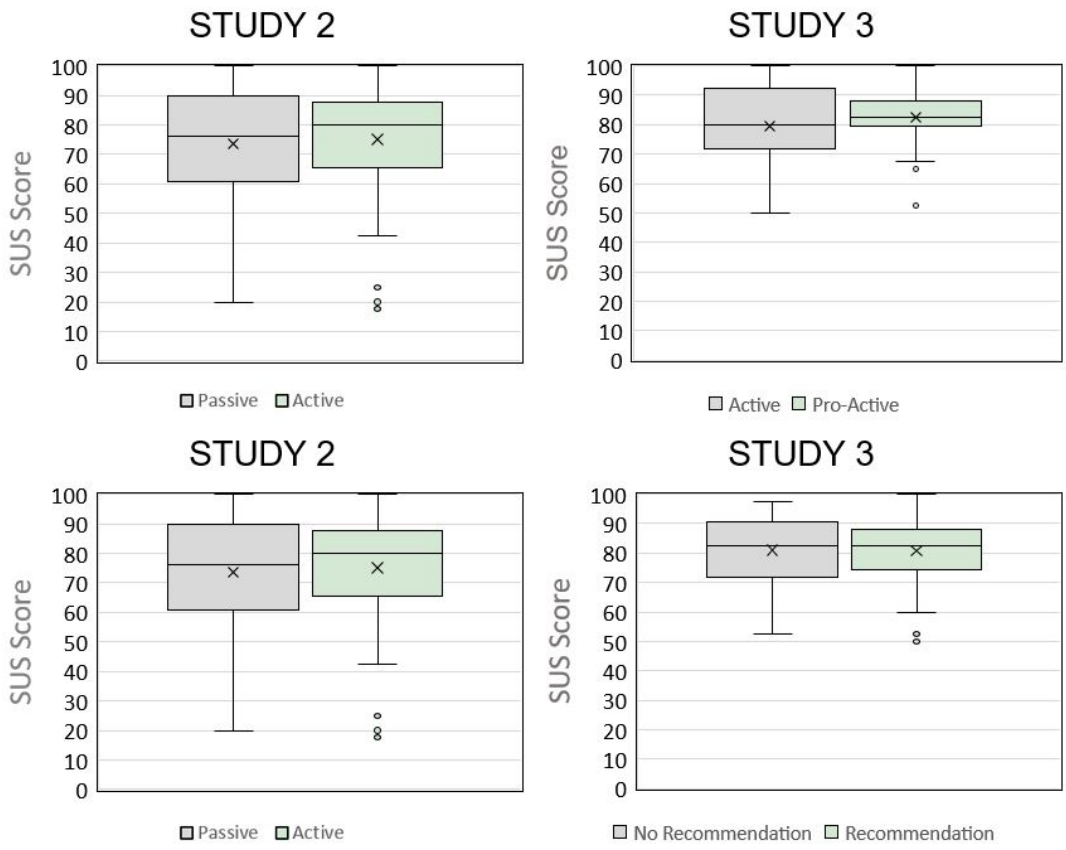


Figure 7.5: Comparison of SUS scores between Study 2 and Study 3.

7.3.3 Comparison of Impact on Performance

7.3.3.1 Study 1 vs. Study 2

Figure 7.6 illustrates the difference in performance (number of Pareto Optimal flights booked) between Studies 1 and 2 when using different types of CAs. In Study 1 we observe a statistically significant difference between Passive and Active Agents ($p < 0.001$). In Study 2 there was a big drop in the number of participants who managed to select a Pareto Optimal flights; although there is a noticeable difference between Passive and Active agents, it does not amount to statistical significance ($p = 0.077$).

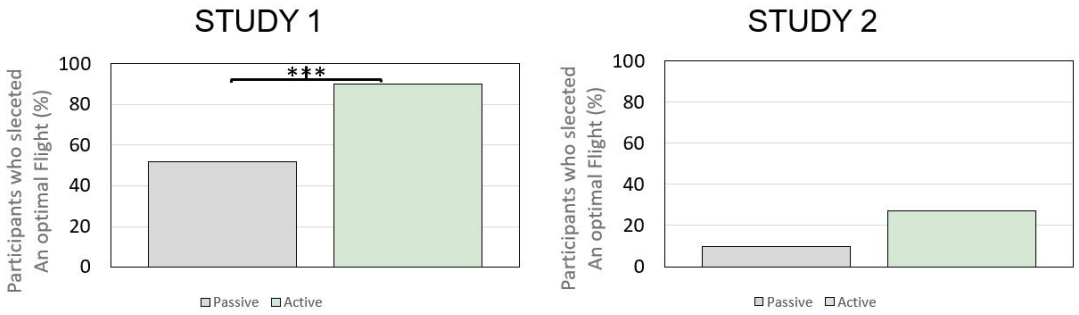


Figure 7.6: Comparison of agent performance for Study 1 and Study 2. The performance is considered in terms of percentage of Pareto Optimal flights selected by participants.

7.3.3.2 Study 2 vs. Study 3

Figure 7.7 presents the difference in participants' performance between Studies 2 and 3. We did not observe a statistically significant difference between any agent pairs in Study 2 and Study 3. In Study 3 there was less of a difference between Active and ProActive agents ($p = 0.18$), and between No Recommendation and Recommendation Agents ($p = 0.115$). Interestingly, more participants managed to get a Pareto Optimal flight in the in Study 3, using an Active agent (28/46) than in Study 2 (13/48). This could have been caused by the fact that participants in Study 3, which was run though Prolific, were more motivated as their payment was subject to performance (the Prolific reward system is further described in Section 7.4).

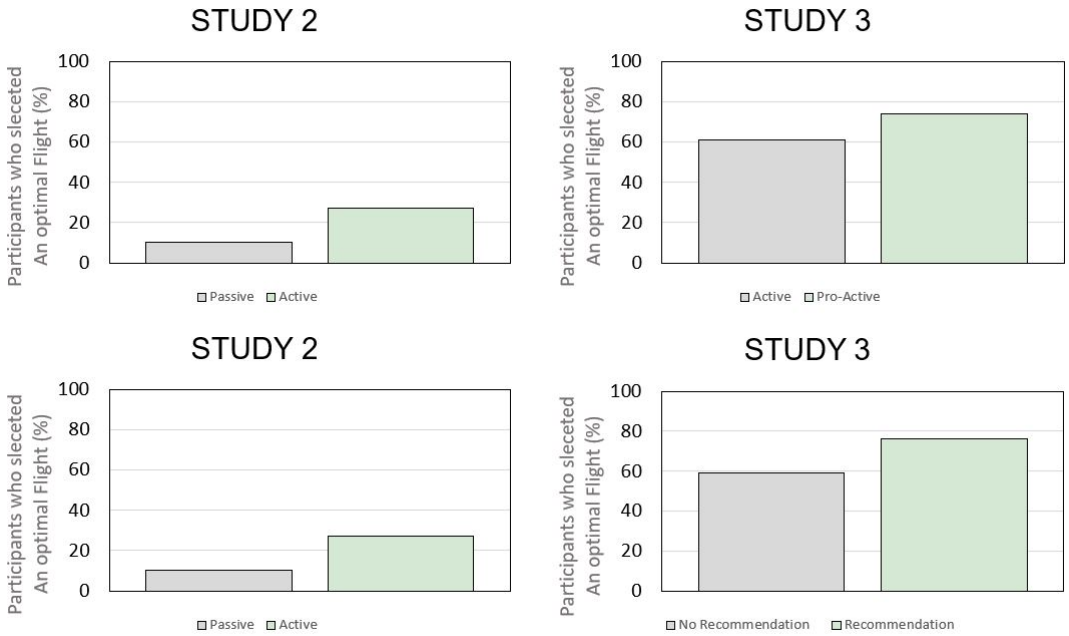


Figure 7.7: Comparison of agent performance for Study 2 and Study 3. Performance is considered in terms of percentage of Pareto Optimal flights selected by participants.

7.3.4 Comparison of Impact on Interaction Times

7.3.4.1 Study 1 vs. Study 2

Figure 7.8 presents a comparison of interaction times for Study 1 and Study 2. It can be seen that interaction times are much shorter in Study 1 compared to Study 2. In Study 1 there is a statistically significant difference between Passive and Active agents ($p < 0.001$). In Study 2, no statistically significant difference between Passive and Active agents was observed ($p = 0.341$). Interestingly, in Study 2, participants spent more time interacting with Active agents than with Passive ones. This could indicate greater involvement in the study as participants wanted to find flight options that offered a 'good balance' between money and travel time.

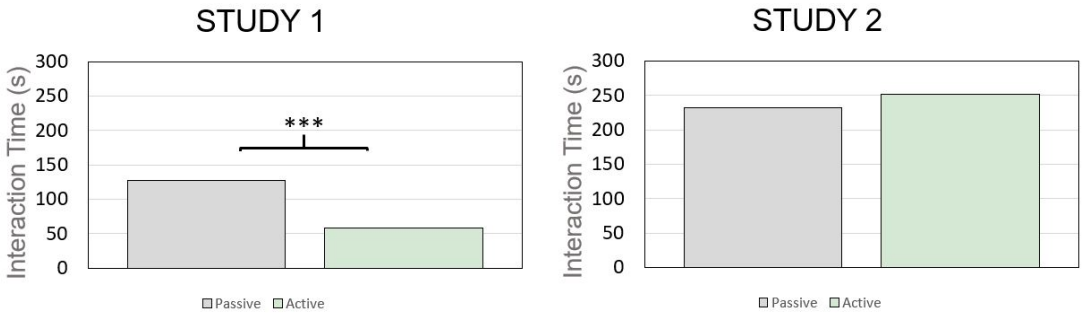


Figure 7.8: Comparison of interaction times for Study 1 and Study 2.

7.3.4.2 Study 2 vs. Study 3

Figure 7.9 presents a comparison in interaction times between Study 2 and Study 3 for different types of agents. While in Study 2, we did not observe a statistically significant difference in interaction times between Passive and Active agents ($p = 0.341$) nor for Summarising and Listing Agents ($p = 0.555$) in Study 3 there was a statistically significant difference between Active Agents and Proactive Agents ($p = 0.0011$) and between No Recommendation and Recommendation Agents ($p = 0.008$). It can be seen that the median interaction time for the Active agents went up from 252 seconds in Study 2 to 273 seconds in Study 3.

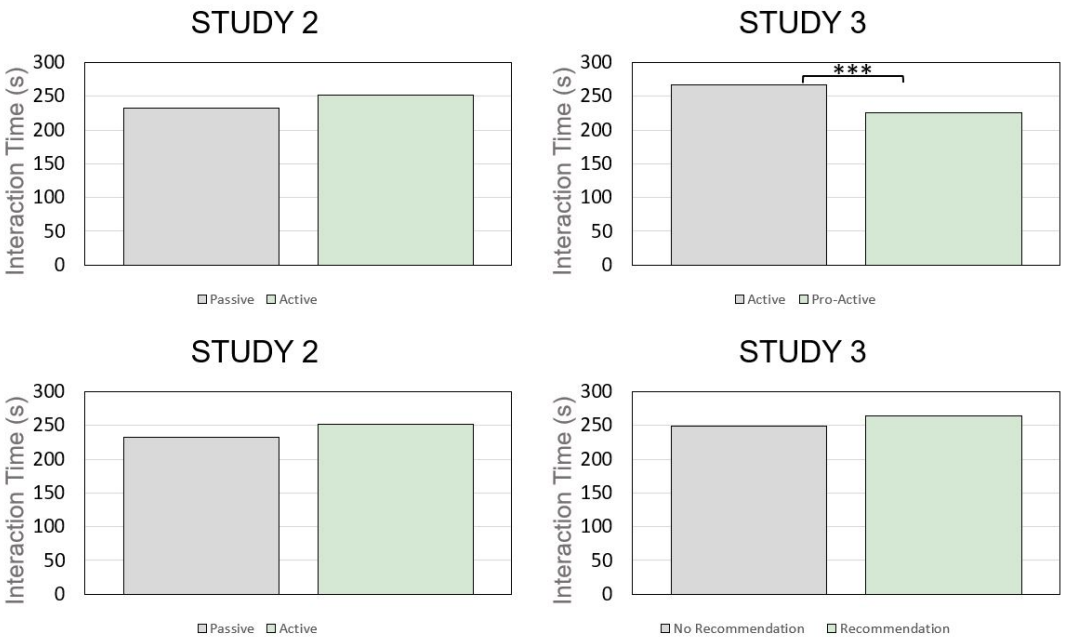


Figure 7.9: Comparison of interaction times for Study 2 and Study 3.

7.3.5 Global Comparison across Studies 1-3

The relationship between participants' performance and other metrics of user search experience is presented in Figure 7.10 (cognitive workload), Figure 7.11 (satisfaction) and Figure 7.12 (interaction time). For cognitive workload, it can be seen that the relationship between NASA TLX scores and performance is negative - in general the lower cognitive workload scores coincide with better performance (the higher percentage of selected Pareto Optimal flights). In terms of satisfaction, the reverse trend can be observed - the higher SUS scores tend to coincide with better performance. As for interaction time, with the exception of Study 2, shorter interactions are indicative of better performance. In general, these three trends are indicative of the beneficial impact that CAs with increased conversational support can have on user search experience in a voice-only goal-oriented task.

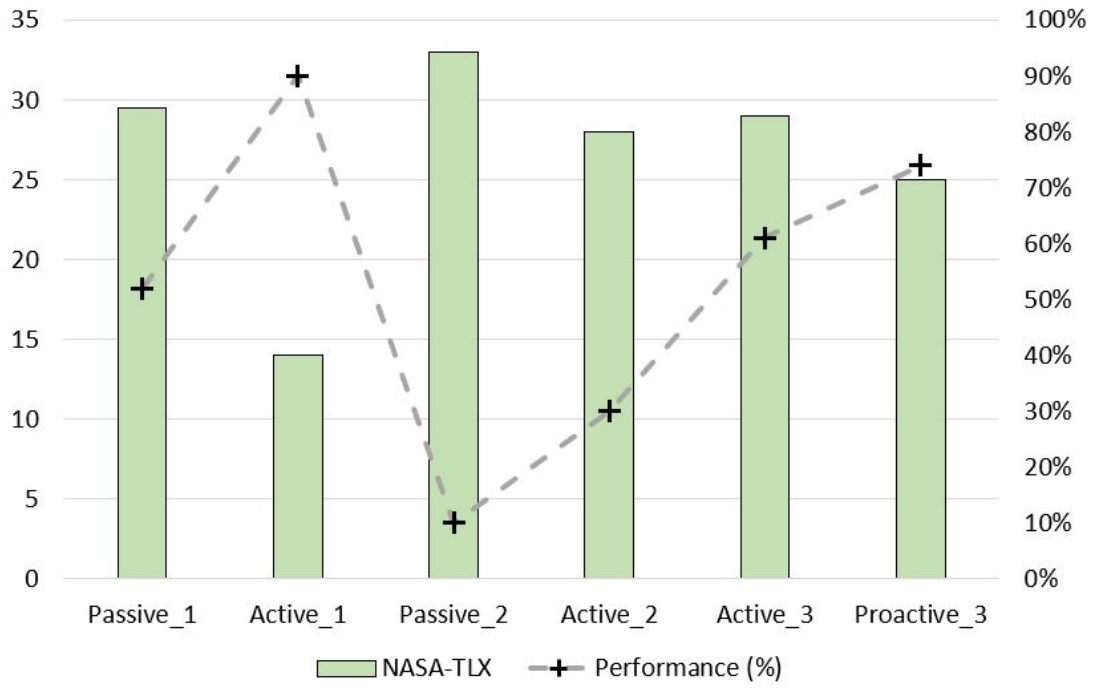


Figure 7.10: Comparison of Nasa TLX scores and participants' performance across Studies 1-3.

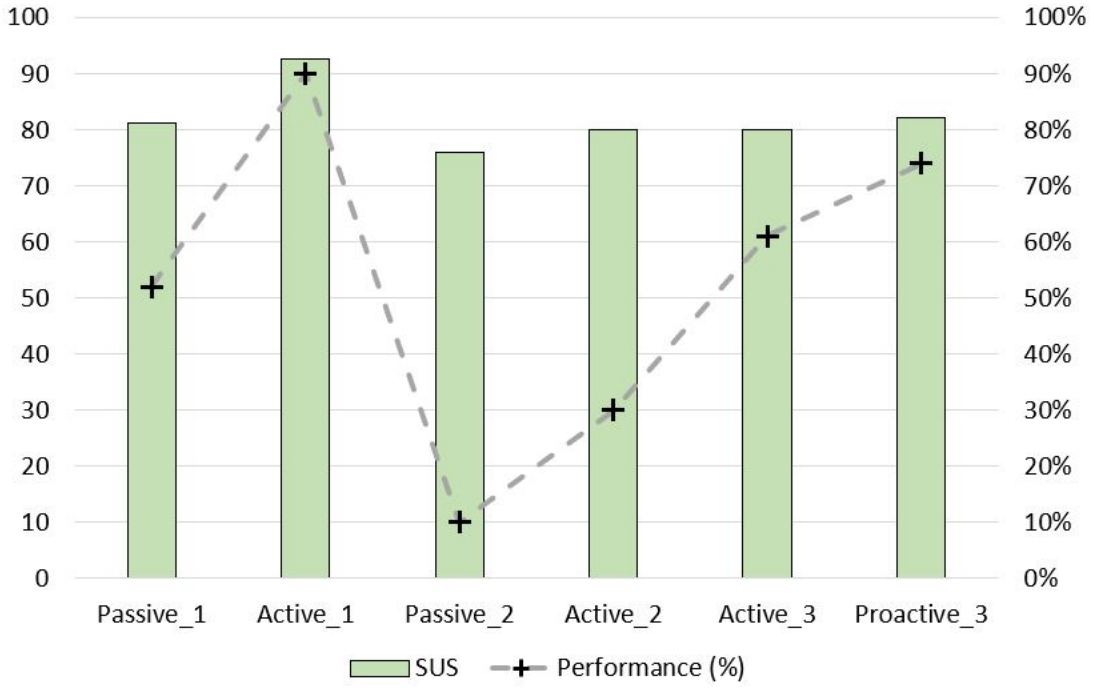


Figure 7.11: Comparison of SUS scores and participants' performance across Studies 1-3.

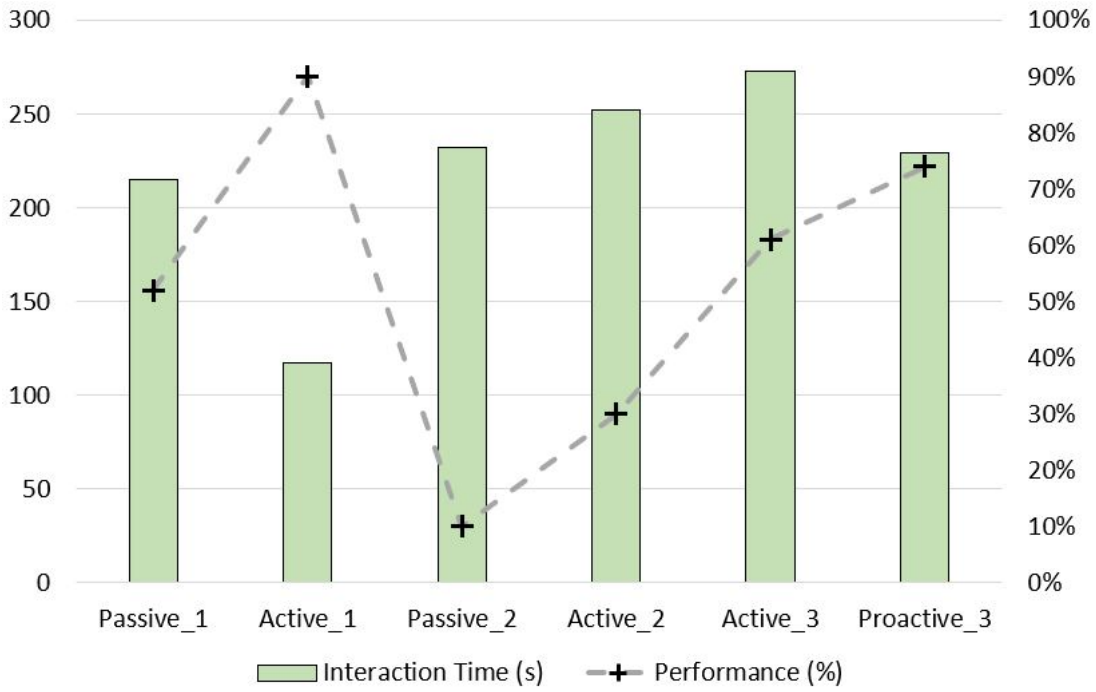


Figure 7.12: Comparison of integration time and participants' performance across Studies 1-3.

7.4 Reflections on Mode of Experiment Delivery

A comparison of Study 2 and Study 3 indicated little variance in user search experience scores for all metrics except participants' performance. As we can see in Figure 7.7 the proportion of Pareto Optimal flights selected by participants is substantially larger in Study 3 than in Study 2 for all types of compared agents. Potentially, this could be attributed to the fact that participants who were recruited via Prolific (Study 3) were used to the system where payment is based on task-completion time (see Figure D.3) and participant were offered additional incentives if they perform well (see Figure D.4). To migrate this bias, we specified that each participant would be rewarded with £10 upon successful completion of the study. Overall, it is reassuring that despite the change in the study's mode of delivery (i.e. in-lab vs. online) and in the person acting as wizard (external person vs. experimenter), we did not observe substantial differences in user search experience scores.

7.5 Reflections on Task Design

Based on the results of our comparative analysis of Studies 1 and 2, we can make the following observation regarding the impact of task-complexity on the evaluation of voice-only, goal oriented CAs. **(1) Tension between search criteria:** When the wording of the task highlighted the trade-off between the search metrics (e.g. flight cost vs. flight duration), participants were more involved in the task and explored the search space more thoroughly. Consequently, participants were less likely to satisfice (settle for a minimally acceptable option) and spent more

time interacting with the CA. The set-up featured in Study 2 which was more restrictive and included a larger number of search criteria than Study 1, provided us with more conversational material for analysis and, in turn, richer insights into participants' behaviour. **(2) Realistic Constraints:** When the search task provided participants with an additional constraint (e.g. preferred time of arrival), motivated by the background story (i.e. reasons for meeting arrival time), we observed longer interaction times than when there was no constraint. It may indicate that as participants tried to mitigate the negative consequences of missing the recommended arrival time, (e.g. getting stuck in the traffic, paying extra for check-in, etc.) their involvement in the search task increased.

7.6 Contributions and Limitations

Through three WOZ studies, the current thesis sought to explore the impact of four variables (i.e. cognitive workload, satisfaction, performance and interaction time) that relate to the style of a CA interaction and the level of its support on user search experience. The ultimate goal of this exploration was to learn how users can be best supported when using a voice-only agent for a goal-oriented task to achieve optimal results. In this section, we will highlight the contributions of the current work set against previous research and concede its limitations.

7.6.1 Contributions

This thesis offers several empirical insights that can guide the design of CA evaluation tasks to better understand the impact of different conversational interaction strategies on user search experience. Specific contributions of the current work to the research focused on conversational search are listed below.

Firstly, the thesis provides an empirical bridge to the theory of conversational search by exploring the impact of the two core elements of a truly conversational agent proposed by Radlinski and Craswell, namely (1) memory and (2) conversational initiative [126]. More specifically, we explore how these two elements can help to minimise the cost of interaction (cognitive workload) and maximise success in a voice-only search task (selecting optimal flight option). In previous works, the criteria of cost and task success were applied as fundamental metrics determining overall user satisfaction (cf. [167]); in the current thesis, they are considered with a focus on the trade-off between providing enough information and not overwhelming the user (a challenge of conversational search highlighted by Trippas et al. [158]).

Secondly, the thesis proposes a spectrum of conversational strategies (discussed in Section 3.3.1) that defines a CA in terms of its conversational involvement and style of presenting information. Using the spectrum of conversational strategies provides a basis to systematically vary the conversational behaviour of CAs to make their evaluation consistent and comparable across different users and scenarios. While previous work on the evaluation of goal-oriented dialogue

systems ensured consistency by using fully automated algorithms (e.g. [117,119,141,166]), it was limited by the performance of individual modules of the system (especially automatic speech recognition) that impacted on task completion. On the other hand, human-human (H-H), voice-only interactive information seeking studies, where pairs of participants acted as intermediaries and seekers (e.g. [150,157]), were not subject to such system performance limitations. Nonetheless, due to their nature, H-H studies provide little control over the consistency of information presentation strategies due to individual differences between each pair of participants. Our approach of using WOZ to emulate conversational strategies allowed us to combine the merits of both approaches, i.e. (1) ensuring consistency by providing the wizard with interaction algorithms, while (2) mitigating problems related to the individual performance of system modules. Consistency was achieved by defining interaction strategies for each agent to determine its conversational behaviour (discussed in Section 4.3.6, Section 5.3.6 and Section 6.3.6) and testing them though pilot studies before deployment.

Thirdly, this thesis introduces a novel CA evaluation metric that is based on the concept of Pareto Optimality (discussed in Section 3.4.2.1). We drew inspiration from Borlund's simulated work tasks [22] to create a series of travel scenarios that introduced a set of search criteria that determined which options were considered as optimal. The search space for each task was balanced to ensure that the trade-offs between different options are comparable across search scenarios and that participants need to engage in exploration in order to find the optimal solution (see Figure 3.6 for details). Participants' performance was measured by evaluating how their selected option compared with other available alternatives within the search space. Our proposed metric can be applied in other product or service seeking scenarios that require a seeker to compromise between different attributes. The required search criteria (illustrated in Figure 7.13) can be defined to determine what makes the desired options 'optimal.'

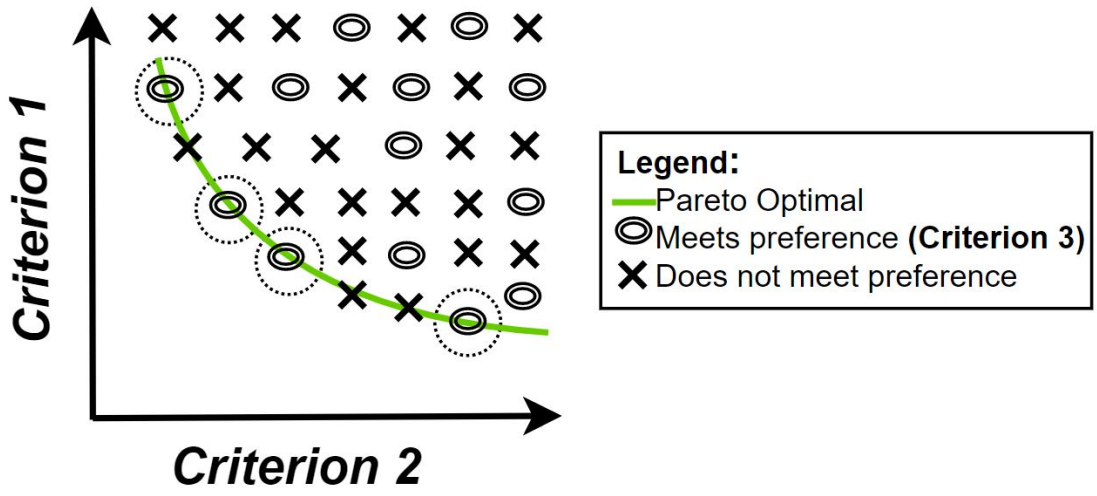


Figure 7.13: Search space with Pareto Optimal options marked with dotted circles.

To summarise, the key contribution of the current thesis is that, by combining three WOZ studies, it empirically validates the Theoretical Framework of Conversational Search [126] and verifies the postulated benefits of active and pro-active conversational support [156]. The theory of conversational search is systematically tested by using scenarios with a balanced search space that ensures comparability and an even spread of search options. We combine metrics from the fields of human-computer interaction and information retrial, previously used in the evaluation of dialogue system and conversational agents [85, 87, 96, 98, 103, 119, 167, 176], and contextualise them by incorporating participants' feedback obtained through semi-structured interviews. Overall, our quantitative and qualitative findings reveal the beneficial impact of an agent's statefulness (conversational memory) across all aspects of user search experience, and indicate that as a CA's initiative grows, participants are able to achieve better performance and faster task completion.

Despite the contributions of the current research, however, it also has several limitations, which are outlined in the next sub-section.

7.6.2 Limitations

While the three WOZ studies presented in the current thesis allowed us to consistently and systematically test different conversational strategies, our approach is subject to some limitations. Firstly, in Study 2 and Study 3 (Chapter 5 and Chapter 6), no statistically significant differences were found regarding the impact of CAs on user cognitive workload and satisfaction. The lack of significance could be linked to participants' perceptions of the agents; in Study 3, 11 out of 23 participants reported that they could not see any difference between the CAs (see Section 6.5.2 for more details). This could be due to the within-groups experimental design used in the current research, where each participant interacts with multiple conversational agents. As indicated by the semi-structured interviews, having completed the initial tasks, several participants developed an interaction strategy that they applied to other CAs to improve performance (see P3, P5 and P23 comments in Section 5.5.3.1). While using a between-groups design, where each participant interacts with only one type of agent, could have addressed this problem and prevented the 'carry over effect', it could have also caused comparability issues due to the different cognitive abilities of participants.

Secondly, participants' feedback regarding the challenges of conducting a voice-only search requiring a comparison of various options (see Section 5.5.3.2), poses questions about the suitability of the flight-booking task (especially when it requires exploration of an extensive search space). As indicated by P21 in Section 6.5.2.3, displaying results on the screen could have allowed participants to make more informed choices. While this point is valid, and recent research has explored how proactive agents can support users by providing results via different devices (i.e. combining voice and graphical feedback through smart speakers and computer

screen) [110], the current thesis foucused exclusively on voice-only interaction and exploration of its challenges.

Thirdly, while many participants appreciated the increased involvement of conversational agents in the search process (as indicated by the comments in Section 5.5.3.1 and Section 6.5.2.1), some found it unpredictable (see P1, page 62), intrusive (see P21 and P24, page 90), or suspicious (P9, page 117). Differences in participants' expectations regarding the required level of CA support highlight the need for CA personalisation both in terms of elicitation and revealment of information. In the current thesis, for the sake of consistency, the explored conversational strategies were rigid and did not adapt to the different interaction styles preferred by participants. In order to overcome this limitation, future research could elicit a participant's preferences regarding the CA's behaviour at the beginning of the experiment and then adjust the interaction style accordingly. An example of such personalisation is the 'User Model Summarise and Refine' approach [40] that tailors the information presentation method to user requirements.

Finally, we concede that while a series of subjective and objective metrics have been combined to provide a multi-dimensional overview of user search experience (see Section 3.4), there are other potentially informative variables that were not considered. Examples of such variables are trust and user search agency (i.e. the degree of control over the search process). Recent research by Andolina et al. [4] revealed that while participants appreciated the proactive support of a CA, they felt less in control of the search process. Measuring trust and perceived search agency could provide further meaningful insights to better understand participants' motivations for making specific choices during interaction (e.g. following or rejecting CA's recommendations). This could be achieved through the combination of direct measures such as single-item questionnaires and behavioural analysis [6]. It should be noted, however, that building trust between the CA and the user is a gradual process which requires time (cf. Radlinski and Craswell [126]). Therefore, to ensure higher reliability, it would be best to test it through longitudinal studies.

7.7 Conclusions

The cross-comparison of the three WoZ studies that contributed to this PhD brings us to two main conclusions. Firstly, a comparison of Studies 1 and 2 indicates that search tasks, that add tension between different search criteria encourage a more thorough exploration of the search space and higher participant involvement. Interestingly, we observed that more constrained tasks led to more exploration which is evidenced by longer interaction times. However, the higher participant involvement came at a cost of increased cognitive workload and lower satisfaction with the agent. Secondly, a comparison of Studies 2 and 3 revealed that the mode of experimental delivery (i.e. in-lab vs. online) did not seem to impact on user search experience scores. This finding is reassuring as it indicates that the proposed metrics are robust and stable regardless of the method of their delivery.

Part II of the thesis presented 3 Wizard of Oz studies that explored the impact of different types of conversational agents on user search experience. Chapter 4 explored the role of memory, while Chapters 5 and 6 focused on the role of agent's mixed-initiative - exploring active (Chapter 5) and proactive (Chapter 6) conversational involvement. The last chapter of the Part II (Chapter 7) provided a reflection on the impact of different experimental setups on obtained user search experience scores. In Part III, in the final chapter of this thesis (Chapter 8), we will summarise conclusions derived from 3 Wizard of Studies and discuss possible directions for future work.

PART III:

Conclusions

Chapter 8

Conclusions

Voice interfaces are becoming increasingly prevalent. Commercial systems are currently frequently used for factoid queries and starting to provide support for simple goal oriented tasks such as booking services or product search. Given the increased usage of voice interfaces for information retrieval tasks, it is timely and relevant to investigate how such interfaces impact on user search experience in terms of: cognitive workload, satisfaction, performance, and time required to carry out the task. The current PhD thesis featured three Wizard of Oz studies whose goal was to examine how different types of conversational agents impact on user search experience in goal oriented scenarios. The research presented in this thesis provides insights on cognitive-, satisfaction-, performance-, and interaction time-related implications of different conversational agents that vary in terms of their memory and levels of conversational initiative. The current work was motivated by the increasing overlap of the fields of HCI, Information Retrieval and the growing use of voice only conversational assistants and their commercial deployment for goal-oriented tasks. Previous research on conversational assistants identified the need for agents to provide more proactive support in order to increase usability (e.g. [126,154]). However, there is a lack of research that evaluates the impact of such support empirically. In line with the switch in evaluative focus from system usability to subjective user experience (as postulated by Dix [42]) - we evaluated user search experience from multiple angles (cognition, satisfaction, performance, and interaction time) while accounting for participants' feedback to get a more complete overview of the impact of different conversational agents on user search experience. The work presented in this thesis seeks to offer a more robust understanding of the causal effect of using different conversational agents by evaluating them in simulated search tasks. By conducting three interactive user studies, we aimed to identify how conversational agents affect users, and how agents can be engineered to lead to the best user search experience.

8.1 Impact of Conversational Agents on User Search Experience

The overarching question that the research presented in this thesis sought to address was: '**How much would a truly conversational agent with the ability to preserve state (memory) and conversational initiative improve user search experience compared to a stateless, passive voice search system?** The question was addressed through three Wizard of Oz studies (described in Chapter 4, Chapter 5 and Chapter 6).

The first study (discussed in Chapter 4), explored the role of a conversational agent's memory on user search experience in a goal-oriented task. The empirical examination indicated that the ability of the agent to preserve conversational state significantly improves user search experience for all of its constituent aspects (as discussed in Section 4.4). We observed that using a conversational agent with memory was less cognitively taxing, led to higher satisfaction, improved performance (indicated by a higher number of optimal flight choices) and led to shorter task completion times as compared to a stateless conversational agent. Interestingly, we also observed that participants expressed a more positive sentiment towards the stateful agent by using more polite language and thanking the agent more frequently.

The second study (discussed in Chapter 5), focused on mixed-initiative by investigating how different levels of an agent's involvement in elicitation (passive vs. active) and revealment (passive vs. active) impacted on user search experience. We observed that although there was little difference in how different agents were perceived by participants (i.e. cognitive effort and satisfaction metrics (see Table 5.3)), participants performed significantly better when using Active CAs. The active conversational support led to a higher number of optimal flight options being selected (see Table 5.4) and less time and money being wasted by participants (see Table 5.5). The study provided empirical evidence of the benefits of active conversational support in a voice-only, goal-oriented task.

The third study (discussed in Chapter 6), further explored the role of mixed-initiative by focusing on the proactive support of a conversational agent in eliciting user search criteria and recommending results (proactive revealment). Although, as illustrated in Section 6.4, we found no significant differences in subjective measures (cognitive load and satisfaction) or participants' performance, the results indicate that proactive support enabled participants to complete their search task significantly faster (cf. Table 6.8) *while* maintaining a good level of performance (i.e. a high proportion of Pareto Optimal flights being booked (see Table 6.4 for details)). The third study provided evidence that moving the conversational initiative more towards the agent (active to proactive support) can yield further benefits for user search experience.

Taken together all three studies indicate that a CA's memory and increased level of conversational initiative (proactive support) lead to improved user search experience. Throughout all the three studies conducted, the measures employed in this research to evaluate 'user search experience' showed consistent results. Agents providing a higher level of support consistently led to a better user search experience.

8.1.1 Design Recommendations for Conversational Agents

Based on the results obtained from our Wizard of Oz studies, we would like to highlight several important factors that should be considered when designing CAs for goal-orientated search through a voice only channel.

1. **Be Proactive:** While a CA should generally help a user to orient their search direction by proactively eliciting their search criteria and providing alternative results suggestions (proactive revealment), this should not always be universally applied (see point 2 below). This recommendation is based on the superior performance of the Proactive CAs (see Tables 6.4 and 6.5) and supported by the feedback provided by participants (see Table 6.9 and Section 6.5.2.1).

2. **Provide control over support:** Regardless of the Proactive nature of a CA, the user should have the option to adjust the level of support (agent's initiative) to carry out an individual query refinement when needed (switching filtering suggestions on and off). 'Controlling the Scope of the Query' was one of the functionalities most frequently requested by more than half of our participants (see Table 5.5).

3. **Introduce bookmarking:** The CA should retain flights in its memory for future comparison to reduce cognitive load and improve comparability of flights. This, in turn, will lead to the user making a more informed choice. – This recommendation is motivated by aspects of interactions that were identified by participants as challenging (see Table 5.5 and Section 6.5.2.3 for details).

4. **Provide the minimally salient details:** Rather than providing attribute ranges (unless explicitly requested), a CA should start with providing flights details based on a specific attribute (e.g. price or departure time). Otherwise, the user is likely to be overwhelmed by additional details. This recommendation is based on our NASA TLX results (see Tables 5.3 and 6.3 for details).

8.2 Contributions

Taken together, the research presented in this thesis makes the following contributions. Firstly, it provides empirical (interaction-based) insights into the impact of different conversational agents on user search experience in a voice-only, goal-oriented task - validating the Theoretical Framework of Conversational Search proposed by Radlinski and Craswell [126] which highlighted the crucial importance of agent memory and mixed-initiative. Secondly, it proposes a novel method for evaluating goal oriented conversational agents, combining a set of subjective and objective metrics that provide insights into user search experience. Thirdly, it proposes a method of evaluating conversational search performance in respect to the Pareto frontier which provides a way to evaluate whether participants engage in a satisficing or maximising search behaviour. Finally, it demonstrates that the design and complexity of the simulated search task has a bearing on the experimental outcomes.

8.3 Limitations

Although the research presented in this thesis provides much needed insight into impact of conversational agents on user search experience, there are limitations that need to be noted. One of the limitations of this research is the relative homogeneity of the samples in terms of age distributions. These are however reflective of the population being studied in terms of the demographics of conversational interfaces users.[1]

A further limitation of this research lies in the controlled experimental nature of the methods used. The desire for controlled and consistent conditions in terms of providing users with predefined search scenarios does not truly reflect the realism of the interaction, however it is an inherent difficulty of experiment-based research [42]. Nonetheless, the controlled nature of the research presented in this thesis is also one of its main strengths. Experiment based research is vitally important as it allows for standardisation and provides control over potential confounds and aspects that can be included in the design to improve the ecological validity CAs evaluation. For example, participants were given tasks that were relevant to the representative user group (flight search scenarios with background stories to make them more relatable). Additionally, efforts were made to control the effects of condition order, task order and other confounds which may have influenced the measures in the research in each experiment (more details about the specific efforts of confound reduction in each study can be found in Chapters 4, 5 and 6).

We are also mindful of the fact that the within-subjects experimental design used in our studies could have made the distinction of individual agent strategies more difficult, as indicated by participants' comments during semi-structured interviews. While a between-subjects design

[1] As indicated in a recent report: https://voicebot.ai/2019/06/21/voice-assistant-demographic-data-young-consumers-more-likely-to-own-smart-speakers-while-over-60-bias-toward-alexa-and-siri/ (last accessed on the 7th November 2020)

could have addressed this issue it would have made a comparison of results between the groups more difficult as participants would not have been exposed to different agent strategies. A between-subjects design would also have necessitated larger sample sizes which would have made the recruitment impossible given the limitations imposed by the experimental budget.

Another limitation of this work that needs to be acknowledged is learning effect involved in using different CAs. In the current work, we tried to control for learning effects by using a Latin square rotation of tasks and agents. It is worth noting that we did not observe any significant differences in behaviour or performance stemming from the ordering. However, we do acknowledge that further investigation is required to understand how quickly participants can learn to efficiently and effectively use CAs to perform tasks.

Regardless of the above limitations, the insights gathered from three interactive user studies allow for more certain causal conclusions about the effect of specific variables of a conversational search agent on user search experience, namely (1) memory (agent statefulness), (2) Active elicitation and revealment strategies and (3) Proactive elicitation and revealment strategies. Combining qualitative and quantitative methods arguably leads to higher ecological validity. Importantly, experiment-based research presented in the current thesis offers the benefit of replicable methodology so that findings using the same methods across different scenarios can be compared and contrasted.

8.4 Future Work

There are several research avenues that could be explored in future work on the interactive evaluation of goal-oriented conversational agents.

Firstly, it would be important to consider how user search experience and agent preferences differ between participants based on their cognitive abilities. To reflect that, any future experiment design should involve questionnaires and memory tests in the evaluation procedure.

Secondly, it would also be interesting to investigate the role of performance based incentives on the number of Pareto Optimal flights selected by participants. In the current work all participants received the same financial reward for their participation. While motivating participants by providing performance-based incentives could lead to greater search involvement, it is important to consider if such a way of motivating participants could consequently lead to more biased experimental outcomes and compromise ecological validity. Thus, it is important to carefully design search tasks and rewards that could reflect a more realistic user behaviour which is more reflective of real-life.

Thirdly, research should also consider other goal-oriented domains that lie outside the flight booking tasks, to investigate if they can yield similar results. Specifically, the evaluation should look into the concept of Pareto Optimality in more detail to explore tensions between different

aspects of products where participants need to make a compromise. One potential domain for evaluation is the use of recommender systems for commercial products and services.
Fourthly, since people are generally more used to graphical interfaces, it would be interesting to see if participants can improve their performance over time as their experience with using a voice-only conversational search system increases. This would necessitate conducting further longitudinal studies using a conversational agent prototype.
Fifthly, as the proactivity of agents and their conversational involvement comes at the expense of user agency in the search process, future work should consider the ethical implications of using proactive conversational agents and their social acceptability. This could include issues of transparency of results presentation and explainability of actions taken by the agent during the search process.
Finally, future experiments could also explore how experimental outcomes are affected by different types of synthetic voices that vary in prosodic qualities. In Study 1 we used a simple prompt console to provide search results via female synthetic voice with a Scottish accent. However, with a growing number of APIs, experimenting with a large spectrum of voices and models, and creating more complex search scenarios is becoming ever more feasible.

8.5 Closing Remarks

We would like to close with a quotation from Dix that highlights the complementarity of quantitative and qualitative measures in system evaluation.

> "Quantitative end-to-end measures are good at telling you *whether* there is an effect and h*ow strong* it is, but it is the qualitative data that helps you understand *why* you are seeing the phenomenon." Dix [42, p.19]

In this spirit, the current thesis combined both quantitative and qualitative measures to provide a better understanding of different conversational agents on user search experience. We believe that this approach has strong potential to inform the design of CAs that are not only efficient and usable but also user-centered.

Bibliography

[1] Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[2] Aliannejadi, M., Zamani, H., Crestani, F., and Croft, W. B. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2019), pp. 475–484.

[3] Anand, A., Cavedon, L., Joho, H., Sanderson, M., and Stein, B. Conversational search (dagstuhl seminar 19461). In *Dagstuhl Seminar 19461* (2020), Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[4] Andolina, S., Orso, V., Schneider, H., Klouche, K., Ruotsalo, T., Gamberini, L., and Jacucci, G. Investigating proactive search support in conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference* (2018), ACM, pp. 1295–1307.

[5] Andrews, P., and Quarteroni, S. Extending conversational agents for task-oriented human-computer dialogue. In *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. IGI Global, 2011, pp. 177–202.

[6] Antonak, R. F., and Livneh, H. Direct and indirect methods to measure attitudes toward persons with disabilities, with an exegesis of the error-choice test method. *Rehabilitation Psychology 40*, 1 (1995), 3.

[7] Asri, L. E., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., Mehrotra, R., and Suleman, K. Frames: A corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057* (2017).

[8] Austin, J. L. *How to do things with words*, vol. 88. Oxford university press, 1975.

[9] Avula, S., Chadwick, G., Arguello, J., and Capra, R. Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 conference on human information interaction & retrieval* (2018), pp. 52–61.

[10] Aylett, M. P., and Pidcock, C. J. The cerevoice characterful speech synthesiser sdk. In *IVA* (2007), pp. 413–414.

[11] Azzopardi, L., Dubiel, M., Halvey, M., and Dalton, J. Conceptualizing agent-human interactions during the conversational search process. In *The Second International Workshop on Conversational Approaches to Information Retrieval* (2018).

[12] Baddeley, A. The episodic buffer: a new component of working memory? *Trends in cognitive sciences 4*, 11 (2000), 417–423.

[13] Baddeley, A. D., and Hitch, G. Working memory. In *Psychology of learning and motivation*, vol. 8. Elsevier, 1974, pp. 47–89.

[14] Baeza-Yates, R., Ribeiro-Neto, B., et al. *Modern information retrieval*, vol. 463. ACM press New York, 1999.

[15] Barko-Sherif, S., Elsweiler, D., and Harvey, M. Conversational agents for recipe recommendation. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (2020), pp. 73–82.

[16] Bates, M. J. Where should the person stop and the information search interface start? *Information Processing & Management 26*, 5 (1990), 575–591.

[17] Belkin, N. J. Anomalous states of knowledge as a basis for information retrieval. *Canadian journal of information science 5*, 1 (1980), 133–143.

[18] Berg, M. M. *Modelling of Natural Dialogues in the Context of Speech-based Information and Control Systems*. AKA, 2014.

[19] Biernacki, P., and Waldorf, D. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research 10*, 2 (1981), 141–163.

[20] Bigné, E., Aldás, J., and Hyder, A. Engagement with travel web sites and the influence of online comparative behaviour. In *Cultural Perspectives in a Global Marketplace*. Springer, 2015, pp. 26–33.

[21] Borlund, P. Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation 56*, 1 (2000), 71–90.

[22] Borlund, P. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research. An International Electronic Journal 8*, 3 (2003).

[23] Bradley, J. V. Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association 53*, 282 (1958), 525–528.

[24] Brooke, J., et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry 189*, 194 (1996), 4–7.

[25] Button, G., Lee, J. R., Coulter, J., and Sharrock, W. *Computers, minds and conduct*. Blackwell Publishers, Inc., 1995.

[26] Chen, B., and Metz, C. Google's duplex uses ai to mimic humans (sometimes). *The New York Times* (2019).

[27] Christakopoulou, K., Radlinski, F., and Hofmann, K. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), ACM, pp. 815–824.

[28] Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., and Cowan, B. The state of speech in hci: Trends, themes and challenges. *arXiv preprint arXiv:1810.06828* (2018).

[29] Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., et al. What makes a good conversation?: Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, p. 475.

[30] Cleverdon, C. W., Mills, J., and Keen, E. M. Factors determining the performance of indexing systems,(volume 1: Design). *Cranfield: College of Aeronautics* (1966), 28.

[31] Cohen, M. H., Cohen, M. H., Giangola, J. P., and Balogh, J. *Voice user interface design*. Addison-Wesley Professional, 2004.

[32] Cohen, P. R. Back to the future for dialogue research: A position paper. *arXiv preprint arXiv:1812.01144* (2018).

[33] Commarford, P. M., Lewis, J. R., Smither, J. A.-A., and Gentzler, M. D. A comparison of broad versus deep auditory menu structures. *Human Factors 50*, 1 (2008), 77–89.

[34] Cook, M. J., Cranmer, C., Finan, R., Sapeluk, A., and Milton, C.-A. 15 memory load and task interference: hidden usability issues in speech interfaces. *Engineering Psychology and Cognitive Ergonomics: Volume 1: Transportation Systems* (2017).

[35] Cowan, B. R., et al. Understanding speech and language interactions in hci: The importance of theory-based human-human dialogue research. In *Designing speech and language interactions workshop, ACM conference on human factors in computing systems, CHI* (2014).

[36] Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., and Bandeira, N. What can i help you with?: infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2017), ACM, p. 43.

[37] Croft, W. B., and Thompson, R. H. I3r: A new approach to the design of document retrieval systems. *Journal of the american society for information science 38*, 6 (1987), 389–404.

[38] Dahlbäck, N., Jönsson, A., and Ahrenberg, L. Wizard of oz studies—why and how. *Knowledge-based systems 6*, 4 (1993), 258–266.

[39] Daronnat, S., Azzopardi, L., Halvey, M., and Dubiel, M. Impact of agent reliability and predictability on trust in real time human-agent collaboration. In *Proceedings of the 8th International Conference on Human-Agent Interaction* (2020), pp. 131–139.

[40] Demberg, V., and Moore, J. D. Information presentation in spoken dialogue systems. In *11th conference of the european chapter of the association for computational linguistics* (2006).

[41] Demberg, V., Winterboer, A., and Moore, J. D. A strategy for information presentation in spoken dialog systems. *Computational Linguistics 37*, 3 (2011), 489–539.

[42] Dix, A. Human–computer interaction: A stable discipline, a nascent science, and the growth of the long tail. *Interacting with computers 22*, 1 (2010), 13–27.

[43] Doyle, J. Rationality and its roles in reasoning. *Computational Intelligence 8*, 2 (1992), 376–409.

[44] Doyle, P. R., Edwards, J., Dumbleton, O., Clark, L., and Cowan, B. R. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (2019), pp. 1–12.

[45] Dubiel, M. Becoming digital: Toward a post-internet society. *Journal of Enabling Technologies* (2018).

[46] Dubiel, M. Towards human-like conversational search systems. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (2018), pp. 348–350.

[47] Dubiel, M. Towards human-like conversational search systems. *CHIIR '18: 2018 Conference on Human Information Interaction & Retrieval* (2018).

[48] Dubiel, M., Cervone, A., and Riccardi, G. Inquisitive mind: a conversational news companion. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (2019), pp. 1–3.

[49] Dubiel, M., Halvey, M., and Azzopardi, L. A survey investigating usage of virtual personal assistants. *arXiv preprint arXiv:1807.04606* (2018).

[50] Dubiel, M., Halvey, M., Azzopardi, L., Anderson, D., and Sylvain, D. Conversational strategies: Impact on search performance in a goal-oriented task. In *3rd International Workshop on Conversational Approaches to Information Retrieval (CAIR'20)* (2020), ACM, pp. 1–7.

[51] Dubiel, M., Halvey, M., Azzopardi, L., Aylett, M., Wester, M., and Braude, D. A. Improving conversational dynamics with reactive speech synthesis. In *Voice-based Conversational UX Studies and Design Workshop* (2018).

[52] Dubiel, M., Halvey, M., Azzopardi, L., and Daronnat, S. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *The Second International Workshop on Conversational Approaches to Information Retrieval* (2018).

[53] Dubiel, M., Halvey, M., Azzopardi, L., and Daronnat, S. Interactive evaluation of conversational agents: reflections on the impact of search task design. In *ACM SIGIR International Conference on the Theory of Information Retrieval 2020* (2020).

[54] Dubiel, M., Nakayama, M., and Wang, X. Evaluating synthetic speech workload with oculo-motor indices: preliminary observations for japanese speech. *BIOSIGNALS 2021* (2021).

[55] Dubiel, M., Oplustil, P., Halvey, M., and King, S. Persuasive synthetic speech: Voice perception and user behaviour. — (2020), 8.

[56] Ellis, D. A behavioural approach to information retrieval system design. *Journal of documentation* (1989).

[57] Feldman, S. E. The answer machine. *Synthesis Lectures on Information Concepts, Retrieval, and Services 4*, 3 (2012), 1–137.

[58] Fereday, J., and Muir-Cochrane, E. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods 5*, 1 (2006), 80–92.

[59] Flanagan, J. C. The critical incident technique. *Psychological bulletin 51*, 4 (1954), 327.

[60] Frummet, A., Elsweiler, D., and Ludwig, B. Detecting domain-specific information needs in conversational search dialogues. In *University of Regensburg* (2019).

[61] Gao, J., Galley, M., Li, L., et al. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval 13*, 2-3 (2019).

[62] Garfinkel, H. *Studies in Ethnomethodology*. Prentice-Hall, 1967.

[63] Georgila, K., Henderson, J., and Lemon, O. Learning user simulations for information state update dialogue systems. In *Ninth European Conference on Speech Communication and Technology* (2005).

[64] Georgila, K., Lemon, O., and Henderson, J. Automatic annotation of communicator dialogue data for learning dialogue strategies and user simulations. In *Ninth Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL: DIALOR)* (2005), Citeseer.

[65] Gilbert, N., Wooffitt, R., and Fraser, N. Organising computer talk. In *Computers and conversation*. Elsevier, 1990, pp. 235–257.

[66] Go, K., and Carroll, J. M. The blind men and the elephant: Views of scenario-based system design. *interactions 11*, 6 (2004), 44–53.

[67] Grice, et al. Logic and conversation. *1975* (1975), 41–58.

[68] Hart, S. G. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (2006), vol. 50, Sage Publications Sage CA: Los Angeles, CA, pp. 904–908.

[69] Hart, S. G., and Staveland, L. E. Development of nasa-tlx (task load index). *Advances in psychology 52* (1988), 139–183.

[70] Henderson, J., Lemon, O., and Georgila, K. Hybrid reinforcement/supervised learning for dialogue policies from communicator data. In *IJCAI workshop on knowledge and reasoning in practical dialogue systems* (2005), Citeseer, pp. 68–75.

[71] Herse, S., Vitale, J., Johnston, B., and Williams, M.-A. Using trust to determine user decision making & task outcome during a human-agent collaborative task. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (2021), pp. 73–82.

[72] Holmes, J. *An introduction to sociolinguistics*. Routledge, 2013.

[73] Hutto, C. J., and Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media* (2014).

[74] Järvelin, K. User-oriented evaluation in ir. In *PROMISE Winter School* (2012), Springer, pp. 86–91.

[75] JEFFERSON, G. Side sequences. *Studies in social interaction* (1972).

[76] Joho, H., Cavedon, L., Arguello, J., Shokouhi, M., and Radlinski, F. First international workshop on conversational approaches to information retrieval (cair'17). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017), ACM, pp. 1423–1424.

[77] Joho, H., Cavedon, L., Arguello, J., Shokouhi, M., and Radlinski, F. Cair'17: First international workshop on conversational approaches to information retrieval at sigir 2017. In *ACM SIGIR Forum* (2018), vol. 51, ACM New York, NY, USA, pp. 114–121.

[78] Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551* (2017).

[79] Jurafsky, D., and Martin, J. H. Dialog systems and chatbots. *Speech and language processing 3* (2019).

[80] KAYAK. Kayak skill (amazon echo application software]), 2018. Acessed: 10th September 2019, Retrieved from: https://www.amazon.co.uk/KAYAK/dp/B01EILLOXI.

[81] Kelly, D., et al. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval 3*, 1–2 (2009), 1–224.

[82] Kiesel, J., Bahrami, A., Stein, B., Anand, A., and Hagen, M. Clarifying false memories in voice-based search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (2019), pp. 331–335.

[83] Kiesel, J., Lang, K., Wachsmuth, H., Hornecker, E., and Stein, B. Investigating expectations for voice-based and conversational argument search on the web. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (2020), pp. 53–62.

[84] Kirschthaler, P., Porcheron, M., and Fischer, J. E. What can i say? effects of discoverability in vuis on task performance and user experience. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (2020), pp. 1–9.

[85] Kiseleva, J., and de Rijke, M. Evaluating personal assistants on mobile devices. *arXiv preprint arXiv:1706.04524* (2017).

[86] Kiseleva, J., Williams, K., Hassan Awadallah, A., Crook, A. C., Zitouni, I., and Anastasakos, T. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (2016), ACM, pp. 45–54.

[87] Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A. C., Zitouni, I., and Anastasakos, T. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (2016), ACM, pp. 121–130.

[88] Kobie, N. Google's new voice is as good as your own, 2018.

[89] Kuhlthau, C. C. Developing a model of the library search process: Cognitive and affective aspects. *Rq* (1988), 232–242.

[90] Laubheimer, P., and Budiu, R. Intelligent assistants: Creepy, childish, or a tool? users' attitudes toward alexa, google assistant, and siri. *Nielsen Norman Group* (2018).

[91] Lee, S.-s., Lee, J., and Lee, K.-p. Designing intelligent assistant through user participations. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (2017), ACM, pp. 173–177.

[92] Leviathan, Y., and Matias, Y. Google ai blog: Google duplex: an ai system for accomplishing real-world tasks over the phone. *Google Blog* (2018).

[93] Levin, E., Narayanan, S., Pieraccini, R., Biatov, K., Bocchieri, E., Fabbrizio, G. D., Eckert, W., Lee, S., Pokrovsky, A., Rahim, M., et al. The at&t-darpa communicator mixed-initiative spoken dialog system. In *Sixth International Conference on Spoken Language Processing* (2000).

[94] Lindquist, E. F. Design and analysis of experiments in psychology and education. *American Psychological Association* (1953).

[95] Lison, P., and Meena, R. Spoken dialogue systems: the new frontier in human-computer interaction. *XRDS: Crossroads, The ACM Magazine for Students 21*, 1 (2014), 46–51.

[96] Longo, L. Subjective usability, mental workload assessments and their impact on objective human performance. In *IFIP Conference on Human-Computer Interaction* (2017), Springer, pp. 202–223.

[97] Loper, E., and Bird, S. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1* (2002), Association for Computational Linguistics, pp. 63–70.

[98] Luger, E., and Sellen, A. Like having a really bad pa: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 5286–5297.

[99] Macagno, F. Types of dialogue, dialectical relevance, and textual congruity. *Anthropology & Philosophy 8*, 1-2 (2000), 101–121.

[100] Macagno, F., and Bigi, S. Analyzing the pragmatic structure of dialogues. *Discourse Studies 19*, 2 (2017), 148–168.

[101] Marchionini, G. *Information seeking in electronic environments*. Cambridge university press, 1997.

[102] McDuff, D., Thomas, P., Czerwinski, M., and Craswell, N. Multimodal analysis of vocal collaborative search: a public corpus and results. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (2017), ACM, pp. 456–463.

[103] McTear, M., Callejas, Z., and Griol, D. Evaluating the conversational interface. In *The Conversational Interface*. Springer, 2016, pp. 379–402.

[104] Moore, J. D., Foster, M. E., Lemon, O., and White, M. Generating tailored, comparative descriptions in spoken dialogue. In *FLAIRS Conference* (2004), pp. 917–922.

[105] Mortensen, D. How to design voice user interfaces. *Interaction Design Foundation* (2017).

[106] Mu, J., and Sarkar, A. Do we need natural language?: Exploring restricted language interfaces for complex domains. In *CHI Extended Abstracts* (2019).

[107] Nass, C., Steuer, J., and Tauber, E. R. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1994), pp. 72–78.

[108] Nielsen, J. Interviewing users. *Nielsen Norman Group* (2010).

[109] Nolan, C. Interstellar. Warner Bros. Pictures (International), 2014.

[110] Nouri, E., Sim, R., Fourney, A., and White, R. W. Proactive suggestion generation: Data and methods for stepwise task assistance. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020), pp. 1585–1588.

[111] Novielli, N., and Strapparava, C. Dialogue act classification exploiting lexical semantics. In *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. IGI Global, 2011, pp. 80–106.

[112] Oddy, R. N. Information retrieval through man-machine dialogue. *Journal of documentation 33*, 1 (1977), 1–14.

[113] Pearl, C. *Designing Voice User Interfaces: Principles of Conversational Experiences.* " O'Reilly Media, Inc.", 2016.

[114] Pelikan, H. R., and Broth, M. Why that nao? how humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (2016), pp. 4921–4932.

[115] Perez-Marin, D. *Conversational agents and natural language interaction: Techniques and effective practices: Techniques and effective practices.* IGI Global, 2011.

[116] Pieraccini, R., Tzoukermann, E., Gorelov, Z., Gauvain, J.-L., Levin, E., Lee, C.-H., and Wilpon, J. G. A speech understanding system based on statistical representation of semantics. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* (1992), vol. 1, IEEE, pp. 193–196.

[117] Polifroni, J., Chung, G., and Seneff, S. Towards the automatic generation of mixed-initiative dialogue systems from web content. In *Eighth European Conference on Speech Communication and Technology* (2003).

[118] Polifroni, J., Hirschman, L., Seneff, S., and Zue, V. Experiments in evaluating interactive spoken language systems. Tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE LAB FOR COMPUTER SCIENCE, 1992.

[119] Polifroni, J., and Walker, M. Intensional summaries as cooperative responses in dialogue: Automation and evaluation. In *Proceedings of ACL-08: HLT* (2008), pp. 479–487.

[120] Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems* (2018), pp. 1–12.

[121] Porcheron, M., Fischer, J. E., and Sharples, S. "do animals have accents?": talking with agents in multi-party conversation. *CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social* (2017).

[122] PWC. 2015 iata global passenger survey. *IATA Survey* (2015).

[123] Qu, C., Yang, L., Croft, W. B., Zhang, Y., Trippas, J. R., and Qiu, M. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (2019), ACM, pp. 25–33.

[124] R., J. Informing the design of spoken conversational search. *CHIIR '18: 2018 Conference on Human Information Interaction & Retrieval* (2018).

[125] Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2 (1989), 257–286.

[126] Radlinski, F., and Craswell, N. A theoretical framework for conversational search. In *CHIIR 2017* (2017), ACM, pp. 117–126.

[127] Reeves, S. Some conversational challenges of talking with machines. In *Companion of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (2017), pp. 431–436.

[128] Reichheld, F. F. The one number you need to grow. *Harvard business review 81*, 12 (2003), 46–55.

[129] Revlin, R. *Cognition: Theory and practice*. Macmillan, 2012.

[130] Roller, S., Boureau, Y.-L., Weston, J., Bordes, A., Dinan, E., Fan, A., Gunning, D., Ju, D., Li, M., Poff, S., et al. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442* (2020).

[131] Ruthven, I. Interactive information retrieval. *Annual review of information science and technology 42*, 1 (2008), 43–91.

[132] Sacks, H. *Lectures on Conversation*. Oxford: Blackwell, 1992.

[133] Salton, G. Evaluation problems in interactive information retrieval. *Information Storage and Retrieval 6*, 1 (1970), 29–44.

[134] Saracevic, T. The stratified model of information retrieval interaction: Extension and applications. In *Proceedings of the annual meeting-american society for information science* (1997), vol. 34, LEARNED INFORMATION (EUROPE) LTD, pp. 313–327.

[135] Sauro, J. Measuring usability with the system usability scale (sus), 2011.

[136] Schatzmann, J., Weilhammer, K., Stuttle, M., and Young, S. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Engineering Review 21*, 2 (2006), 97–126.

[137] Schatztnann, J., Stuttle, M. N., Weilhammer, K., and Young, S. Effects of the user model on simulation-based learning of dialogue strategies. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* (2005), IEEE, pp. 220–225.

[138] Schmandt, C. *Voice communication with computers: conversational systems.* Van Nostrand Reinhold Co., 1994.

[139] Schulz, H., Zumer, J., Asri, L. E., and Sharma, S. A frame tracking model for memory-enhanced dialogue systems. *arXiv preprint arXiv:1706.01690* (2017).

[140] Sculley, J. *Odyssey: Pepsi to Apple, a journey of adventure, ideas, and the future.* Harper & Row Publishers, Inc., 1987.

[141] Seneff, S., and Polifroni, J. Dialogue management in the mercury flight reservation system. In *ANLP-NAACL 2000 Workshop: Conversational Systems* (2000).

[142] Shackel, B. Usability–context, framework, definition, design and evaluation. *Interacting with computers 21*, 5-6 (2009), 339–346.

[143] Simon, H. A. Rational choice and the structure of the environment. *Psychological review 63*, 2 (1956), 129.

[144] Sitter, S., and Stein, A. Modeling the illocutionary aspects of information-seeking dialogues. *Information Processing & Management 28*, 2 (1992), 165–180.

[145] Skyscanner.net. Skyscanner flight search (amazon echo application software]), 2018. Accessed on 10th September, Retrieved from: https://tinyurl.com/y8fx2n3o.

[146] Stalnaker, R. Common ground. *Linguistics and philosophy 25*, 5/6 (2002), 701–721.

[147] Suro, J. 5 ways to interpret a sus score, sep 2018.

[148] Sutton, S. J., Foulkes, P., Kirk, D., and Lawson, S. Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, p. 603.

[149] Thomas, P., Czerwinski, M., McDuff, D., Craswell, N., and Mark, G. Style and alignment in information-seeking conversation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (2018), pp. 42–51.

[150] Thomas, P., McDuff, D., Czerwinski, M., and Craswell, N. Misc: A data set of information-seeking conversations. In *SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17)* (2017), vol. 5.

[151] Tokuda, K., Zen, H., and Black, A. W. An hmm-based speech synthesis system applied to english. In *IEEE Speech Synthesis Workshop* (2002), pp. 227–230.

[152] Toms, E. G. Task-based information searching and retrieval. *Interactive information seeking, behaviour and retrieval* (2011), 43–59.

[153] Tortoreto, G., Stepanov, E. A., Cervone, A., Dubiel, M., and Riccardi, G. Affective behaviour analysis of on-line user interactions: Are on-line support groups more therapeutic than twitter? *ACL 2019* (2019), 79.

[154] Trippas, J. Spoken conversational search: audio-only interactive information retrieval. *PhD Thesis* (2019).

[155] Trippas, J. R. Spoken conversational search: Information retrieval over a speech-only communication channel. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2015), ACM, pp. 1067–1067.

[156] Trippas, J. R., Spina, D., Cavedon, L., Joho, H., and Sanderson, M. Informing the design of spoken conversational search: perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (2018), ACM, pp. 32–41.

[157] Trippas, J. R., Spina, D., Cavedon, L., and Sanderson, M. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (2017), ACM, pp. 325–328.

[158] Trippas, J. R., Spina, D., Sanderson, M., and Cavedon, L. Results presentation methods for a spoken conversational search system. In *Proceedings of the First International Workshop on Novel Web Search Interfaces and Systems* (2015), ACM, pp. 13–15.

[159] Trippas, J. R., Spina, D., Thomas, P., Sanderson, M., Joho, H., and Cavedon, L. Towards a model for spoken conversational search. *Information Processing & Management 57*, 2 (2020), 102162.

[160] Trippas, J. R., Thomas, P., Spina, D., and Joho, H. Third international workshop on conversational approaches to information retrieval (cair'20) full-day workshop at chiir 2020. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (2020), pp. 492–494.

[161] Vakulenko, S., Markov, I., and de Rijke, M. Conversational exploratory search via interactive storytelling. *arXiv preprint arXiv:1709.05298* (2017).

[162] Vakulenko, S., Revoredo, K., Di Ciccio, C., and de Rijke, M. Qrfa: A data-driven model of information-seeking dialogues. In *European Conference on Information Retrieval* (2019), Springer, pp. 541–557.

[163] Viera, A. J., Garrett, J. M., et al. Understanding interobserver agreement: the kappa statistic. *Fam med 37*, 5 (2005), 360–363.

[164] Vtyurina, A., and Fourney, A. Exploring the role of conversational cues in guided task support with virtual assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), ACM, p. 208.

[165] Vtyurina, A., Savenkov, D., Agichtein, E., and Clarke, C. L. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017), ACM, pp. 2187–2193.

[166] Walker, M., Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S., Narayanan, S., Papineni, K., et al. Darpa communicator dialog travel planning systems: The june 2000 data collection. In *Seventh European Conference on Speech Communication and Technology* (2001).

[167] Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. Paradise: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004* (1997).

[168] Walton, D., and Krabbe, E. C. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press, 1995.

[169] Walton, D. N., and Walton, D. N. *Informal logic: A handbook for critical argument*. Cambridge University Press, 1989.

[170] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv preprint arXiv:1703.10135* (2017).

[171] Ward, N. G., and DeVault, D. Ten challenges in highly-interactive dialog system. In *2015 AAAI Spring Symposium Series* (2015).

[172] White, R. W., and Ruthven, I. A study of interface support mechanisms for interactive information retrieval. *Journal of the American Society for Information Science and Technology 57*, 7 (2006), 933–948.

[173] Whitenton, K., and Budiu, R. The paradox of intelligent assistants: Poor usability, high adoption, 2019.

[174] Williams, J., Raux, A., and Henderson, M. The dialog state tracking challenge series: A review. *Dialogue & Discourse 7*, 3 (2016), 4–33.

[175] Wilson, T. D. Models in information behaviour research. *Journal of documentation 55*, 3 (1999), 249–270.

[176] Winterboer, A., and Moore, J. D. Evaluating information presentation strategies for spoken recommendations. In *Proceedings of the 2007 ACM conference on Recommender systems* (2007), ACM, pp. 157–160.

[177] Wissbroecker, J., and Harper, F. M. Early lessons from a voice-only interface for finding movies. *arXiv preprint arXiv:1808.09900* (2018).

[178] Wolters, M., Georgila, K., Moore, J. D., Logie, R. H., MacPherson, S. E., and Watson, M. Reducing working memory load in spoken dialogue systems. *Interacting with Computers 21*, 4 (2009), 276–287.

[179] Wooffitt, R. *Conversation analysis and discourse analysis: A comparative and critical introduction*. Sage, 2005.

[180] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. The microsoft 2016 conversational speech recognition system. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (2017), IEEE, pp. 5255–5259.

[181] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., and Zweig, G. Towards human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2017).

[182] Yankelovich, N., Levow, G.-A., and Marx, M. Designing speechacts: Issues in speech user interfaces. In *CHI* (1995), vol. 95, pp. 369–376.

[183] Zhang, Y., Chen, X., Ai, Q., Yang, L., and Croft, W. B. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (2018), ACM, pp. 177–186.

[184] Zue, V., Seneff, S., Polifroni, J., Phillips, M., Pao, C., Goodine, D., Goddeau, D., and Glass, J. Pegasus: A spoken dialogue interface for on-line air travel planning. *Speech Communication 15*, 3-4 (1994), 331–340.

Appendix A

Participant Information Sheets and Consent Forms

A.1 Information Sheet for Study 1

Participant Information Sheet

Name of department: Computer and Information Sciences
Title of the study: Interacting with Virtual Personal Assistants via Voice

Ethics Approval No.: 611

Dear Participant,
You are invited to take part in a research study a research study conducted by Mr Mateusz Dubiel (a PhD student at The University of Strathclyde). The current information sheet describes the nature and content of the study. Before you decide whether to participate, it is important for you to understand why the study is being done, and what it will involve. Please take time to read the following information carefully. If you have any questions or need clarification on any aspect of the study, please ask.

What is involved?
The title of the current study is 'Interacting with Virtual Personal Assistants via Voice'. The study aims to explore usability of two different Virtual Personal Assistants (further referred to as 'assistants') for online search tasks completed using dialogue. Usability is the degree to which an assistant is fit to be used for the purpose it was created – in the context of the study the purpose is making flight reservations. The goal of the study is to find out your perceptions of two different assistants and their usability.

If you agree to take part in this study, you will be asked to complete four search tasks (two for each assistant). After completing the tasks for each assistant, you will be given two short feedback questionnaires to fill in. Each task will include a different search scenario. During the task, you will be interacting with the assistant to ask for information and provide required details when requested by the assistant. Your goal for each of the tasks is to complete your reservation in the most efficient manner, as per instructions provided for each of the tasks. Please note that the assistant does not handle interruption - so you need to wait for it to finish speaking before you can start to speak. The agent only provides audio feedback – there are no graphics involved. Your interactions will be audio recorded with your permission.

Once you have completed all of the search tasks you will be invited to take part in a short post-study discussion. The aim of the discussion is to provide you an opportunity to give feedback about the experiment and ask any questions that you may have. The discussion will take no longer than 10 minutes. Please note that you are welcome to ask questions before search and after search sessions but not during the main experiment. Upon completion of the study, you will be given a £10 Shopping voucher.

The estimated completion time for the whole study is about 60 minutes

Anonymity and Confidentiality
All information collected about you during the course of the study will be anonymised, and used exclusively for research purposes. A unique ID number (e.g. S1) will be used to identify you.

No personal details will be used in the current study. All data collected will be held in a secure and safe manner in accordance with the Data Protection Act 1998, and will only be accessible by the researchers: Mr. Mateusz Dubiel, Dr. Martin Halvey, and Dr. Leif Azzopardi

The following steps will be taken to ensure anonymity and confidentiality of information:

1. Audio recordings will be kept on a password-protected computer. The audio recordings will be transcribed and stored on The University of Strathclyde secure storage system *StrathCloud*

2. Consent Forms will be stored in a locked cabinet located in Computer and Information Science Department of The University of Strathclyde for the period of five years. Once this period is over the forms will be shredded and recycled.

Right to Withdraw
Your participation in the study is voluntary and you have the right to withdraw from it at any time if you wish without providing a reason.

Research Team contact details:

Lead Researcher	First Supervisor	Second Supervisor
Mr. Mateusz Dubiel University of Strathclyde 16, Richmond Street, Glasgow G1 1XQ. Email: mateusz.dubiel@strath.ac.uk	Dr. Martin Halvey University of Strathclyde 16, Richmond Street, Glasgow G1 1XQ. Email: martin.halvey@strath.ac.uk	Dr. Leif Azzopardi University of Strathclyde 16, Richmond Street, Glasgow G1 1XQ. Email: leif.azzopardi@strath.ac.uk 0141 548 3617

Consent Declaration

I have read and understand the procedures involved in this study; and I am happy to take part.

Name of Participant	Date	Signature
_____	_____	_____

Name of Researcher	Date	Signature
_____	_____	_____

The results of this study will be published as part of conference and journal submissions. If you would like to receive a copy once the results have been published, please provide your email address:

_____

Ethical Approval

This investigation has been granted an ethical approval by the Departmental Ethics Committee (Application ID: 611). If you have any questions/concerns, during or after the investigation, or wish to contact an independent person to whom any questions may be directed or further information may be sough from, please contact: Secretary to the University Ethics Committee, Research & Knowledge Exchange Services, University of Strathclyde, Graham Hills Building, 50 George Square, Glasgow, G1 1QE, Telephone: 0141 548 3707, email: ethics@strath.ac.uk

A.2 Information Sheet for Study 2

Participant Information Sheet

Name of department: Computer and Information Sciences
Title of the study: Evaluating Conversational Search Agents for Flight Booking

Ethics Approval No.: 791

Dear Participant,
My name is Mateusz Dubiel and I am a researcher in the Department of Computer and Information Sciences at the University of Strathclyde. My contact email is mateusz.dubiel@strath.ac.uk. I am currently conducting research into information retrieval and conversational search where I am investigating different conversational agents for flight search. Please take time to read the following information carefully. If you have any questions or need clarification on any aspect of the study, please ask.

What is the purpose of this investigation?
The aim of this experiment is to evaluate different conversational search agents for finding flights.

Do you have to take part?
Participation in the experiment is voluntary and it is your decision to take part in this investigation or not. You can refuse to participate and you can withdraw from the study at any time without any consequences.

What will you do in the project?
The study will consist of four search tasks. For each task, you will be talking to a flight search agent that will search flights. Your goal for each of the task is to find and select a flight as specified in task scenario. The task is considered to be completed once you have informed the agent about your selected flight. Your interactions will be recorded with your permission.

During the experiment, you will be asked to complete a number of questionnaires about your experience with the agent used. At the beginning of the study, you will also be asked to fill in a short demographics survey along with your experiences using flight booking websites (e.g. *SkyScanner, Expedia* etc.) and virtual personal assistants (*Amazon Echo, Google Home* etc.). At the end of the experiment, you will be invited to take part in a short discussion that will be an opportunity for you to ask questions and share your feedback.

Upon completion of the study, you will be given a £10 shopping voucher.

Who can take part in the project?
To qualify for the study, you must be fluent in English, over 18 and with no hearing impairment.

What are the potential risks to you in taking part?
There are no serious risks associated with this experiment, which will be conducted in an office environment.

What happens to the information in the project?
All information and data collected during the experiments will be anonymised and no personally identifying information will be associated with the data. All data collected in the experiment will be retained and used in project publications. Your data will remain confidential and your name will not appear in any published documents relating to the research conducted. The University of Strathclyde is registered with the Information

Commissioner's Office who implements the The General Data Protection Regulation (GDPR). All personal data on participants will be processed in accordance with the provisions of GDPR.

Thank you for reading this information – please ask any questions if you are unsure about what is written here.

What happens next?
If you are happy to be involved in this project, then please read and then complete the following consent form and then we can arrange the time and dates of the sessions. Otherwise, we thank you for your time and attention.

When the investigation has been completed and you have indicated as such on the consent form, we will provide a summary of our research findings to you.

Research Team contact details:

Lead Researcher	First Supervisor	Second Supervisor
Mr. Mateusz Dubiel University of Strathclyde 16, Richmond Street, Glasgow G1 1XQ. Email: mateusz.dubiel@strath.ac.uk	Dr. Martin Halvey University of Strathclyde 16, Richmond Street, Glasgow G1 1XQ. Email: martin.halvey@strath.ac.uk	Dr. Leif Azzopardi University of Strathclyde 16, Richmond Street, Glasgow G1 1XQ. Email: leif.azzopardi@strath.ac.uk 0141 548 3617

This investigation was granted ethical approval by the University of Strathclyde Ethics Committee.

If you have any questions/concerns, during or after the investigation, or wish to contact an independent person to whom any questions may be directed or further information may be sought from, please contact:

Computer & Information Sciences
Livingstone Tower
26 Richmond Street
Glasgow, G1 1XHTelephone: 0141 548 3707
Email: enquiries@cis.strath.ac.uk

Consent Form

Name of department: Computer and Information Sciences
Title of the study: Evaluating Conversational Search Agents for Flight Booking

Please indicate whether you agree to the following statements (please initial each box if you agree):

- ☐ I confirm that I have read and understood the information sheet for the above project and the researcher has answered any queries to my satisfaction.
- ☐ I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of completion, without having to give a reason and without any consequences. If I exercise my right to withdraw and I don't want my data to be used, any data, which has been collected about me, will be destroyed.
- ☐ I understand that I can withdraw from the study any personal data (i.e. data which identify me personally) at any time.
- ☐ I understand that anonymised data (i.e. data which do not identify me personally) cannot be withdrawn once they have been included in the study.
- ☐ I understand that any information recorded in the investigation will remain confidential and no information that identifies me will be made publicly available.
- ☐ I consent to being a participant in the project.
- ☐ I consent to being audio and/or video recorded as part of the project.

Name of Participant	**Date**	**Signature**
_____	_____	_____
Name of Researcher	**Date**	**Signature**
_____	_____	_____

The results of this study will be published as part of conference and journal submissions. If you would like to receive a copy once the results have been published, please provide your email address:

_____

A.3 Information Sheet for Study 3

Participant Information Sheet

Name of department: Computer and Information Sciences
Title of the study: *Evaluating the Impact of Conversational Agent's for Flight Search*

Ethics Approval No.: *1155*

Dear Participant,
My name is Mateusz Dubiel, I am a researcher in the Department of Computer and Information Sciences at the University of Strathclyde. Currently, I am conducting research into information retrieval and conversational search where I am investigating different conversational agents for flight search. Please take time to read the following information carefully. If you have any questions or need clarification on any aspect of the study, please ask.

What is the purpose of this investigation?
The aim of this study is to evaluate different conversational search agents for finding flights.

Do you have to take part?
Participation in the study is voluntary. You can refuse to participate, and you can withdraw from the study at any time. However, to receive compensation you would need to diligently complete all provided tasks.

What will you do in the project?
The study will consist of several interactive search tasks that will be conducted via Zoom. For each task, you will be talking to a flight search agent that will search flights. Your goal for each of the task is to find and select a flight as specified in task scenario. The task is considered to be completed once you have informed the agent about your selected flight. Your interactions will be audio and video-recorded with your permission.

During the study, you will be asked to fill in questionnaires about your experience with the conversational agent used. All of the questionnaires will be deployed in *Qualtrics* (an online survey platform) and accessible via links sent to you at upon completion of each search task. At the beginning of the study, you will also be asked to fill in a short demographics survey along with your experiences using flight booking websites (e.g. SkyScanner, Expedia etc.) and conversational assistants (Amazon Echo, Google Home etc.) The demographics survey will also be deployed via *Qualtrics*. At the end of the study, you will be invited to take part in a short online discussion (conducted via Zoom) that will be an opportunity for you to ask questions and share your feedback.

Upon completion of the study, you will be given a link that will enable you to claim your £10 payment.

Who can take part in the project?
To qualify for the study, you must be fluent in English, over 18 and with no hearing impairment.

What are the potential risks to you in taking part?
There are no known risks associated with this study.

What happens to the information in the project?
All information and data collected during the study will be anonymised and no personally identifying information will be associated with the data. All data collected in the study will be retained and used in project publications. Your data will remain confidential and your name will not appear in any published documents relating to the research conducted. The University of Strathclyde is registered with the Information Commissioner's Office who implements the The General Data Protection Regulation (GDPR). All personal data on participants will be processed in accordance with the provisions of GDPR

What happens next?
If you are happy to be involved in this study, then please read and then complete then read the consent form and pick one of the available experiment slots.

Research Team contact details:

Lead Researcher	First Supervisor	Second Supervisor
Mr. Mateusz Dubiel University of Strathclyde 16, Richmond Street, Glasgow G1 1XQ. Email: mateusz.dubiel@strath.ac.uk	Dr. Martin Halvey University of Strathclyde 16, Richmond Street, Glasgow G1 1XQ. Email: martin.halvey@strath.ac.uk	Dr. Leif Azzopardi University of Strathclyde 16, Richmond Street, Glasgow G1 1XQ. Email: leif.azzopardi@strath.ac.uk

This investigation was granted ethical approval by the University of Strathclyde Ethics Committee.

If you have any questions/concerns, during or after the investigation or wish to contact an independent person to seek further information, please contact:

Computer & Information Sciences
Livingstone Tower
26 Richmond Street
Glasgow, G1 1XHTelephone: 0141 548 3707
Email: enquiries@cis.strath.ac.uk

Evaluating Conversational Agents for Flight Search

Welcome to the study.

Just to remind you, the Zoom call is recorded.

Please read the following terms and click on '>>' button if you agree with all of terms and wish to proceed.

☐ I consent to audio and video data being recorded during the study.

☐ I understand that anonymised data (i.e. data which do not identify me personally) cannot be withdrawn once they have been included in the study.

☐ I understand data provided by me will used for analysis, and that the results of such analysis may be presented at conferences and/or other scientific events.

☐ Upon successful completion of the study (completing all search tasks), I will receive a confirmation link that will enable me to claim the payment for taking part in the study.

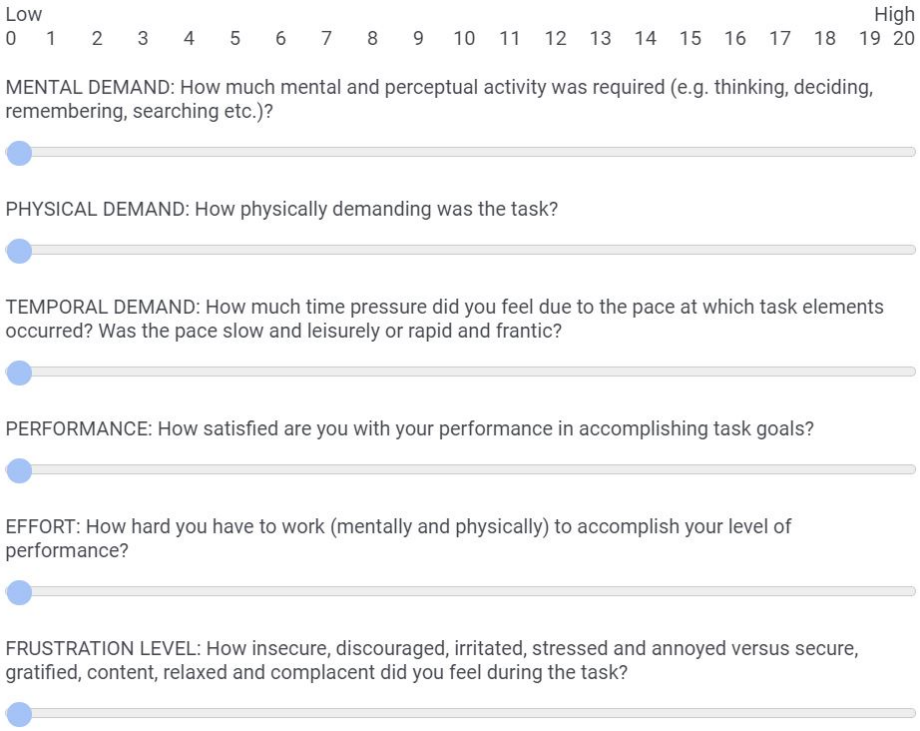
☐ I consent to participate in the study.

>>

Appendix B

Questionnaires

B.1 NASA TLX Questionnaire



B.2 SUS Questionnaire

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
1. I think that I would like to use this system frequently.	○	○	○	○	○
2. I found the system unnecessarily complex.	○	○	○	○	○
3. I thought the system was easy to use.	○	○	○	○	○
4. I think that I would need the support of a technical person to be able to use the system.	○	○	○	○	○
5. I found the various functions in this system were well integrated.	○	○	○	○	○
6. I thought there was too much inconsistency in this system.	○	○	○	○	○
7. I would imagine that most people would learn to use this system very quickly.	○	○	○	○	○
8. I found the system very cumbersome to use.	○	○	○	○	○
9. I felt very confident using the system.	○	○	○	○	○
10. I needed to learn a lot of things before I could get going with the system.	○	○	○	○	○

B.3 SSES Questionnaire

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly Agree
1. I am satisfied with how the system informed me about available flights (the options were clearly presented).	○	○	○	○	○
2. I am satisfied with overview of available flights that the system presented.	○	○	○	○	○
3. I think that there may be better (cheaper and faster) flights that the system did not tell me about.	○	○	○	○	○
4. I am happy with the speed in which the system helped me to find my flight.	○	○	○	○	○

Appendix C

Interactive Dialogues - Excerpts

C.1 Passive-Summary Agent (PS)

Angus: Hello, my name is Angus. I am a flight search agent and I will help you to find your flight. To get started please provide me with your destination and day of departure.

Participant: Emm, I am going to Rome on a Friday in September. But it doesn't specify any specific Friday yet.

Angus: One moment, which date would you like to check first?

Participant: Friday fourteenth.

Angus: One moment please. There are thirty flights available. The prices range from one hundred and forty pounds to two hundred and fifteen pounds and flight durations range from three hours to six hours fifteen

Participant: Can you check Friday the twenty first instead?

Angus: One moment please. There are twenty-five flight available. The prices range from one hundred pounds to one hundred and seventy pounds and the flight durations range from four hours to six hours fifteen [. . .]

C.2 Active-Summary Agent (AS)

Calum: Hello my name is Calum, I am a flight search agent and I will help you to find your flight. To get started, please provide me with your destination and day of departure.

Participant: Hi Calum, the destination is London and the date of departure is between Wednesday the twelfth and Friday the fourteenth of December.

Calum: Ok, which day would you like to check first?

Participant: Wednesday.

Calum: One moment please. There are twenty-nine flights available. The prices range from seventy pounds to one hundred and forty and the durations range from one hour thirty to four hours fifteen. Would you like to filter by price?

Participant: Yes.

Calum: Ok, how would you like to filter?

Participant: Cheapest first, please.

Calum: One moment please. Do you have a maximum price in mind?

Participant: Ninety.

Calum: There are six flights available. The prices range from seventy pounds to ninety pounds and the durations range from three hours to four hours. Would you like to filter by duration?

Participant: No, Can I filter by arrival time?

Calum: Ok, how would you like to filter?

Participant: Arrive at or before ten a.m.

Calum: There are no flights that arrive at or before ten am. You may want to increase the arrival time.

Participant: Are there any that arrive at or before eleven am?

Calum: No. You may want to increase the arrival time.

Participant: Can I try Thursday? [. . .]

C.3 Passive-Listing Agent (PL)

Blair: Hello, my name is Blair, I am a flight search agent and I will help you to find your flight. To get started, please provide me with your destination and day of departure.

Participant: I would like to get Stockholm on Monday, the fifth of November.

Blair: One moment please. There are twenty-nine flights available. The cheapest flight costs one hundred and seventy pounds and takes four hours forty five and the fastest flight costs two hundred and twenty pounds and takes two hours thirty minutes.

Participant: What is the time of arrival for the cheapest flight?

Blair: The cheapest flight arrives at fifteen zero five.

Participant: Ok, how about Tuesday, the sixth of November?

Blair: One moment please. There are twenty seven flights available. The cheapest flight costs one hundred seventy pounds and takes three hours thirty and the fastest flight costs two hundred and thirty pounds and takes two hours.

Participant: What is the arrival time for the cheapest flight?

Blair: The cheapest flight arrives at eighteen zero zero.

Participant: And how about Wednesday, the seventh of November?

Blair: One moment, please. There are twenty-eight flights available. The cheapest flight costs one hundred and thirty pounds and takes four hours forty-five minutes and the fastest flight costs one hundred and eighty pounds and takes three hours.

Participant: What is the arrival time for the cheapest flight?

Blair: The arrival time for the cheapest flight is fourteen forty-five

Participant: And what is the price, again?

Blair: One hundred and thirty pounds.

Participant: And what was the arrival for the one on Tuesday the sixth of November that is the cheapest?

Blair: That was eighteen zero zero.

Participant: And the price?

Blair: One hundred and seventy pounds

Participant: Ok, and is there a cheaper flight... no, sorry. What is the cheapest option on the seventh again?

Blair: The cheapest flight on the seventh is one hundred and thirty pounds, it an Air France flight it departs at ten zero zero and arrives at fourteen forty-five, it has flight duration of four hours forty-five minutes.

Participant: Is there one that is below one fifty and arrives closer to seven pm?

Blair: One moment, please. The closest flight would be fifteen thirty it is a British Airways flight that departs at ten fifteen and arrives at fifteen thirty and flight duration is five hours fifteen and it costs one hundred and thirty pounds.

Participant: I would like to book that one.

C.4 Active-Listing Agent (AL)

David: Hello, my name is David. I am flight search agent and I will help you to find your flight. To get started please provide me with your destination and day of departure.

Participant: Stockholm, Monday the fifth.

David: One moment please, there are twenty nine flights available. The cheapest flight costs one hundred and seventy pounds and takes four hours forty five and the fastest flight costs two hundred and twenty pounds and takes two hours thirty. Would you like to filter by price?

Participant: Yes, please.

David: Ok, how would you like to filter?

Participant: Lowest to highest, or between... Well, between two... Sorry can you repeat the cheapest price again, please?

David: It is one hundred and seventy-five pounds.

Participant: Could you filter between two hundred and three hundred pounds?

David: Ok, one moment please. There are twenty-three flights available. The cheapest flight costs two hundred and ten pounds and takes three hours forty-five and the fastest flight costs two hundred twenty pounds and takes two hours thirty. Would you like to filter by duration?

Participant: Emm, no.

David: Ok, would you like to filter by departure time?

Participant: Yes, please.

David: Ok, how would you like to filter?

Participant: Emm, departing before three pm. [. . .]

C.5 Active-Recommendation Agent (AR)

Euan: Hello, my name is Euan. I am flight search agent and I will help you to find your flight. To get started please provide me with your destination and day of departure.

Participant: I want to go to Reykjavik and I am flying on the sixth of April.

Euan: Sure, one moment. Okay, so we have twenty seven flights on that day. Prices range from one hundred ninety to three hundred pounds and flight durations range from two and a half hours to five and a half hours. Would you like to filter that by price?

Participant: Can we filter by time? Please get rid of anything that arrives after three thirty pm?

Euan: We have six flights in that range. rices range from one hundred ninety to three hundred pounds and flight durations range from two and a half hours to five and a half hours. Would you like to filter that by price?

Participant: Yes, how long does the cheapest flight take?

Euan: That flight takes five hours fifteen minutes and it costs one hundred ninety pounds. However, if you spend thirty pounds more you can get there half an hour quicker.

Participant: OK, could you do the same search for the next day? [...]

C.6 ProActive-Listing Agent (ProL)

Frank: Hello, my name is Frank. I am flight search agent and I will help you to find your flight. To get started please provide me with your destination and day of departure.

Participant: I would like to book a flight for Las Palmas in the Canary Islands, Spain. I am hoping to fly on Wednesday, the eighth of July.

Frank: Sure, are you more concerned about the cost of the flight or about the duration of travel?

Participant: The cost of the flight

Frank: Ok, do you have any requirements regarding departure or arrival time?

Participant: I would like to arrive before six pm.

Frank: Ok, one moment please. The two cheapest flights that arrive before six pm are as follows. The first one is a KLM flight, departure six fifty am, arrival two zero five pm, it costs three hundred seventy pounds and takes seven hours and fifteen minutes. The second option is a BA flight, departure time ten am and arrival at four forty-five pm, it costs three hundred

seventy pounds and takes six hours forty-five minutes. Would you like to select any of these flights?

Participant: Yes, I would like to go for the first option.

C.7 ProActive-Recommendation Agent (ProR)

Graham: Hello, my name is Graham. I am flight search agent and I will help you to find your flight. To get started please provide me with your destination and day of departure.

Participant: Hi, I am looking to book a flight to Reykjavik, Iceland on Friday the sixth of April, please.

Graham: Sure, are you more concerned about the cost of the flight or about the duration of travel?

Participant: Duration of travel.

Graham: Ok, do you have any requirements regarding departure or arrival time?

Participant: It would be ideal to arrive before four pm, please.

Graham: Okay, so the quickest flight on that day that arrives before four pm, is Air France departure time eight fifteen am, arrival eleven fifteen am it costs two hundred and sixty five pounds and it takes three hours. However, if you are willing to spend forty-five minutes longer travelling, you can save fifteen pounds. Would you be interested in any of these flights?

Appendix D

Additional Material

D.1 Kayak Travel Skill for Amazon Echo.

Alexa: Can you repeat the destination?

User: Chicago.

Angus: Ok, so you are going to Chicago, are you looking for a flight, hotel or car?

User: A flight.

Alexa: Got it, where do you want to fly out from?

User: Edinburgh

Alexa: Got it, when are do you want to fly out?

User: Tomorrow.

Alexa: Ok, when do you want to fly back to Edinburgh?

User: It is a one-way flight

Alexa: Ok, looking for a one-way flight from Edinburgh, UK to Chicago, Illinois, United States. Non-stop flight start at 1498 pounds. Flights with one stop start at 1139 pounds. The least expensive flight is one stop flight with Delta for 1191 pounds. Would you like to do another search?

D.2 Guidance on user studies during Covid-19 Pandemic

Staff

University Ethics Committee Covid-19 Guidance

Guidance issued 17 March 2020

- **Where there is absolutely NO face-to-face contact between researchers and participants** there are unlikely to be new ethical issues arising as a result of Covid-19. Research should be able to proceed. If there are issues relating to Covid-19 the lead researcher must contact the Ethics Committee that approved the research (School/Department or University) to discuss any the additional risks or changes to the research protocol already approved.

- **Where there is face-to-face contact between researchers and participants**:

  - studies that have not yet begun should be postponed.

  - research that has already started which involves groups of people (whether vulnerable at-risk participants or not) must be put on hold.

  - research that has already started with individuals rather than groups of people must be put on hold. However, where possible research could move to online alternatives (including *inter alia* Skype/Zoom, online questionnaires, postal questionnaires).

    Online interview methods have different ethical concerns compared to face-to-face in, for example, monitoring participants' state during the interview and how data are captured. The researcher must contact the Ethics Committee that approved the research (School/Department or University) to ask for a written change in protocol from a face-to-face to online engagement.

    The University Ethics Committee will consider any protocol change through its normal fast track process of convenors action.

  - guidance must be sought from the University Ethics Committee (which will be guided as required by the University Safety Team) before restarting research.

  - participants who have already been recruited should be informed that the research has been postponed and that they will be contacted again about a resumption of activity.

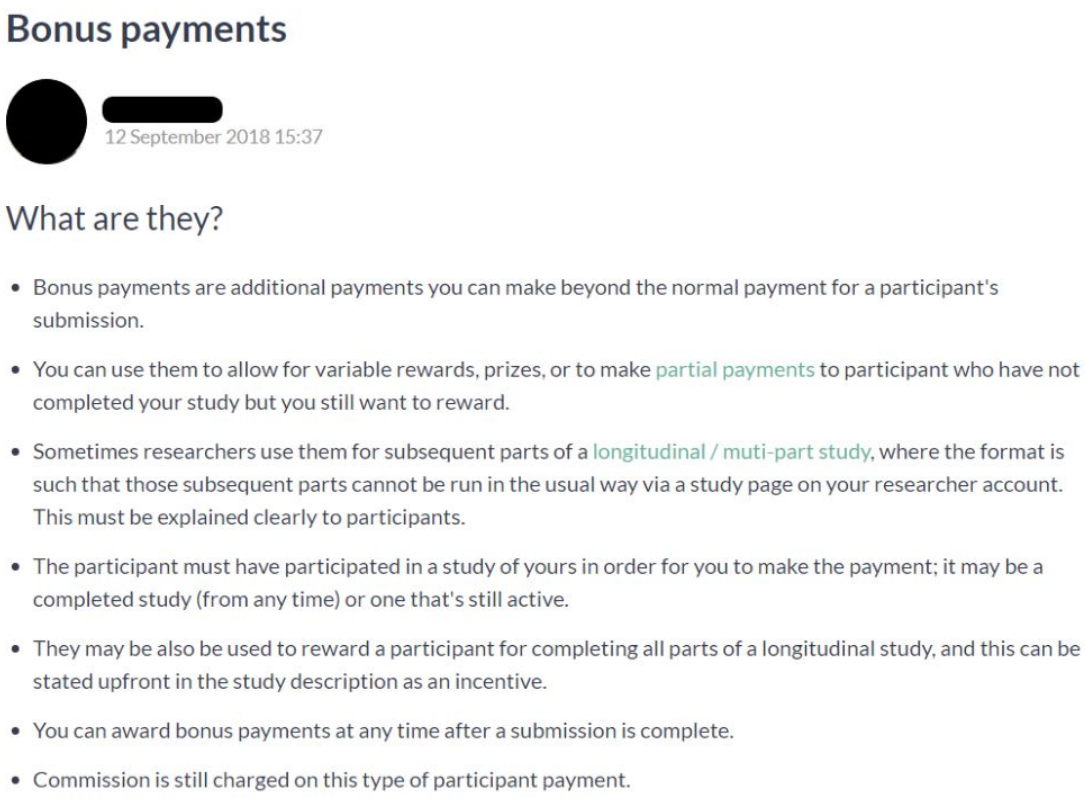
  - if the research is sponsored by external funders they should be notified of any likely delay to completion.

For further information please contact:

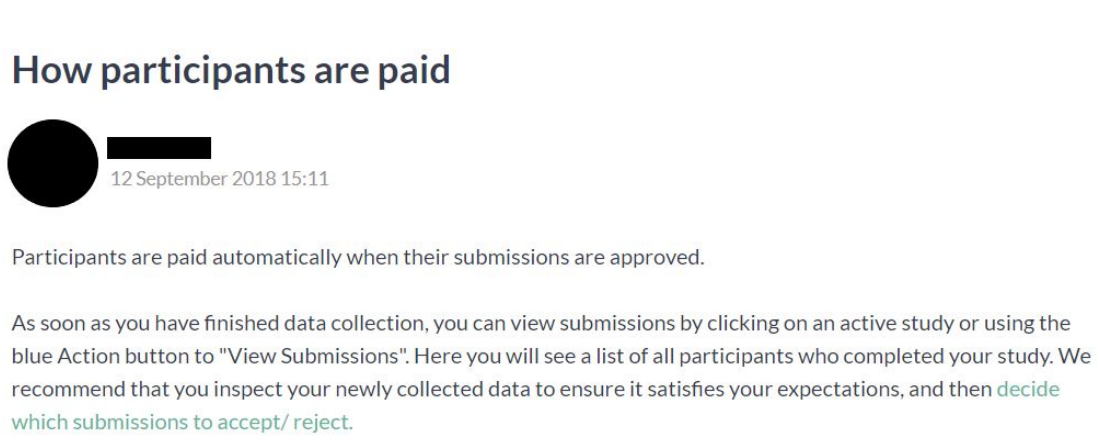
Professor Philip Winn
Convener of the University Ethics Committee
philip.winn@strath.ac.uk

D.3 Prolific Payment Estimation

D.4 Prolific Bonus Payment Information



D.5 Prolific Accepting/Rejecting Submissions



How participants are paid

[redacted]

12 September 2018 15:11

Participants are paid automatically when their submissions are approved.

As soon as you have finished data collection, you can view submissions by clicking on an active study or using the blue Action button to "View Submissions". Here you will see a list of all participants who completed your study. We recommend that you inspect your newly collected data to ensure it satisfies your expectations, and then decide which submissions to accept/ reject.

Appendix E

Search Scenarios

E.1 Search Scenarios - Study 1

You are planning to visit your friend who lives in **Bristol**. You will be flying from Glasgow Airport. Your total budget is **100 pounds**. You can travel either on the 11th or 12th of November.
Indicative Request: You want to find the cheapest possible deal but your flight needs to leave on, or before 11am.
Note: Please wait for the agent to finish before you start to speak.

You will be attending a conference in **Manchester**. You will be flying from Glasgow Airport. Your total budget is **120 pounds**. You can travel either on the 11th or 12th of November.
Indicative Request: You want to find the cheapest possible deal but your flight needs to leave on, or before 3pm.
Note: Please wait for the agent to finish before you start to speak.

You are going for a weekend break in **Liverpool**. You will be flying from Glasgow Airport. Your total budget is **120 pounds**. You can travel either on the 4th or 5th of November.
Indicative Request: You want to find the cheapest possible deal but your flight needs to leave on, or before noon.
Note: Please wait for the agent to finish before you start to speak.

A friend from **Cardiff** invited you for a wedding. You will be flying from Glasgow Airport. Your total budget is **100 pounds**. You can travel either on the 2nd or 3rd of December.
Indicative Request: You want to find the cheapest possible deal but your flight needs to leave on, or before 11am.
Note: Please wait for the agent to finish before you start to speak.

E.2 Search Scenarios - Study 2

You are planning to visit your friends who live in **Rome**. You want to get there over a weekend (on either **Friday the 14th, Friday the 21st, or Friday the 28th of September**. Since you only will be staying there for a few days, you do not want to waste too much time travelling but also want save money, so you have more money to spend during your holidays. The traffic in Rome tends to be heavy from 2pm onward, so you would prefer to get a flight that arrives around 10am.
Indicative Request: Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
Note: Please wait for the agent to finish before you start to speak.

You will be attending a student conference in **Stockholm**. You will be travelling there on either **Monday the 5th, Tuesday the 6th, or Wednesday the 7th of November**. Your university advised you that you will be allocated money from your conference fund that you will use to fund other events till the end of your academic course. To be able to attend more events in the future, you want to save money while not spending too long getting there. The student dorms where you will be staying charge extra for late check-in, so you will be aiming to arrive at around 7pm to be able to check in to your accommodation on time.
Indicative Request: Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
Note: Please wait for the agent to finish before you start to speak.

You will be flying to **London** for an interview for an internship with a local start-up company. You will be travelling there on either **Wednesday the 12th, Thursday the 13th, or Friday the 14th of December**. Since the company will not reimburse you for your travel, you need to find a cheap flight. You will only be spending one night in London so you also want to get there quick. Your interview will be after lunch so you would like to arrive around 10am to get there on time.
Indicative Request: Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
Note: Please wait for the agent to finish before you start to speak.

You are flying to **New York** to meet your cousin. You will be travelling there on either **Sunday the 16th, Sunday the 23rd, or Sunday the 30th of September**. Although your parents have decided to treat you and pay for your flight, you still want to avoid overspending. Find a flight that will get you to New York quick but without spending too much money. Your cousin offered to pick you up from the airport but he starts his afternoon work shift at 3pm so you would prefer your flight to arrive in New York around noon.
Indicative Request: Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
Note: Please wait for the agent to finish before you start to speak.

E.3 Search Scenarios - Study 3

You will be flying to **Las Palmas in Canary Islands, Spain,** to join your friends who are currently spending their holidays there. You will be travelling there on either **Wednesday the 8th , Thursday the 9th or Friday 10th of July.** Since you will only be staying in Las Palmas for a few days, you do not want to spend too much time travelling. You also want to save money so that you can spend more during your holidays. Your friends usually go to watch music concerts in the evening. You would prefer to get a flight that arrives to Las Palmas before 6pm so you can check into your accommodation and then join your friends.
Indicative Request: Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
Note: Please wait for the agent to finish before you start to speak.

You will be attending a student conference in **Lisbon, Portugal**. You will be travelling there on either **Monday the 12th, Tuesday the 13th, or Wednesday the 14th of October**. Your university advised you that you will be allocated funds from your conference fund. To be able to attend more events in the future, you want to save money while not spending too long getting there. Since the student dorms where you will be staying do not offer late check in you would prefer arrive before 8pm to be able to check in to your accommodation on time and avoid extra charges.
Indicative Request: Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
Note: Please wait for the agent to finish before you start to speak.

You will be flying to **Reykjavik, Iceland** to visit your cousin. You will be travelling there on either **Friday the 6th, Saturday the 7th, or Sunday the 8th of April**. Since you were recently made redundant, you need to find a cheap flight. However, you will be only spending three nights in Reykjavik so you also want to get there quickly. Your cousin told you that she can pick you up from the airport if you arrive before 4pm, as she starts her shift in the café afterwards. You would prefer not to miss the opportunity for getting a lift.
Indicative Request: Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
Note: Please wait for the agent to finish before you start to speak.

Your aunt who lives in **Madeira, Portugal** invited you to visit her for a holiday break. You will be travelling there on either **Sunday the 2nd, Monday the 3rd, or Wednesday the 4th of August**. Although your parents have decided to treat you and pay for your flight you still want to avoid overspending. Find a flight that will get you to Madeira quick without overspending. Your aunt offered to take you to a local cuisine festival, you would prefer to get a flight that arrives before 2pm to make the most out of your trip.
Indicative Request: Explore available flights to find a flight that offers a good balance between price and travel time (a cheap flight with short travel time)
Note: Please wait for the agent to finish before you start to speak.