Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Master thesis

Harmony in Melody: A Mediator Concept for Song Recommendation

Zorica Kačarević

Studiengang:	M.Sc. Computational Linguistics
Prüfer*innen:	Prof. Dr. Thang Vu
	Dr. Antje Schweitzer
Betreuer:	Lindsey Vanderlyn
Beginn der Arbeit:	10.08.2022
Ende der Arbeit:	30.06.2023

Erklärung (Statement of Authorship)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein. [1]

(Zorica Kačarević)

[1] Non-binding translation for convenience: This thesis is the result of my own independent work, and any material from work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completely nor partially been published before. The submitted electronic version is identical to this print version.

Abstract

This thesis discusses the adaptation of dialogue systems based on user personality. With the increasing integration of computers into people's private and professional lives the importance of effectively navigating computer systems seems to be a valid concern. As not all user interfaces can be used intuitively a bridge to close the gap between the user's expectations and the system's capabilities seems necessary. Therefore, this thesis offers the conceptualization of an adaptive dialogue system that acts as a bridge between a system and a user. For this thesis, the online music streaming platform Spotify will be used as the system's domain in question. Further, the proposed system will adapt to a user's personality. Therefore, personality traits are discussed and two traits will be chosen as axes for adaptation. Based on these, four policies will be designed to match each combination possible. To test whether adaptation based on personality makes a difference to the user, a Wizard of Oz study will be conducted. Therefore, users will be asked to create a playlist using the dialogue system. The study will then be evaluated and discussed. Interactions with the dialogue system will then be annotated to create a database as a result of this thesis.

Contents

List of Figures

List of Tables

1 Introduction

I would like to start by asking a question: *What if people were all alike?* Unfortunately, there is no definitive answer I can provide. In my imagination though, there would not be much variance, individuality, or change as people would probably enjoy the same things like food, music, and clothes. Nevertheless, in my opinion, the small differences in personality, personal likes and dislikes as well as varying experiences are what makes life interesting.

While it seems obvious to state that the personalities of children and adults differ, it should be noted that individuals' personalities may vary just as much within one of these groups (cf. Shiner and Caspi, 2003). Further, people's personalities may vary across cultures or change over time (cf. Costa Jr et al., 2001; McCrae and Costa, 2003; Cobb-Clark and Schurer, 2012). Consequentially, a person's taste, for example in music, can change over time. As of today many songs have been published and new songs are being added regularly. Those songs can be classified into varying genres and subgenres. Online platforms like `Youtube` or `Spotify` offer services to listen to music, search for new songs based on artists or genres, or create playlists. Nevertheless, to this day there is no natural language interface included in these platforms. Rather, people need to search for artists or songs using a keyword-like search. Although effective, it would be more practical to have the ability to talk to such platforms directly. For example, if a song's melody is remembered but the lyrics are not, would it not be helpful to be able to hum the melody to find a song to which the lyrics cannot be remembered? Another point to consider would be if a natural language interface is included, what else could such a system offer? Could it adapt to a user based on their preferences or personality?

To answer these questions, this thesis aims to conceptualize an adaptive dialogue system that will mediate between the online music platform Spotify and a user. Another main focus of this thesis will be to adapt the system based on the user's personality. Therefore, personality traits that can serve as axes for adaptation need to be picked first and then combined to create varying policies. Further, each policy should aim to take different approaches to song recommendation in order to further enhance the user's diverse personality. Lastly, a user study will be conducted to test the developed policies and create a database of manually annotated dialogues based on these interactions.

1.1 The Need for Adaptation in Dialogue Systems

As computers have become more involved in people's everyday lives, being able to properly use a system has become an important and expected skill. Nevertheless, not every interface provided to users is intuitive and easy to use (Kiani et al., 2019). Further, varying levels of expertise, experiences with similar systems as well as a user's willingness to engage with a new system can influence whether such interactions end successfully or not. Therefore, a bridge needs to be created – between a user's expectations of how the system should work and the system itself as well as its limitations and options. To do so, a conversational mediator, i.e. an adaptive dialogue system, is proposed.

1.2 Research Questions

Within the scope of this thesis, I follow the hypothesis that an individual's personality can arguably influence their life in varying ways. For example, being extroverted opens many possibilities to meet new friends or have interesting discussions with strangers while being introverted on the other hand would likely lead to a more distant behaviour towards strangers. Therefore, having a more natural way to communicate with a dialogue system based on one's personality, should make a difference in how users interact with the system as well as influence its perceived likability. More precisely, throughout this thesis I aim to answer the following questions:

- Does a user like a dialogue system more, if it's designed to allow the user to be 'more themselves', i.e. for a very open, very proactive user: being more proactive and relying on recommendations rather than having to answer multiple questions first? Put differently, is a dialogue system more likeable based on the approach it takes?

- Is there a difference between liking the interaction with the dialogue system and liking the playlist?
- Are there any perceived differences in the system's usability based on the approach it takes?

To answer these questions, I will address how an adaptive dialogue system should look to ensure a natural dialogue for its users. Therefore, I investigate which criteria could influence and trigger adaptation, what axes to adapt, and whether adaptation can be used to mediate effectively between users and the system.

1.3 Structure of this Work

To answer the proposed questions mentioned in section 1.2, I will first establish a foundation by discussing dialogue systems and their architecture as well as briefly examine dialogue systems as conversational mediators in section 2.1. Following this, section 2.2 aims to provide insight into how dialogue systems can adapt to users. Section 2.3 will then explore different approaches to recommender systems. This is followed by a discussion on personality and traits in section 2.4.

The main focus of this thesis will be a user study. Therefore, section 3 is dedicated to the conceptualization of an adaptive dialogue system and in particular, the design of its policies. Further, a pilot study conducted to improve these policies is discussed. Section 4 aims to describe the design and conduct of the user study. Following this, section 5 will provide information on participants while section 6 focuses on evaluating the participant's responses to the opposing policies used. The results will then be discussed in section 7. Finally, section 8 concludes this thesis.

2 Background and Related Work

In this chapter, I discuss the background and work related to dialogue systems, as well as adaptation based on user preferences and personality. Therefore, section 2.1

discusses the standard architecture of task-based dialogue systems first. Following this, section 2.2 will briefly discuss how dialogue systems can adapt to user preferences. Section 2.3 will then focus on different approaches used for recommender systems. Finally, section 2.4 will offer an introduction to the field of personality and personality traits as well as discuss some traits in more detail.

2.1 Dialogue Systems

Dialogue Systems or *Conversation Agents* like *Amazon Alexa* or *Siri* have proven to be useful tools for people to use. For example, people might as Siri about the weather or order a book using Amazon Alexa. However, before speech recognition was integrated into dialogue systems, text-based systems like Weizenbaum's *ELIZA* were considered state-of-the-art. As one of the first dialogue systems developed, ELIZA used pattern matching to converse with humans. What started with pattern matching has now developed into a huge field of research. Newer approaches incorporate Reinforcement Learning or Deep Learning techniques to improve the conversational strategies (cf. Jurafsky and Martin, 2009; Li et al., 2020; Chen et al., 2017). For example, a more recently developed dialogue system, *ChatGPT* uses artificial intelligence to learn how to generate text more effectively in human-computer interactions (cf. Lund and Wang, 2023). Consequently, not just ChatGPT, but Siri, Amazon Alexa, and ELIZA can all be described as "[...] a program which makes natural language conversation with a computer possible" (Weizenbaum, 1966; p. 36).

In general, dialogue systems can be divided into open-ended (or chitchat) and task-based systems (cf. Chen et al., 2017; Yan et al., 2017; Li et al., 2020; Su et al., 2016). While open-ended dialogue systems could be efficiently described as chatbots without the need to fulfill a certain task, task-based systems on the other hand are implemented to follow a certain goal (cf. Chen et al., 2017). For example, finding a restaurant or booking a flight are tasks that could be completed using a task-based system. For this thesis, the focus will be on task-based systems. Depending on the application, there are varying levels of complexity dialogue systems can have. However, because dialogue systems are used to facilitate some form of human-computer communication, all systems will need to accept some kind of user input like text or speech, process it with respect to the given task and generate an output (cf. Jurafsky and Martin, 2009). To understand how conversing works, this chapter will focus on providing the standard architecture of a task-based Dialogue System. Therefore, *Natural Language Understanding (NLU)*, *Belief State Tracker (BST)* and *User State Tracker (UST)* as parts of the *Dialogue Manager* or *Policy*, and *Natural Language Generation (NLG)* will be discussed. It should be noted that recent studies lean towards Reinforcement Learning or Neural Network approaches (cf. Wen et al., 2016; Li et al., 2020). Nevertheless, this thesis will be focused on a more general structure rather than discuss implementation approaches.

2.1.1 Natural Language Understanding

The *Natural Language Understanding (NLU)* module is responsible for translating a user utterance into `User Acts` the system is able to understand. User Acts can be described as a user's intent. If the user input includes speech an additional *Automatic Speech Recognition (ASR)* module is needed to translate speech into text first. Then the user utterance needs to be analyzed and relevant information needs to be extracted. To map the utterance, regular expressions could be used to match keywords of a user utterance to its corresponding user act. For example, the sentence *Can you recommend a song* could be parsed as user act `Request()` with their respective slot and value being `song=random`. Mapping the sentence into a machine-readable entity is part of the NLU's task. A sentence is analyzed, the intent is identified and together with other needed information, it is stored for further processing (cf. Chen et al., 2017; Jurafsky and Martin, 2009; Li et al., 2020).

2.1.2 Belief State Tracker

The `Belief State Tracker (BST)` manages the input and stores information of previous user utterances provided by the NLU. In other words, the BST keeps track of the dialogue, updates the belief state accordingly, and, optionally, the certainty of correctly interpreting the user's actions (cf. Chen et al., 2017; Li et al., 2020). Although it is not necessary for determining the next system action, some systems' BSTs keep a history of previous system actions for disambiguation purposes (cf. Li et al., 2020).

2.1.3 Dialogue Manager/Policy

The *Dialogue Manager* or *Policy* then decides on the next `System Action` based on the BST's certainty and the current user act. A system action can be described as the system's intent to provide the user with the best answer possible. For example, following the NLU example above, an appropriate system action could be `Recommend_Song(song=song)`.

One possible approach to creating a dialogue system's policy would be a handcrafted one where the next system act is decided based on a set of handcrafted rules determining which system act is to occur in which scenario. These rules could include looking into the belief state to determine whether a sufficient amount of information has already been provided or whether more information needs to be requested from the user. Another approach would be to utilize reinforcement learning and teach the policy how to react instead of using rules (Li et al., 2020; Chen et al., 2017).

2.1.4 Natural Language Generation

The *Natural Language Generation (NLG)* module is responsible to provide a natural output to the user based on the system act. The NLG can be, in the simplest approach, template based. Meaning that each system act leads to a predefined sentence within the NLG template. For example, the system act `Recommend_Song` would be mapped to an NLG template including *I would like to recommend this song to you:*. This natural sentence is then presented to the user to ensure more natural communication. If speech output is needed, the sentence then would be given to a *text-to-speech synthesizer (TTS)* to generate a spoken utterance.

2.1.5 Dialogue Systems as Conversational Mediators

The field of human-computer interaction (HCI) strives to improve the usability of a system. Most approaches tend to focus on improving the (graphical) user interface. But with the rising complexity of such systems and therefore the need for instruction for use, an alternative approach seems to be inevitable to minimize the user's frustration trying to learn how to use a system. Research on *Help Systems* aims to provide the user with a cooperative setting, offering help if needed (cf. Fischer et al., 1985). Therefore, this system needs to have some sort of understanding of the user's goals and knowledge. Since this proves to be a rather complex task, some researchers propose to create an *Adaptive Help System* using AI in order to better adapt to the user and react more efficiently. Either way, for example, the user's actions need to be tracked in order to decide on what the user needs help with and provide a help message.

A more recent dialogue system, ChatGPT for example offers users a variety of possibilities. For example, I asked whether it could help me with cooking *spaghetti carbonara* and was provided with a decent recipe for me to follow. When I asked for some pop songs, it recommended several for me to choose from. But when I prompted it with potentially controversial inputs like *I am proud to be white* versus *I am proud to be black* it became evident that ChatGPT seems to be rather biased in its answer. While I was reminded to be mindful of other races and cultures for being *proud to be white*, I was on the contrary praised for being *proud to be black*.

Generally speaking, if used properly, systems like ChatGPT can be useful Help Systems when searching for something specific like a recipe, but information or statements on other topics should be looked at with a grain of salt. Nevertheless, if considered as a mediator between users and the world wide web, their value becomes evident. For example, when searching for a specific recipe while dietary restrictions need to be met can be rather exhausting. In such cases, dialogue systems can be a great source of help filtering large amounts of information one encounters online while providing the user with a small but useful portion of it.

2.2 Adaptive Dialogue Systems

Adaptive dialogue systems are in essence dialogue systems that adapt their responses and behavior based on user inputs. Adaptation is achieved utilizing Reinforcement Learning or Deep Learning approaches, where a system is trained to adapt its responses and behaviour based on user feedback and in the case of a Reinforcement Learning approach rewarded if successful (cf. Chen et al., 2017; Litman and Pan, 2002). Generally speaking, the dialogue system's policy is adapted to better suit a user's needs as the goal of adaptive dialogue systems is to provide more engaging and personalized interactions.

Adaptive dialogue systems aim to understand user intents, generate appropriate responses, and dynamically adapt to different dialogue contexts and user preferences. For example, in the context of a technical support system, one crucial point of adaptation would be to generate responses based on the user's knowledge level (cf. Janarthanam and Lemon, 2014). Gathering information such as a user's knowledge level as well as their preferences and behaviour is referred to as `User Modeling` (cf. Janarthanam and Lemon, 2014; Litman and Pan, 2002). For example, when I asked a state-of-the-art language model, i.e. ChatGPT[2] *How do you adapt your way of communication?* the following response was given to me:

> As an AI language model, I don't have the ability to adapt my way of communication in the same way that a human would. However, I can adjust the style and content of my responses based on the input I receive and the feedback I'm given.
>
> For example, if a user provides me with a lot of technical terms or jargon, I can adjust my responses to use similar language and terminology. Alternatively, if a user prefers simpler language, I can adjust my responses to use more basic vocabulary and sentence structures.
>
> Additionally, if a user provides feedback indicating that they prefer a

[2]Personal communication, April, 2023

> certain tone or style of communication, such as a more formal or conversational tone, I can adjust my responses to match that preference.
>
> Overall, while I may not be able to adapt my way of communication in the same way that a human would, I can still adjust my responses to better suit the needs and preferences of the user.

In other words, ChatGPT's level of adaptation is rather focused on adjusting to a user's knowledge base and preferred way of communicating, which can be linked to adjustments in policy and NLG respectively. However, it remains challenging to reach a human-like level, as some preferences cannot be described but need to be felt. For example, the quality of an actor may influence how we feel about a movie. If they perform their role in a way that touches one's emotion, that movie is more likely to be enjoyed even though it might not usually be the genre of choice. Still, that same actor might not evoke the same emotions in another movie even though their quality of acting remains the same, rather it would be the unique combination of skill and topic that would be hard to replicate in a different movie. Nevertheless, adaptive dialogue systems represent a significant advancement in adjusting to a user's needs and responding in a more personalized manner.

2.3 Recommender Systems

Recommender systems are, generally speaking, algorithms designed to provide personalized recommendations to users, helping them discover relevant items or content based on their preferences and past behavior. These systems have become an integral part of various online platforms, such as e-commerce websites, streaming services, social media platforms, and more (cf. Ricci et al., 2015; Leskovec et al., 2020). For example, when I watch a show on the streaming platform `Netflix`, different shows with similar content are recommended to me to watch next. "Recommendation systems suggest items of interest and enjoyment to people based on their preferences" (Bennett et al., 2007).

In order to understand how recommender systems work, it is crucial to know that a vast amount of data is required (Bennett et al., 2007). Therefore, one goal of recommender systems is to overcome information overload and assist users in navigating through available options (Ricci et al., 2015; p. 2). By analyzing user data, such as for example preferences on music or movies, previous purchases, ratings of songs and movies, and demographic information, recommender systems can learn to understand individual preferences and generate recommendations based on them. This information is stored in a `User Model` (Ricci et al., 2015; Billsus and Pazzani, 1997; Fischer, 2001) and accessed to customize recommendations for users.

There are several types of recommender systems, including collaborative filtering, content-based filtering, and hybrid approaches. Although other approaches exist such as Knowledge-based or Demographic (Burke, 2007), for the purpose of this thesis only the three initially mentioned will briefly be discussed.

Collaborative filtering techniques use the information provided by many users and their preferences in order to make recommendations to a single user. Collaborative filtering is a pattern-based approach, meaning that similarities between users in regards to previously liked or purchased items yield similar recommendations (cf. Ricci et al., 2015; Burke, 2007).

Content-based filtering, on the other hand, is based on information previously provided by the user in regard to an item. It analyzes the content of items, such as textual descriptions, genres, or tags, and matches them with user preferences. For example, if a user has indicated to like action movies by giving them a high rating, content-based recommender systems will then recommend similar movies within the same spectrum of preferences the user has shown (cf. Ricci et al., 2015; Burke, 2007).

Hybrid recommender systems combine different approaches in order to maximize their benefits. When combining collaborative and content-based filtering the aim is to overcome the limitations of each method individually and offer more accurate and diverse recommendations. Therefore, instead of exclusively relying on data provided by many users or only moving within the scope of one user, offering both options to a user to choose from seem to be a viable option for recommendations (cf. Ricci et al., 2015). Nevertheless, combining two or more models does not exclude the same types of models to be combined. For example, two different content-based models could be combined as well (cf. Burke, 2007).

How successful a recommender system is at providing suggestions to the user is heavily reliant on the quality and amount of data available. Without feedback from one or more users, none of the systems briefly discussed above would be able to achieve the goal of providing useful recommendations. Therefore, data and information need to be collected from users which comes with ethical considerations that need to be taken into account. Nevertheless, this will not be further discussed but needed to be mentioned as user data should also be handled with utmost caution and care.

In conclusion, recommender systems play an important role in enhancing a user's experience. They can help reduce information overload, for example, if a user would have to go through millions of potential songs to listen to compared to being offered a small amount tailored to a user's preference or help users discover new items they potentially like.

2.4 Personality and Traits

Although the term *Personality* is colloquially used to describe a person's behaviour, character, and way of reacting to and engaging with varying circumstances, there seems to be no consensual definition of personality amongst psychologists (cf. Feist and Feist, 2006; Roberts and Mroczek, 2008). For example, I would describe myself as outgoing, nice and extroverted with friends but shy, suspicious and introverted around strangers. This inconsistency in behaviour could be explained by past experiences, upbringing or other external influences. Therefore, personality seems to be multifaceted. In broad terms, personality seems to be a unique combination of these elements combined with social and environmental influences that define an individual's distinctive pattern of thinking, feeling, and acting and what makes each person special in their own way (Feist and Feist, 2006; p. 4).

However, although there is no agreement on the definition of *Personality* amongst personality theorists Feist and Feist (2006), following Roberts and Mroczek (2008), argue that an individual's *Personality* can be described in terms of *Traits* combined with *Characteristics*. Traits such as *openness to experience* can change in time. Although considered mostly stable in adulthood, recent studies have shown that traits can continue to change at any stage of life (McCrae and Costa, 2003; Cobb-Clark and Schurer, 2012; cf.). Characteristics on the other hand are being described as "unique qualities of an individual that include attributes such as temperament, physique, and intelligence" (Feist and Feist, 2006; p. 4)[3].

In other words, an individual's *Personality* can be described using patterns of *Traits*. Trait theories of personality emphasize the measurement and categorization of specific traits to understand and predict human behavior. These theories aim to identify the fundamental building blocks of personality by breaking them down into distinct dimensions. Traits are typically assessed through self-report questionnaires or behavioral observations, and they can vary in intensity from person to person. One prominent framework to study and understand personality is the *Five-Factor Model (FFM)*, also known as the *Big Five* (cf. Rammstedt et al., 2010; Almlund et al., 2011; Johnson, 2014; McCrae and Costa, 2003; Costa Jr and McCrae, 2006)[4].

2.4.1 The Five-Factor Model (Big Five)

The Five-Factor Model (FFM) categorizes personality into five broad dimensions: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. Each dimension represents a spectrum, with individuals falling somewhere along the continuum for each trait (cf. Johnson, 2014; McCrae and Costa, 2003; Costa Jr and McCrae, 2006). Each dimension will be explained briefly in the following paragraphs.

[3] As this thesis focuses on adaptation through personality traits rather than characteristics, the latter will not be discussed further.

[4] It should be noted that, although not further discussed, other models and questionnaires to model personality exist, such as for example the Myer-Briggs Type Indicator.

Openness (to Experience) refers to an individual's inclination towards new experiences, attentiveness to inner feelings, and (intellectual) curiosity. Those high in openness tend to be imaginative, creative, and receptive to novel ideas and perspectives. On the other hand, individuals low in openness may prefer routine, tradition, and familiarity (cf. Brislin and Lo, 2006; Rothmann and Coetzer, 2003; McCrae and Costa, 2003; Almlund et al., 2011).

Conscientiousness relates to one's level of organization, responsibility, and self-discipline. Highly conscientious individuals are often reliable, efficient, and diligent in their work and personal lives. They tend to set and pursue goals with determination and strive for excellence. Conversely, those with low conscientiousness may exhibit a more spontaneous and relaxed approach to life, sometimes struggling with self-control and consistency (cf. Rothmann and Coetzer, 2003; Brislin and Lo, 2006; McCrae and Costa, 2003; Almlund et al., 2011).

Extraversion encompasses sociability, assertiveness, and enthusiasm. Extroverts thrive in social situations, gain energy from interactions with others, and enjoy being the center of attention. They tend to be outgoing, talkative, and seek external stimulation. In contrast, introverts tend to be more reserved, and introspective and prefer quieter, less stimulating environments (cf. Rothmann and Coetzer, 2003; Brislin and Lo, 2006; McCrae and Costa, 2003; Almlund et al., 2011).

Agreeableness reflects an individual's level of compassion, empathy, and cooperativeness. Those high in agreeableness are typically considerate, kind, and accommodating. They value harmonious relationships, are good listeners, and often prioritize the needs of others. People low in agreeableness may be more assertive, independent, and direct in expressing their opinions and desires (cf. Rothmann and Coetzer, 2003; Brislin and Lo, 2006; McCrae and Costa, 2003; Almlund et al., 2011).

Neuroticism refers to the degree of emotional stability and reactivity. Individuals high in neuroticism tend to experience more negative emotions such as anxiety, worry, and mood swings. They may be more prone to stress and exhibit higher emotional sensitivity. Conversely, individuals low in neuroticism are generally more emotionally resilient, calm, and stable (cf. Rothmann and Coetzer, 2003; Brislin and Lo, 2006; McCrae and Costa, 2003; Almlund et al., 2011).

It is important to note that personality traits are not fixed or unchangeable. While there is a genetic predisposition towards certain traits (cf. Rothmann and Coetzer, 2003), individuals have the capacity to develop and grow in various aspects of their personality throughout their lives (cf. McCrae and Costa, 2003; Cobb-Clark and Schurer, 2012).

Understanding personality traits, for example with the help of a framework like the Five-Factor Model, is a key ingredient towards choosing appropriate axis for an adaptive dialogue system. Nevertheless, the FFM traits should be understood as taxonomies rather than absolute traits (cf. DeYoung et al., 2007). They are constantly reevaluated and adapted. For example, the `HEXACO Personality Inventory` offers a sixth trait, namely Honesty-Humility (cf. Lee and Ashton, 2004) which is linked to a person's degree of sincerity, fairness and modesty (cf. Anglim and O'connor, 2019).

2.4.2 The Proactive Personality Scale

Another interesting personality trait to consider is *Proactivity*. Bateman and Crant (1993) developed the `Proactive Personality Scale (PPS)` mostly focusing on proactivity as a "dispositional construct that identifies differences among people in the extent to which they take action to influence their environment" (Bateman and Crant, 1993; p. 103). Put differently, people's efforts in completing a task or proposing a new one can be linked to their proactive behaviour. "To be proactive is to take the initiative [...]" (Bateman and Crant, 1999; p. 63). Although initially proposed and empirically tested in a workplace setting, proactivity applies to the person, not the job description. For example, two individuals may hold the same job at the same company but tackle it rather differently. While the proactive person would likely actively try to finish a task or find more things to do, the not-proactive one would be more passive about those tasks but still complete them (cf. Bateman and Crant, 1999).

With respect to the topic of this thesis, proactivity seems to be a valuable trait for adaptation as conversing with a dialogue system can become monotone rather quickly when answering the same questions repeatedly. Hence, proactivity can be a means to achieving a goal more efficiently.

3 Conceptualizing an Adaptive Dialogue System

Since this thesis's aim is to conceptualize an adaptive dialogue system and understand how it is perceived by users, the main focus will be on the dialogue system's policy. Prior to working on conceptualizing the dialogue system, a domain needed to be selected. It was decided to focus on song recommendation and use the already existing website *Spotify*[5] as a database for song recommendations. Further, `System Actions` as well as `User Actions` needed to be developed. A complete list can be found in A.3.

3.1 Designing the Policies

As the dialogue system aims to be adaptive to the user's personality, an appropriate axis for adaptation needed to be established first. The choice was made to use `Openness to Experience` from the Five-Factor Model described in section 2.4 as users likely will expect a song-recommending system to be able to suggest songs based on their preferences. For example, if a user specifies to like *Pop* music but without specifying a particular artist, the system would be required to offer songs of random, and to the user possibly unknown, pop artists. Therefore, whether a user is open to experiences, i.e. experiencing new and unknown songs, seems to be a crucial starting point for adaptation. The next step was to find a validated scale to measure a user's *openness*. It was decided to use the IPIP validated scale

[5] `https://open.spotify.com/search`

Adventurousness[6] based on Goldberg's representation of Costa and McCrae's (1992) Revised NEO Personality Inventory (NEO-PI-R Facets) (cf. Goldberg et al., 1999; Costa and McCrae, 1992)[7] as it provided interesting questions to user's openness to adapt to unknown songs like for example *'(I) Prefer variety to routine'* and *'(I) Am a creature of habit.* Other scales on openness or adventurousness focused for example on discovering new countries like *Adventurousness*-scale developed by Johnson (2014) and were therefore deemed to be not suitable for the purpose of this study.

As the second axis for adaptation, `Extraversion` was considered first. However, as no appropriate scale could be found for the purpose of this study and as users needed to engage with a dialogue system rather than with another human, it was decided to instead use another trait. After some discussion it was agreed upon using the `Proactivity Personality Scale (PPS)` by Bateman and Crant (1993). The full scale can be viewed in A.2. The PPS involves questions like *I enjoy facing and overcoming obstacles to my ideas* and *When I have a problem, I tackle it head-on* which would provide insight on how likely a user is to proactively take control over the situation. For example, proactive users would be expected to name songs they liked rather than waiting for the system to discover their taste in music.

As users would be asked to provide information on their personality traits (cf. 4 prior to interacting with the dialogue system, four policies needed to be created as well as accepted `System` and `User Acts` needed to be established.

As can be seen in figure 1 users will be categorized into one of four policies based on their self-reported personality traits. Users with a high value in proactivity and a high value in openness would be assigned the `HPHO` policy, while users with low values would be assigned the `LPLO` policy. These two policies were considered to be extreme cases. Respectively, users with a high proactivity score but a low openness score were assigned to the `HPLO` policy and those scoring low on proactivity but high on openness were given the `LPHO` policy. Each policy was designed to match a user's self-reported traits, meaning that the policies were designed to support their

[6]The scale can be found here: `https://ipip.ori.org/newNEOKey.htm#Adventurousness`
[7]See also: `https://ipip.ori.org/newMultipleconstructs.htm`

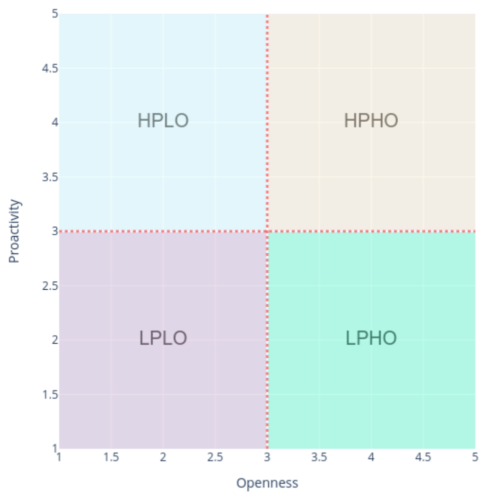


Figure 1: Assigning dialogue system policy based on personality traits – HPHO (High Proactivity, Low Openness), HPLO (High Proactivity, Low Openness), LPHO (Low Proactivity, High Openness), LPLO (Low Proactivity, Low Openness)

personality. For example, if a user reports being highly proactive and highly open, the system offers suggestions but expects the user to offer details as to why they like or dislike a song that has been recommended. On the other hand for highly proactive but not open users, the system would not offer suggestions unless prompted to do so and instead directly ask the user which songs they would like to add. In the following paragraphs differences in the four policies will be discussed in more detail.

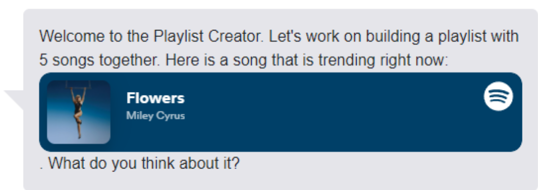


Figure 2: Welcome message for HPHO users including one song recommendation and an open question for the user to answer.

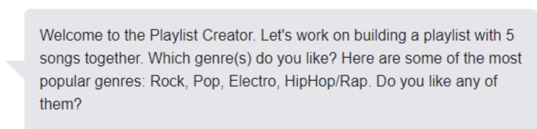


Figure 3: Welcome message for LPLO users including information on popular genres together with a question about their preferences.

HPHO policy: For the HPHO policy, users were expected to provide detailed feedback to open questions and recommendations due to their proactive personality. Further, they were given immediate recommendations based on their feedback. As can be seen in figure 2 users were welcomed by the system with a song recommendation followed by an open question *What do you think about it?*. Further, songs were directly added to the playlist if a user indicated to like a song unless the user indicated otherwise. After offering three songs from three genres (pop, rock, and metal) to establish a baseline, users were offered one-song recommendations based on their preferences. If no preferences were established, the system would proceed to ask for preferred genres or artists.

LPLO policy: For the LPLO policy, users were expected to not provide a lot of information unless prompted otherwise or asked directly. Therefore, users were given selections of two to three songs to vote on after narrowing down their preferences. Therefore, their welcome message, as can be seen in figure 3 consisted of a question about genre preferences followed by a recommendation on popular genres. After preferences were established, users received two-song recommendations based on the information provided. If no preferences could be established, random suggestions were made. Prior to adding a song to the playlist, users were asked *Would you like to add it to your playlist?* unless they indicated to directly add it to the playlist. After adding a song to the playlist, users were given the option to change the genre or artist: *Would you like to add more songs from this artist or would you like to change the artist or genre?*. Based on previous answers given, the system would either switch to the next artist or genre provided or ask the user which genre or artist they would prefer to switch to while again providing options for them to choose from.

HPLO policy: For the HPLO policy, users were expected to provide a lot of information to the system and suggest songs directly. Therefore, the welcome message, as can be seen in figure 4, included a direct question on which song to add *What song should I add first?* and no recommendations as users were expected to be not very open for suggestions. Therefore, the system would not ask general questions

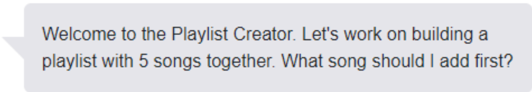


Figure 4: Welcome message for HPLO users including a direct question as to which song to add.

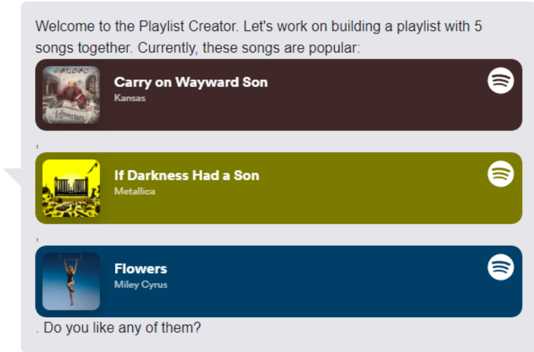


Figure 5: Welcome message for LPHO users including three song recommendations of different genres to select from.

as to what genres or artists were liked unless the user was unwilling to provide information and rejected multiple suggestions that were previously requested. Songs were again directly added to the playlist if users expressed liking them and didn't state to not include them in the playlist.

LPHO policy: For the LPHO policy, users were expected to be very open to recommendations and were therefore greeted with popular songs from three different genres (i.e. Rock, Metal, and Pop) as can be seen in figure 5. As users were not expected to offer a lot of information, they were asked whether they liked any of the songs *Do you like any of them?*. If a user liked any of the three songs, the respective genre and artist were noted as being liked, and other songs within the genre were recommended going further. This continued until a user specified to want recommendations from another genre or artist, or requested a specific song. Again, users were asked whether they wanted the song to be added to their playlist prior to doing so: *Would you like to add it to your playlist?*. If answered positively the song was added, if ignored or negated the song would not be offered again.

In general, all four policies consisted of the same `System Acts` but differ in their respective approach, i.e. offering suggestions and information seeking, as explained above. Nevertheless, some System Acts like the welcome messages will only occur in their respective policy, while others can be triggered by the user although are unlikely to appear, like the varying recommendations acts. For example, HPHO users will be given one song at a time (`Recommend_One_Song()`) as more detailed feedback is expected, while the other three policies will provide at least two songs (`Recommend_Song()`) if asked for recommendations. Nevertheless, HPHO users can trigger `Recommend_Song()` by being vague in their requests and feedback. A complete list of System Acts and their respective NLG template as well as accepted User Acts can be found in A.3.

3.2 Pilot Study

Prior to the main experiment, a pilot study with 2 participants was conducted to test the study design, and in particular, the user and system acts that I had developed so far. Both participants were given the same task, namely to test the capabilities and limitations of the dialogue system. They were asked to imagine the dialogue system would be able to understand most queries. During the study conduct both participants were aware that I was in fact acting as the dialogue system. This decision was made in order to be able to directly address issues and questions and to gain insight into how the participants would expect such a system to act and what additional functionalities would be needed. They were not informed that there were four versions based on their proactivity and openness scores. However, they were told that I wanted to test two contrasting approaches to song recommendation.

Coincidentally, both participants had high proactivity and high openness scores, and therefore, only two out of four policies could be tested effectively, i.e. only HPHO and LPLO were tested while HPLO and LPHO could not. Nevertheless, I decided to adapt the other two policies with this information as well, by fitting proactive behaviour to the HPLO policy and more open behaviour to the LPHO.

This resulted in a revision of the proposed system and dialogue acts, which can be found in A.3, as well as new insights into what could be expected by other users. Furthermore, I included the results of the dialogues in my exemplar policy dialogue and manually annotated them to further enhance `System` and `User Acts`[8].

4 WOZ Study – Design and Conduct

One main focus of the study design was to not overly strain participants and keep them engaged in the study in order to obtain the best results possible. Therefore, designing the dialogue system (cf. section 3) and study had to go hand in hand. After the four policies were established, it was decided to conduct a within-subject study and present participants with the policy that matched their personality (referred to as `Match`) as well as the opposite policy (referred to as `Opposite or Mismatch`). To avoid bias, the starting policies were alternated, i.e. if the first user started with a matching policy and ended with the mismatching one, then the second user was given the mismatching policy first and ended with their match.

Further, the decision was made to conduct a `Wizard of Oz (WOZ)` study. Wizard of Oz studies refer to a methodology used in the field of human-computer interaction and natural language processing to evaluate and refine dialogue systems prior to implementation (cf. Dahlbäck et al., 1993; Dow et al., 2005; Janarthanam and Lemon, 2014). In WOZ studies participants believe to be interacting with a system, while in reality the system's output is controlled by a human *wizard*. The wizard acts like the system following a set of predefined templates while making decisions about the system's turn within the scope of its policy (cf. Dahlbäck et al., 1993; Dow et al., 2005). Dow et al. (2005) explain that the wizard can play different roles by either representing the system completely or by monitoring an existing system's decisions. As there is currently no such system implemented, the wizards, i.e. my, task will be to fully represent the system during interactions with the participants.

Prior to conducting the study, each user was provided with survey instructions which can be viewed in A.4. In order to be as efficient as possible while conducting a WOZ study, participants were first asked to self-report on their *Openness*, *Proactiv-*

[8] Annotated example dialogue for each of the four policies can be found on Github: `https://github.tik.uni-stuttgart.de/dirkvaeth/woz_datacollection.git`

ity, *Familiarity* with similar systems and their *Propensity to Trust* dialogue systems prior to interacting with it.

Familiarity and *Propensity to Trust* are validated subscales designed by Körber (2019). They are part of a set of subscales on *Trust in Automation*[9]. Each scale can be used individually or in combination with other subscales (cf. Körber, 2019). The subscale on familiarity consists of two questions while the subscale on propensity to trust consists of three. In the latter one question needs to be inverted, then for both subscales, the mean is calculated.

As discussed prior in section 3.1, for openness I used the IPIP validated scale *Adventurousness* based on Goldberg's representation of Costa and McCrae's (1992) Revised NEO Personality Inventory (NEO-PI-R Facets) (cf. Goldberg et al., 1999; Costa and McCrae, 1992) while proactivity was measured using the `Proactivity Personality Scale (PPS)` by Bateman and Crant (1993). The openness/ adventurousness scale consists of ten questions of which six need to be inverted prior to calculating the mean score. The proactivity scale on the other hand consists of 17 statements of which only one needed to be inverted prior to calculating the mean value. Both values were then used to decide on the correct policy and the first dialogue started.

Participants were asked to interact with the dialogue system and create a five-song playlist based on their own song preferences. They were informed that the system will accept their input on `genre`, `artist(s)`, `title`, `album`, `year`, and `version`. A valid user input would therefore be for example *I want a song by Eminem* or *Can you play the studio version of this song?* but *What is the weather today?* would not be understood. They were further informed that they could request recommendations or ask to change an already picked song. It was explained that they would receive a short sample from Spotify to listen to. And they were asked to be patient as the system would require some time to react to their requests.

The general task was visible to the user at all times as part of the interface as

[9] The subscales can be accessed on Github: `https://github.com/moritzkoerber/TiA_Trust_in_Automation_Questionnaire`

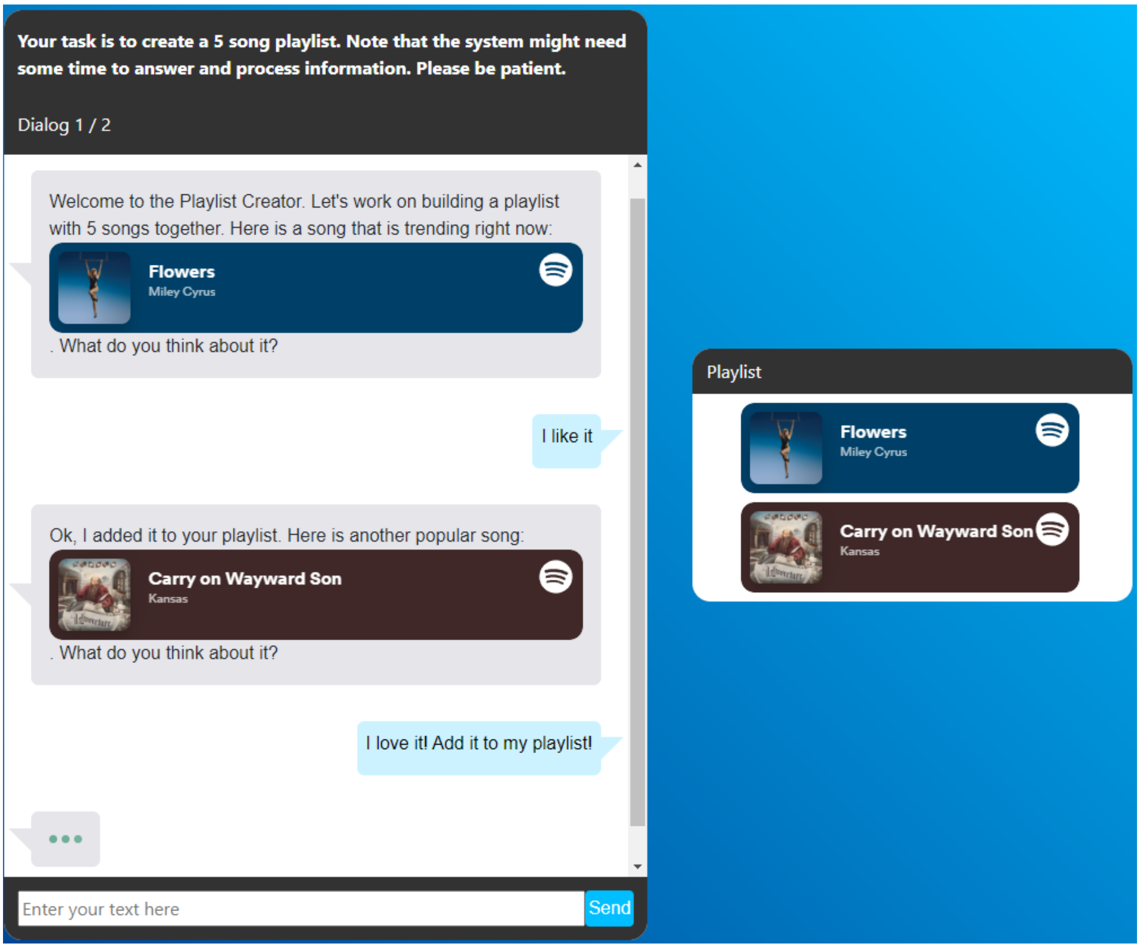


Figure 6: Interface overview for with a short example dialogue and two songs in the playlist.

well as a reminder to please be patient as the system needed some time to answer and process information. As can be seen in 6, users were shown three dots as a sign that the system was thinking and processing information.

After a 5 song playlist was created, users were asked to fill out a questionnaire based on their personal perception of the interaction. Further, they were asked to only rate their interaction with the current dialogue system. They were presented with three more subscales of Körber (2019), namely `Reliability/Competence`, `Understanding/Predictability`, and `Trust in Automation`. The latter is a small subscale consisting of two questions, Reliability/Competence consists of 6 questions of which two needed to be inverted, and Understanding/Predictability includes 4 questions of which again two needed to be inverted. For all three subscales a mean needed to be calculated afterwards.

Further, participants were asked to rate the system's usability following the `System Usability Scale (SUS)` by Brooke et al. (1996). To get appropriate SUS scores, numbers that were provided on a five-point Likert scale had to be summed up first. Then half of the items needed to be reduced by one while the other half needed to be subtracted from five, i.e. inverted. Afterwards, the sum had to be multiplied by 2.5 in order to obtain an overall value ranging between 0 and 100.

Following this users were asked three questions on a five-point Likert scale: whether they liked using the dialogue system, whether they enjoyed communicating with the dialogue system, and whether they were satisfied with the playlist. These values were then used for correlation analysis to see if there was any correlation between liking the system and liking the playlist.

In combination with the three questions, participants were asked to rate their impression of the dialogue systems' likability. Therefore they were given a scale of five points in between contrasting adjectives: *dislike - like, unfriendly - friendly, unkind - kind, unpleasant - pleasant,* and *awful - nice.* Bartneck et al. (2009) refer to this scale as `Godspeed questionnaire` and although more five-point scales belong to this, they were not relevant to this study as they were designed for measuring robots rather than dialogue systems.

Lastly, participants were given the option to include written feedback on what they liked and disliked about the interaction with each dialogue system. They were also asked what they would have changed about the system's behaviour to make the interaction better for them. Lastly, there were given the option to provide any other feedback they had or report problems with the system if there were any.

Except for the self-report questionnaire in the beginning, participants were expected and informed to interact with two dialogue systems and evaluate them separately[10].

5 Participant Data

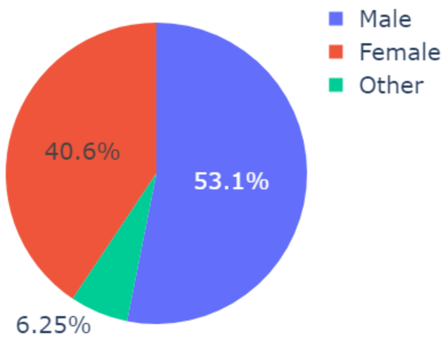
This section aims to describe participants and the information they provided in regard to themselves. In particular, section 5.1 will be devoted to demographic information. Following this, section 5.2 aims to describe how participants' personalities were assigned to the four policy models as well as provide information on distribution across policies. Lastly, section 5.3 provides information on participants' familiarity with dialogue systems as well as their propensity to trust said systems.

5.1 Demographic Information

A total of 33 participants took part in the study. One participant had to be excluded due to mistakenly being assigning the wrong versions of the dialogue system. This left me with a total of 32 participants aged between 20 and 42, with an average age of 26.6. Out of these 32 participants, 13 were female (40.6 %), 17 were male (53.1 %) and 2 (6.25 %) indicated their gender as other. A visual representation of the gender distribution can be seen in figure 7.

Prior to engaging with the dialogue systems, users were asked to fill out a questionnaire which can be seen in A.5. Information about educational background was

[10]Data collected during this study can be found on Github: `https://github.tik.uni-stuttgart.de/dirkvaeth/woz_datacollection.git`



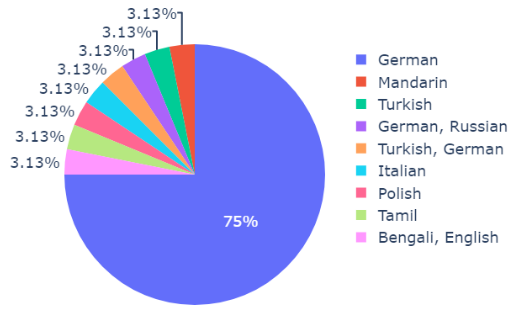


Figure 7: Distribution of participant's indicated gender.

Figure 8: Distribution of participant's native languages.

collected but was not analyzable due to highly varying in language and understanding as some users interpreted the question as being asked about their current employment. With respect to native languages, 29 participants were raised monolingual with 24 indicating German (overall 75 %) as their native language, and 5 indicating others (i.e. Mandarin, Turkish, Polish, Italian and Tamil) as their first languages. The remaining three participants indicated having two native languages (German & Russian, Turkish & German, Bengali & English). Although interesting, no further considerations could be made in regards to native languages. The distribution of native languages can nevertheless be seen in figure 8.

5.2 Modelling Personality

Before proceeding to interact with the dialogue system, participants were asked to indicate how `open` they perceive themselves to be. This was established using ten statements on a 5-point Likert scale[11]. Therefore, I decided to use the IPIP validated

[11]The 5-point Likert scale used ranged from 1 to 5 with the following mapping: 1-strongly disagree, 2-disagree, 3-neither agree nor disagree, 4-agree, 5-strongly agree.

scale *Adventurousness*[12] to get an idea of how open participants are. The scale is based on Goldberg's representation of Costa and McCrae's (1992) NEO-PI-R Facets (cf. Goldberg et al., 1999; Costa and McCrae, 1992)[13] and can be seen in A.1. All statements were slightly altered by adding an *I* at the beginning of the sentence. For example, the statement *Prefer variety to routine* was changed to *I prefer variety to routine*. This was done to highlight that participants should answer these questions in regard to their perception of themselves. The *Adventurousness* scale was chosen, because it fits best with respect to the user's task being to interact with a song-recommending dialogue system.

Further, participants were asked to indicate how `proactive` they perceive themselves to be. As for openness, a 5-point Likert scale[14] was used. As a basis for this I used the *Proactive Personality Scale* developed by Bateman and Crant (1993). The scale consists of 17 statements and can be viewed in A.2. As all statements were already written in first-person, no changes needed to be made.

For `Proactivity` and `Openness` a mean needed to be calculated for each participant. Prior to that, all values marked as *–keyed* in A.1 and A.2 needed to be inverted[15]. After calculating the mean for both traits, values ≥ 3.05 were considered `High`, while values ≤ 3.04 were considered `Low`. For example, if a participant had a proactivity score of 3.05 and an openness score of 2.5, this participant would have been seen as having high proactivity (`HP`) and low openness (`LO`) and therefore would have been assigned the `HPLO` policy as their match and consequently the LPHO policy as their mismatch. A distribution of proactivity and openness scores is shown in figure 9.

As can be seen, the figure is divided into four parts. The top left represents participants that indicated to be highly proactive, but rather less open/adventurous and were therefore labeled as `High Proactivity, Low Openness (HPLO)`. Participants in the top right part perceive themselves as being very proactive and very open and

[12]The scale can be found here: `https://ipip.ori.org/newNEOKey.htm#Adventurousness`
[13]See also: `https://ipip.ori.org/newMultipleconstructs.htm`
[14]See footnote 11
[15]Invert *–keyed* values as follows: 1=5, 2=4, 3=3, 4=2, 5=1

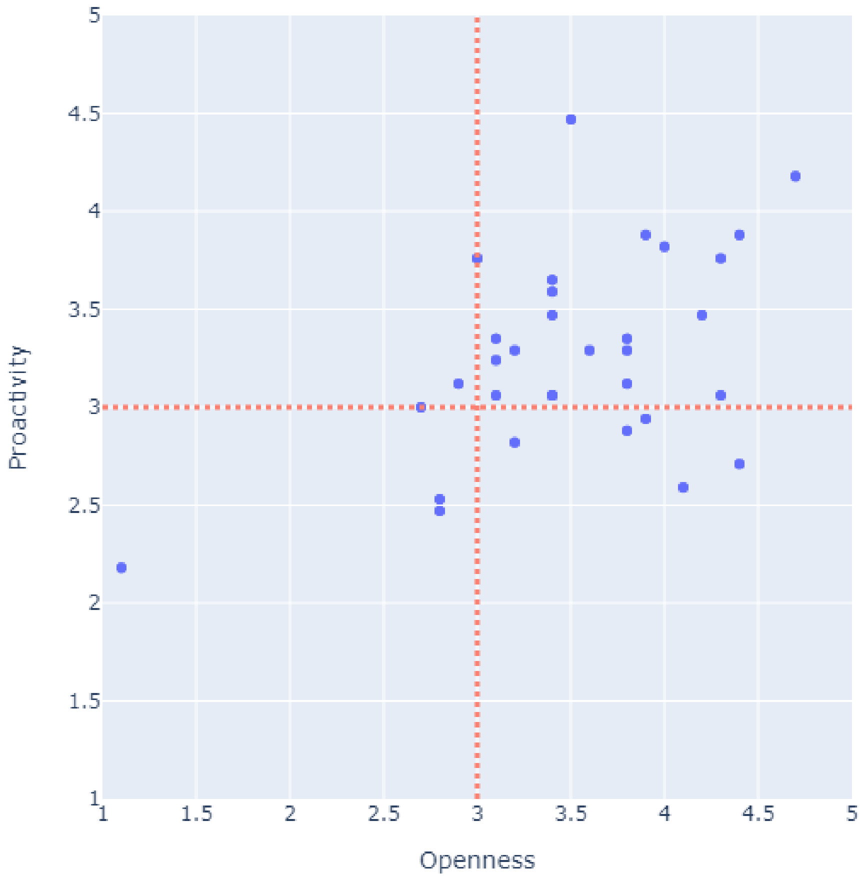


Figure 9: Distribution of user's proactivity and openness scores.

were therefore labeled as `High Proactivity, High Openness (HPHO)`. In the bottom half, participants indicated to be less proactive. Further, participants on the left side considered themselves to be less open/adventurous and were therefore labeled as `Low Proactivity, Low Openness (LPLO)`. Participants on the right who indicated to be more open were labeled as `Low Proactivity, High Openness (LPHO)` respectively.

A total of 21 participants (65.6 %) were labeled as `HPHO`, followed by 5 `LPHO` participants (15.6 %). Then 4 participants (12.5 %) were labeled as `LPLO` and lastly 2 participants were labeled as `HPLO` (6.25 %). Even though two participants cannot be representative or statistically significant, I have decided to nevertheless include them for further evaluation if applicable.

5.3 Familiarity and Propensity to Trust

As mentioned in section 5.1, participants were asked to provide demographic information first. Following these questions participants were asked to provide information on their *Familiarity* with and their *Propensity to Trust* dialogue systems or chatbots on a 5-point Likert scale[16]. *Familiarity* and *Propensity to Trust* are both validated subscales of Körber's work on *Trust in Automation*[17], which can be used individually as well as in combination with other subscales (cf. Körber, 2019).

	HPHO	***HPLO**	**LPHO**	**LPLO**
Familiarity	4.12	4.25	4.1	3.5
Standard Deviation	0.84	0.35	1.24	1.23
Standard Error	0.18	0.25	0.56	0.61
Propensity to Trust	2.59	3.17	3.0	2.67
Standard Deviation	0.41	0.71	0.62	0.72
Standard Error	0.09	0.50	0.28	0.36

Table 1: Mean values for the subscales *Familiarity* and *Propensity to Trust* with respect to the four policy models developed for this study. Values marked with * are considered to be not representative due to a low number of participants.

To test whether any statistical significance can be assigned to these values, a `Multivariate analysis of variance (MANOVA)` was conducted. Unfortunately, the results showed that there were no statistically significant differences across conditions and across policy models (with p-values bigger than 0.05). Therefore, no further statistical tests were applied.

[16]See footnote 11

[17]More information can be found on Github: `https://github.com/moritzkoerber/TiA_Trust_in_Automation_Questionnaire`

For the purpose of this thesis, I decided to look at the subscales individually. Moreover, I decided to check the user's familiarity with and their propensity to trust similar systems prior to engaging with the dialogue system as those values were, in my opinion, not likely to change throughout the course of this experiment. Table 1 provides mean values for each of the two subscales as well as their `standard deviation (SD)` and `standard error of the mean (SEM)` while figures 10 and 11 offer a visual representation of the mean values per policy.

As can be seen in table 1 and figures 10 and 11 HPHO participants (M=4.12, SD=0.84) as well as HPLO (M=4.25, SD=1.24) and LPHO (M=4.1, SD=1.24) seemed to be more familiar with similar systems compared to LPLO participants (M=3.5, SD=1.23). Although the SEM indicates that for LPHO and LPLO the answers provided were more spread out than for the other two models, I would argue that on average most participants were rather familiar with dialogue systems.

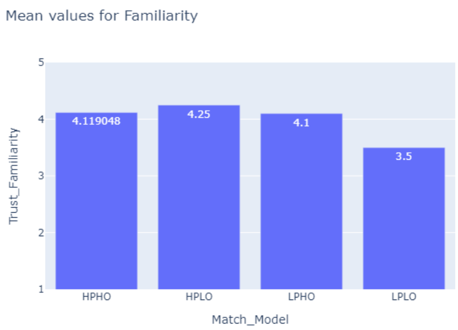


Figure 10: Mean values for the subscale *Familiarity* across the four policy versions.

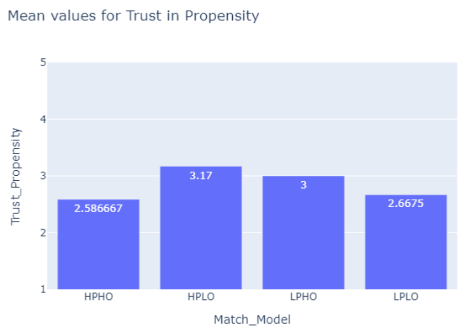


Figure 11: Mean values for the subscale *Propensity to Trust* across the four policy versions.

Interestingly, when it comes to participant's propensity to trust dialogue systems on average they seemed rather neutral. HPHO participants (M=2.59, SD=0.41) seem to rather mistrust dialogue systems compared to their LPLO counterparts (M=2.67, SD=0.72). This difference might be explained due to HPHO having interacted with more similar systems and potentially learning that there are limitations depending on the system. While LPLO participants indicated less familiarity and therefore, their answers could be spread out more depending on the number of interactions they already had.

Overall, participants seem to be rather familiar but neutral in regards to trusting a dialogue system, which might be an interesting starting point for future experiments.

6 Evaluation of WOZ study

As discussed in section 5, survey data from 32 participants was collected and evaluated after one participant had to be excluded from the evaluation. After interacting with each dialogue system, users were asked to rate their experience. They were further asked to only consider the latest interaction. To avoid bias, users were either assigned to start with their matching policy or their opposite. As mentioned in section 5.1, a total of 21 participants (65.6 %) were labeled as `HPHO`, followed by 5 `LPHO` (15.6 %), 4 `LPLO` (12.5 %) and lastly 2 `HPLO` (6.25 %). Even though two participants cannot be representative or statistically significant, I have decided to nevertheless include them for further evaluation if applicable.

It should be noted that for all scales discussed in the following sections, a `Multivariate analysis of variance (MANOVA)` was conducted. Unfortunately, the results showed again that there were no statistically significant differences across conditions and across policy models (with p-values bigger than 0.05). Therefore, no further statistical evaluation methods were applied.

Further, it needs to be mentioned, that for the following sections, the policy version will be mentioned together with the condition. For a better overview, table 2 provides a mapping of the policies match and mismatch/opposite. For example, if a user's self-reported openness and proactivity values indicated that they were highly proactive as well as highly open, they were assigned `HPHO` as their match, i.e. `HPHO-Match`. As their mismatch or as their opposite they were given `LPLO`, i.e. `LPLO-Opposite`. Therefore, it is crucial to keep in mind, that figures 12 to 19 should be compared from the inside out. For example, the blue bar on the left in figure12

Match	Opposite
HPHO	LPLO
$*$HPLO	$*$LPHO
LPHO	HPLO
LPLO	HPHO

Table 2: Mapping of policies to their opposites. Rows marked with a * are considered to be not representative due to a low number of participants.

represents user's `HPHO-Match` values while the orange bar on the right represents their mismatch, `LPLO-Opposite`.

With respect to the two participants in the HPLO-Match policy and their counterparts, i.e. LPHO-Opposite, it needs to be stressed again, that their scores cannot be perceived as statistically significant but will be considered for evaluation if applicable.

6.1 Likability

To answer the question of whether a user likes a dialogue system more based on the approach it takes, the `Likability` scale from the `Godspeed questionnaire` was used (cf. Bartneck et al., 2009).

As can be seen in figure 12 and verified in table 5 participants seemed to perceive both dialogues as equally likable. For example, the HPHO-Match (M=4.28, SD=0.66, SEM=0.14) policy was even perceived as slightly less likable than its counterpart LPLO-Opposite (M=4.4, SD=0.68, SEM=0.15).

6.2 Reliability/Competence, Understanding/Predictability and Trust in Automation

Although not directly part of the research questions, I was curious to know whether participants would perceive differences between both approaches in regards to three

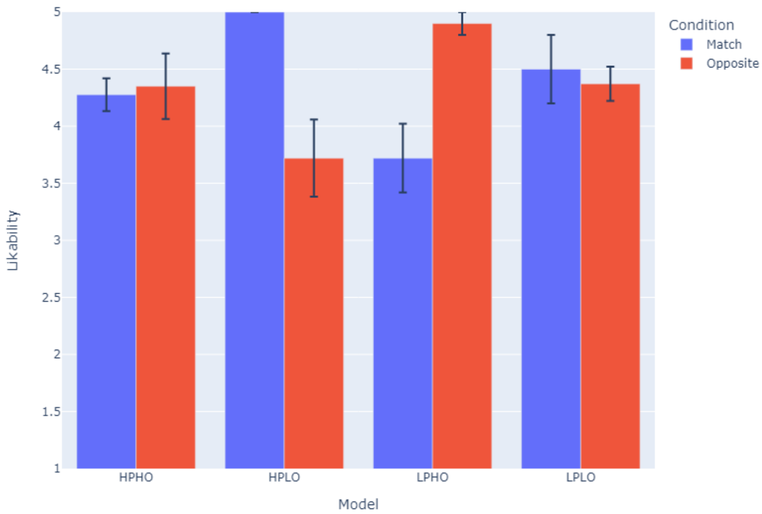


Figure 12: Mean values and standard error for Godspeed Likability: Match and Opposite condition grouped by policy version.

more subscales by Körber (2019). Therefore, I took into account Reliability/Competence, Understanding/Predictability and Trust in Automation.

For Reliability/Competence, users were asked to evaluate their impression of the dialogue systems' capabilities. Unfortunately, no differences were reported. For example, the HPHO-Match (M=3.48, SD=0.61, SEM=0.13) policy was perceived as almost equally capable as its counterpart LPLO-Opposite (M=3.90, SD=0.50, SEM=0.11), as can be seen in figure 13 and verified in table 6.

For the subscale Understanding/Predictability users were asked to report how well they were able to follow and understand the dialogue systems' next step. Figure 14 offers a visual representation while table 7 offers exact values. Interestingly, HPHO-Match (M=3.98, SD=0.76, SEM=0.166) users seemed to better understand how their mismatching policy LPLO-Opposite (M=4.45, SD=0.46, SEM=0.10) works. Similarly, LPHO-Match (M=4.3, SD=0.57, SEM=0.26) users seemed to be able to follow the system's way of thinking better than with their HPLO-Opposite (M=4.05, SD=0.48, SEM=0.22) counterpart.

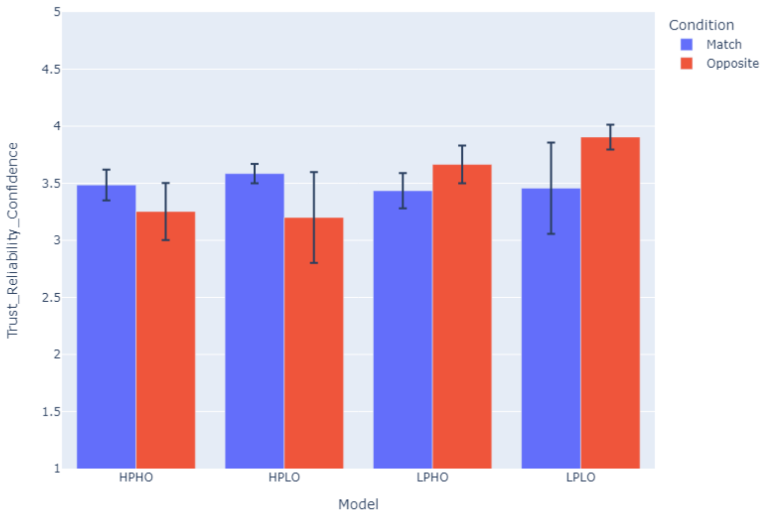


Figure 13: Mean values and standard error for Reliability/Competence: Match and Opposite condition grouped by policy version.

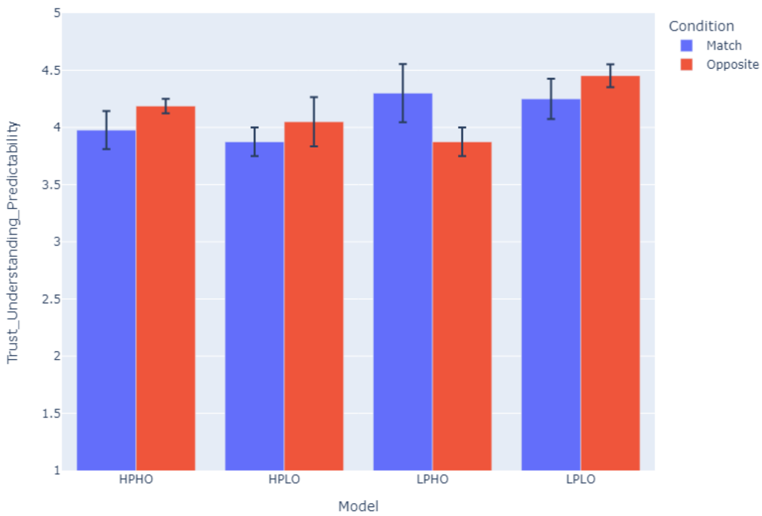


Figure 14: Mean values and standard error for Understanding/Predictability: Match and Opposite condition grouped by policy version.

The subscale Trust in Automation focuses on the user's trust in the system's ability. The only numerical difference reported was for LPHO-Match (M=4.1, SD=0.22, SEM=0.10) and their counterparts HPLO-Opposite (M=3.4, SD=0.96, SEM=0.43). Nevertheless, HPLO-Opposite's standard deviation indicates that answers were rather spread out and might therefore not be an accurate representation. A visual representation of the mean values is provided in figure 15 while numerical information can be found in table 8.

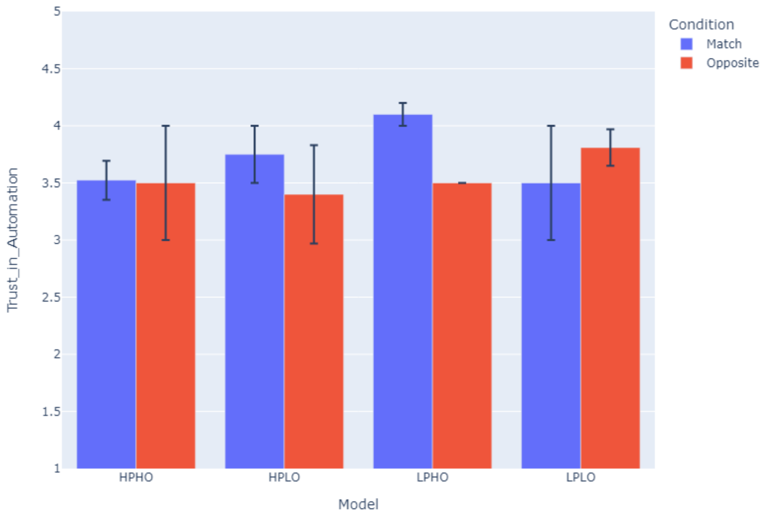


Figure 15: Mean values and standard error for Trust in Automation: Match and Opposite condition grouped by policy version.

6.3 System Usability

To answer the question of whether there are any perceived differences in the system's usability based on the approach it takes, the System Usability Scale by Brooke et al. (1996) was used.

As can be seen in figure 16 and respectively in table 9, LPLO-Match (M=81.25, SD=10.30, SEM=5.15) as well as LPLO-Opposite (M=84.88, SD=8.82, SEM=1.92)

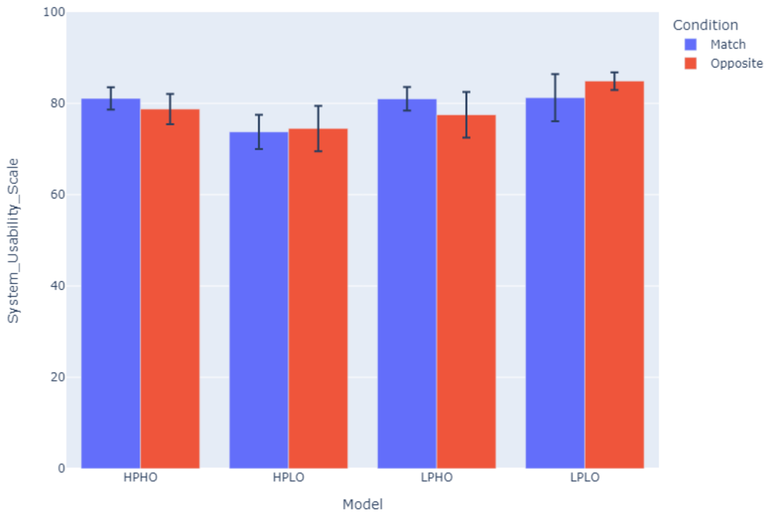


Figure 16: Mean values and standard error for System Usability: Match and Opposite condition grouped by policy version.

were considered to be more user-friendly than their counterparts HPHO-Opposite (M=78.75, SD=6.61, SEM=3.31) and HPHO-Match (M=81.07, SD=11.22, SEM=2.44). Similarly, users perceived LPHO-Match (M=81.0, SD=5.76, SEM=2.57) and LPHO-Opposite* (M=77.5, SD=7.07, SEM=5.00) as slightly better usable compared to their counterparts HPLO-Opposite (M=74.5, SD=11.10, SEM=4.96) and HPLO-Match* (M=73.75, SD=5.30, SEM=3.75).

6.4 Correlation

To answer the question of whether there is a difference between liking/enjoying the interaction with the dialogue system and liking the playlist, users were asked to rate their overall impression. Figures 17, 18 and 19 as well as tables 10, 11 and 12 show that with respect to mean values there seems to be no statistically significant difference.

Nevertheless, Pearson's correlation coefficient was calculated to make sure that relevant correlations were not missed. Table 13 provides correlation values as well as

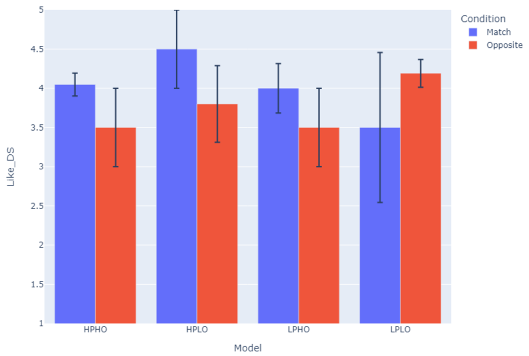


Figure 17: Mean values and standard error for "Overall liking the DS": Match and Opposite condition grouped by policy version.

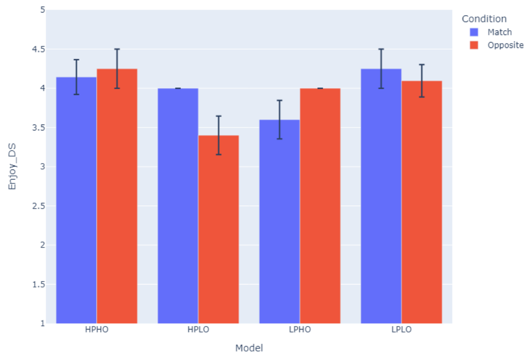


Figure 18: Mean values and standard error for "Overall enjoying the interaction with the DS": Match and Opposite condition grouped by policy version.

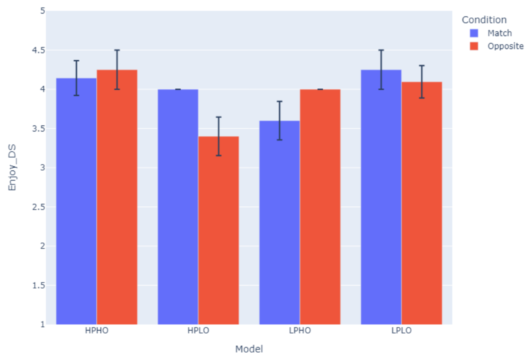


Figure 19: Mean values and standard error for "Overall enjoying the interaction with the DS": Match and Opposite condition grouped by policy version.

p-values across all submitted answers. Although correlation values are rather close to zero, a more fine-grained distinction seems necessary[18].

Therefore, values were first divided with respect to their conditions, namely

[18]With respect to correlation, no visual representation is provided due to a lack of data points.

match and opposite. Table 14 provides an overview of correlation and p-values. Interestingly, there seems to be a moderate correlation for the Match condition between participants enjoying the dialogue system and liking the playlist (ρ=0.498, p=0.004). Further, there seems to be a moderate correlation between participants liking the dialogue system and liking the playlist (ρ=0.46, p=0.008). For the opposite condition, there seems to be no relevant correlation between the dialogue system and the playlist.

To further investigate correlation, the four policies were investigated for correlations between liking the dialogue system, enjoying the dialogue system, and liking the playlist. Table 15 offers an overview of the correlation coefficients with respect to the policy. A strong negative correlation can be observed for the LPHO policy between enjoying the dialogue system and liking the playlist (ρ=-0.73, p=0.062). For the HPHO policy, a moderate correlation (rho=0.568, p=0.003) can be seen between enjoying the dialogue system and liking the playlist. Correlation values for the HPLO and LPLO policies were below ρ=0.5 and will not be discussed further.

6.5 Content Analysis

As participants were asked but not obligated to provide free-form positive and negative feedback on their experience with the dialogue system, the content of them was analyzed manually using Excel[19]. Therefore, labels for all responses provided were created and counted based on the policy model and condition. A selection of positive and negative feedback along with their respective labels is provided in A.8. Further, a selection of more general feedback and answers on what the user would like to change are provided there as well. Figure 20 shows that three labels were assigned significantly more often than others, namely *competent* (21), *good recommendations* (19), and *slow* (19). The latter is related to the users being unaware that they were communicating with a human rather than a dialogue system and was therefore expected. Nevertheless, it is interesting to note that users described the dialogue system as competent and enjoyed the recommendations provided.

[19]Excel provides the option to count occurrences of words.

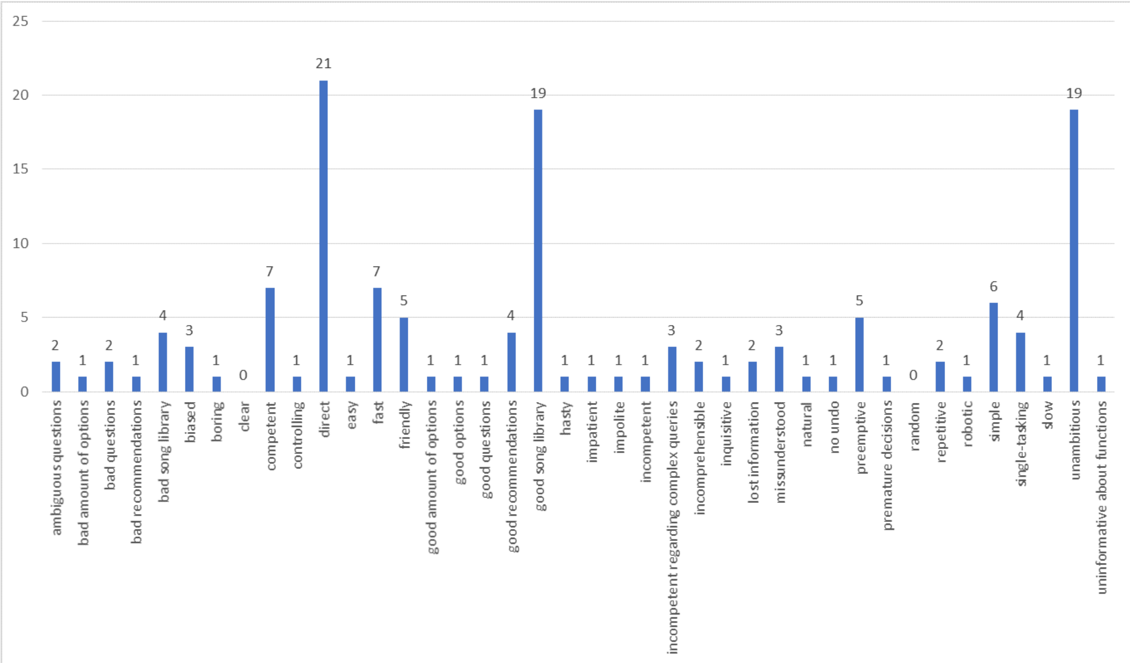


Figure 20: Distribution of assigned content labels in alphabetical order.

7 Results and Discussion

Initially, three questions were asked following the hypothesis that an individual's personality can arguably influence their lives and should therefore be relevant when interacting with dialogue systems as well.

With respect to the question of whether a user likes a dialogue system more based on the approach it takes, the answer seems to be that within the scope of this study, no significant differences could be proven. In other words, as long as the system is able to fulfill its task, users do not seem to be bothered much in terms of likability. This might be explained by users being aware that they are talking to a machine rather than a human being.

For Reliability/Competence, users were asked to evaluate their impression of the dialogue systems' capabilities. Unfortunately, no differences were reported. This might be related to the fact that all users achieved their goal of creating a playlist and therefore might even be a good sign as all policy models yielded the same results in terms of being competent. For the subscale Understanding/Predictability users were asked to report how well they were able to follow and understand the dialogue systems' next step. Interestingly, HPHO-Match (M=3.98, SD=0.76, SEM=0.166) users seemed to better understand how their mismatching policy LPLO-Opposite (M=4.45, SD=0.46, SEM=0.10) works. This can be explained due to the LPLO policy initially asking for liked genres and artists while the HPHO policy offered recommendations based on feedback and could have been perceived as random or unpredictable by users. Similarly, LPHO-Match (M=4.3, SD=0.57, SEM=0.26) users seemed to be able to follow the system's way of thinking better than with their HPLO-Opposite (M=4.05, SD=0.48, SEM=0.22) counterpart. This could be explained by the LPHO model offering a variety of songs while HPLO focused primarily on asking the user for the next song to add. The subscale Trust in Automation focuses on the user's trust in the system's ability. The only numerical difference reported was for LPHO-Match (M=4.1, SD=0.22, SEM=0.10) and their counterparts HPLO-Opposite (M=3.4, SD=0.96, SEM=0.43). Nevertheless, HPLO-Opposite's standard deviation indicates that answers were rather spread out and might therefore not be an accurate representation. Nevertheless, this could be explained by the number of participants. As only 5 participants were assigned the LPHO-Match policy.

To answer the question of whether there are any perceived differences in the system's usability based on the approach it takes, the System Usability Scale by Brooke et al. (1996) was used. The LPLO-Match (M=81.25, SD=10.30, SEM=5.15) as well as LPLO-Opposite (M=84.88, SD=8.82, SEM=1.92) were considered to be more user-friendly than their counterparts HPHO-Opposite (M=78.75, SD=6.61, SEM=3.31) and HPHO-Match (M=81.07, SD=11.22, SEM=2.44). This can be explained with the LPLO policy first establishing a baseline user model by asking the user for preferences. Similarly, users perceived LPHO-Match (M=81.0, SD=5.76, SEM=2.57) and LPHO-Opposite* (M=77.5, SD=7.07, SEM=5.00) as slightly better usable compared to their counterparts HPLO-Opposite (M=74.5, SD=11.10, SEM=4.96) and HPLO-Match* (M=73.75, SD=5.30, SEM=3.75). Again, this could be explained due to the LPHO policy offering suggestions rather than directly asking the user for songs to add to the playlist.

To answer the question of whether there is any difference between liking/enjoying the interaction with the dialogue system and liking the playlist, users were asked to rate their overall impression. Therefore, all values were taken into consideration first which lead to no true correlation. To further investigate, values were then divided with respect to their conditions, namely match and opposite. Interestingly, there seems to be a moderate correlation for the Match condition between participants enjoying the dialogue system and liking the playlist (ρ=0.498, p=0.004). Further, there seems to be a moderate correlation between participants liking the dialogue system and liking the playlist (ρ=0.46, p=0.008). For the opposite condition, there seems to be no correlation between the dialogue system and the playlist. This finding is not surprising, as it was expected that the users would enjoy their matching systems more than their opposites. Nevertheless, it is an interesting observation.

To further investigate correlation, the four policies were investigated for correlations between liking the dialogue system, enjoying the dialogue system and liking the playlist. A strong negative correlation can be observed for the LPHO policy between enjoying the dialogue system and liking the playlist (ρ=-0.73, p=0.062). For the HPHO policy a moderate correlation (rho=0.568, p=0.003) can be seen between enjoying the dialogue system and liking the playlist. It seems like playlists and conversations were enjoyed more if the system was offering more songs to the users rather than

Analyzing the free-form feedback users provided showed that three labels were used significantly more often than others, namely *competent* (21), *good recommendations* (19), and *slow* (19). The latter is related to the users being unaware that they were communicating with a human rather than a dialogue system and was therefore expected. Nevertheless, it is interesting to note that users described the dialogue system as competent and enjoyed the recommendations provided. Users were also asked what they would have liked to add. Some users suggested that it would be great to include music from other platforms as Spotify is rather limited when it comes to foreign songs. Some users would have enjoyed more songs and longer samples to listen to, but overall users were rather contented with the dialogue system.

8 Conclusion

For the purpose of this thesis, dialogue systems have been discussed, followed by an overview of adaptive dialogue systems and recommender systems. Further, personality and traits were discussed. The traits Openness and Proactivity were chosen as axes for adaptation and four policies have been designed in order to test adaptation based on a user's personality. Therefore, a Wizard of Oz study was designed, conducted and evaluated. Unfortunately, no statistically significant results were obtained. Nevertheless, some interesting findings were discussed above and might offer grounds for future research.

In general, users seemed to have enjoyed the experience of creating a playlist using a dialogue system. During debriefings with some participants, it was mentioned that they generally enjoyed having the option to listen to the songs, and some reported having discovered new songs they liked. Further, a small number of participants admitted to not having properly read the provided survey instructions. This lead me to believe that participant bias might be a reason for these rather insignificant results and that it might be worth the effort to try and repeat this study with more as well as paid participants using e.g. Amazon Mechanical Turk rather than acquaintances.

It should be noted that two traits have been used for this study. Including more than two traits might yield vastly different results and should therefore be considered as a possibility for future work. Further, including more possibilities for a user might be helpful. For example, offering information on a song like lyrics or offering more than two songs at the same time when asked for recommendations. Also, the ability to create a song hierarchy could be interesting to include for future research.

9 References

Almlund, M., A. L. Duckworth, J. Heckman, and T. Kautz (2011). Personality psychology and economics. In *Handbook of the Economics of Education*, Volume 4, pp. 1–181. Elsevier.

Anglim, J. and P. O'connor (2019). Measurement and research using the big five, hexaco, and narrow traits: A primer for researchers and practitioners. *Australian Journal of Psychology 71*(1), 16–25.

Bartneck, C., D. Kulić, E. Croft, and S. Zoghbi (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics 1*, 71–81.

Bateman, T. S. and J. M. Crant (1993). The proactive component of organizational behavior: A measure and correlates. *Journal of organizational behavior 14*(2), 103–118.

Bateman, T. S. and J. M. Crant (1999). Proactive behavior: Meaning, impact, recommendations. *Business Horizons 42*(3), 63–70.

Bennett, J., S. Lanning, et al. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, Volume 2007, pp. 35. New York.

Billsus, D. and M. Pazzani (1997). Learning probabilistic user models. In *UM97 Workshop on Machine Learning for User Modeling*. Citeseer.

Brislin, R. W. and K. D. Lo (2006). Culture, personality and people's uses of time: Key interrelationships. *Comprehensive handbook of personality and psychopatology*, 44–64.

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

Burke, R. (2007). Hybrid web recommender systems. *The adaptive web: methods and strategies of web personalization*, 377–408.

Buss, D. M. (1991). Evolutionary personality psychology. *Annual review of psychology 42*(1), 459–491.

Chen, H., X. Liu, D. Yin, and J. Tang (2017). A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter 19*(2), 25–35.

Chen, Y.-N., A. Celikyilmaz, and D. Hakkani-Tur (2017). Deep learning for dialogue systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 8–14.

Cobb-Clark, D. A. and S. Schurer (2012). The stability of big-five personality traits. *Economics Letters 115*(1), 11–15.

Costa, P. T. and R. R. McCrae (1992). Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment 4*(1), 5.

Costa Jr, P. T. and R. R. McCrae (2006). Trait and factor theories. *Comprehensive handbook of personality and psychopatology*, 96–114.

Costa Jr, P. T., A. Terracciano, and R. R. McCrae (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of personality and social psychology 81*(2), 322.

Dahlbäck, N., A. Jönsson, and L. Ahrenberg (1993). Wizard of oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*, pp. 193–200.

de Coning, C., A. Muto, and R. Saraiva (2022). *Adaptive Mediation and Conflict Resolution: Peace-Making in Colombia, Mozambique, the Philippines, and Syria.* Springer International Publishing AG.

DeYoung, C. G., L. C. Quilty, and J. B. Peterson (2007). Between facets and domains: 10 aspects of the big five. *Journal of personality and social psychology 93*(5), 880.

Dow, S., B. MacIntyre, J. Lee, C. Oezbek, J. D. Bolter, and M. Gandy (2005). Wizard of oz support throughout an iterative design process. *IEEE Pervasive Computing 4*(4), 18–26.

Feist, J. and G. J. Feist (2006). Theories of personality.

Fischer, G. (2001). User modeling in human–computer interaction. *User modeling and user-adapted interaction 11*, 65–86.

Fischer, G., A. Lemke, and T. Schwab (1985). Knowledge-based help systems. *ACM SIGCHI Bulletin 16*(4), 161–167.

Gazzaniga, M. S. (2000). Human-the science behind what makes us unique.

Goldberg, L. R. et al. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe 7*(1), 7–28.

Herlocker, J. L., J. A. Konstan, L. G. Terveen, and J. T. Riedl (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS) 22*(1), 5–53.

Jackson, D. N. and P. F. Tremblay (2002). The six factor personality questionnaire.

Jaimes, A. and N. Sebe (2007). Multimodal human–computer interaction: A survey. *Computer vision and image understanding 108*(1-2), 116–134.

Janarthanam, S. and O. Lemon (2014). Adaptive generation in dialogue systems using dynamic user modeling. *Computational Linguistics 40*(4), 883–920.

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of research in personality 51*, 78–89.

Jurafsky, D. and J. H. Martin (2009). *Speech and Language Processing: An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2 ed.). Pearson Education.

Kiani, K., G. Cui, A. Bunt, J. McGrenere, and P. K. Chilana (2019). Beyond "one-size-fits-all" understanding the diversity in how software newcomers discover and make use of help resources. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*, pp. 13–30. Springer.

Lee, K. and M. C. Ashton (2004). Psychometric properties of the hexaco personality inventory. *Multivariate behavioral research 39*(2), 329–358.

Leskovec, J., A. Rajaraman, and J. D. Ullman (2020). *Mining of massive data sets.* Cambridge university press.

Li, C.-Y., D. Ortega, D. Väth, F. Lux, L. Vanderlyn, M. Schmidt, M. Neumann, M. Völkel, P. Denisov, S. Jenne, Z. Kacarevic, and N. T. Vu (2020). Adviser: A toolkit for developing multi-modal, multi-domain and socially-engaged conversational agents.

Litman, D. J. and S. Pan (2002). Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction 12*, 111–137.

Lund, B. D. and T. Wang (2023). Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library Hi Tech News 40*(3), 26–29.

McCrae, R. R. and P. T. Costa (2003). *Personality in adulthood: A five-factor theory perspective.* Guilford Press.

Nadkarni, P. M., L. Ohno-Machado, and W. W. Chapman (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association 18*(5), 544–551.

Putnam, L. L. (2013). Definitions and approaches to conflict and communication. In J. G. Oetzel and S. Ting-Toomey (Eds.), *The SAGE Handbook of Conflict Communication: Integrating Theory, Research, and Practice* (2 ed.)., Chapter 1, pp. 1–40. SAGE Publications, Inc.: Los Angeles, London a.o.

Rammstedt, B., L. R. Goldberg, and I. Borg (2010). The measurement equivalence of big-five factor markers for persons with different levels of education. *Journal of research in personality 44*(1), 53–61.

Ricci, F., L. Rokach, and B. Shapira (2015). Recommender systems: introduction and challenges. *Recommender systems handbook*, 1–34.

Roberts, B. W. and D. Mroczek (2008). Personality trait change in adulthood. *Current directions in psychological science 17*(1), 31–35.

Rothmann, S. and E. P. Coetzer (2003). The big five personality dimensions and job performance. *SA Journal of industrial psychology 29*(1), 68–74.

Shiner, R. and A. Caspi (2003). Personality differences in childhood and adolescence: Measurement, development, and consequences. *Journal of child psychology and psychiatry 44*(1), 2–32.

Su, P.-H., M. Gasic, N. Mrksic, L. Rojas-Barahona, S. Ultes, D. Vandyke, T.-H. Wen, and S. Young (2016). On-line active reward learning for policy optimisation in spoken dialogue systems. *arXiv preprint arXiv:1605.07669*.

Ultes, S., L. M. R. Barahona, P.-H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T.-H. Wen, M. Gasic, et al. (2017). Pydial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pp. 73–78.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM 9*(1), 36–45.

Wen, T.-H., D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young (2016). A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Yan, Z., N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li (2017). Building task-oriented dialogue systems for online shopping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 31.

A Appendices

A.1 Validated scale: Adventurousness

Parameter	**Value?**
+ keyed	Prefer variety to routine.
+ keyed	Like to visit new places.
+ keyed	Interested in many things.
+ keyed	Like to begin new things.
- keyed	Prefer to stick with things that I know.
- keyed	Dislike changes.
- keyed	Don't like the idea of change.
- keyed	Am a creature of habit.
- keyed	Dislike new foods.
- keyed	Am attached to conventional ways.

Table 3: Validated scale on Adventurousness

A.2 Validated scale: Proactivity

Parameter	**Value?**
+ keyed	I am constantly on the lookout for new ways to improve my life.
+ keyed	I feel driven to make a difference in my community, and maybe the world.
+ keyed	Wherever I have been, I have been a powerful force for constructive change.
+ keyed	I enjoy facing and overcoming obstacles to my ideas.
+ keyed	Nothing is more exciting than seeing my ideas turn into reality.
+ keyed	If I see something I don't like, I fix it.
+ keyed	No matter what the odds, if I believe in something I will make it happen.
+ keyed	I love being a champion for my ideas, even against other's opposition.
+ keyed	I excel at identifying opportunities.
+ keyed	I am always looking for better ways to do things.
+ keyed	If I believe in an idea, no obstacle will prevent me from making it happen.
+ keyed	I love to challenge the *status quo*.
+ keyed	When I have a problem, I tackle it head-on.
+ keyed	I am great at turning problems into opportunities.
+ keyed	I can spot a good opportunity long before others can.
+ keyed	If I see someone in trouble, I help out in any way I can.
− keyed	I tend to let others take the initiative to start new projects.

Table 4: Validated scale on Proactivity

A.3 List of System and User Actions

System Acts and associated templates:

- `Welcome()`: Welcome to the Playlist Creator. Let's work on building a playlist with 5 songs together.
- `BadAct()`: I'm sorry, I didn't understand that.
- `Clarify()`: Would you mind being a little more specific?
- `BadSong()`: I'm sorry, I cannot find this song.
- `Bye()`: Thank you, goodbye.
- `Inform_Add_Song()`: [Ok, / Sure / None] I added it to your playlist.
- `Inform_Accept()`: [Sure, no problem. / Sure. / No problem. / Okay. / Of course. / Alright.]
- `Inform_Delete()`: I have deleted [song/genre/artist] for you.
- `Confirm_Choice()`: Okay. Let's go with [genre/artist].
- `Confirm_Happiness()`: How happy are you with your playlist on a scale from 1 (not happy at all) to 5 (very happy)?
- `Confirm_Change()`: Would you like to change a song? If yes, which song would you like to remove?
- `Request_Genre()`: Which genre(s) do you like?
- `Request_Artist()`: Is there an artist you like?
- `Request_Song()`: Is there a particular song you would like to add?
- `Request_Add_Song()`: What song should I add [first/ second/ next]?
- `Request_General_Preference()`: Do you like any of them?

- `Request_Add_More()`: Would you like to add more songs from this artist or would you like to change the artist or genre?
- `Request_Preference()`: Is there a [genre/or/artist] you prefer?
- `Request_Confirm_Song()`: Is this the song you were referring to: [song]?
- `Request_Opinion()`: What do you think about [it/them]?
- `Request_Add_Song()`: Would you like to add [it/them] to your playlist?
- `Recommend_Popular_Genre()`: Here are some of the most popular genres: Rock, Pop, Electro, HipHop/Rap.
- `Recommend_Example_Artist()`: Popular artists in the genre include: [2-3 artists].
- `Recommend_Song()`: Here are some songs I'd like to recommend: [2-3 songs]
- `Recommend_Trending_Song()`: Here is a song that is trending right now: [song]
- `Recommend_One_Song()`: Here is another [popular] song I'd like to recommend: [song]
- `Recommend_More_Songs()`: Here are more songs by this artist: [2-3 sonsgs]
- `Recommend_Popular_Songs()`: Currently, these songs are popular [2-3 songs, different genres].
- `Request_Song_Feedback()`: Is there a specific reason you did not like [it/them/the other songs]? For example the genre or artist?

Accepted User Acts – It should be noted that 'none' corresponds to 'random'

- `Bad_User_Act()`
- `Inform() and Change()`: with slots and values
  - genre=[none, genre]
  - artist=[none, artist, partial_name]
  - song=[none, song, partial_name]
  - other_artist=[none, artist, partial_name]
  - like=boolean
  - title=[none, title, title_part]
  - album=[none, album, year]
  - year=[none, year]
  - version=['live', 'studio', 'acoustic']
  - type=[none, type]
  - happiness[scale=scale]
  - add(song=song)
  - remove(song=song)
  - confirm=boolean
- `Request_Recommend()`:
  - genre=[genre]
  - artist=[artist]
  - song=[song]
  - album[album]
  - _Memory: Request_Recommend_Memory(genre=genre, artist=artist)

- `Request_Add_Song()`:
  - song=[song]
  - artist=artist
- `Confirm_Song()`:
  - true
  - false

A.4 User Survey Instructions

Thank you for participating!

This study is part of my master's thesis at the Institute for Natural Language Processing at the University of Stuttgart, Germany. The aim is to get a better understanding of Human-Computer-Interaction based on user preferences.
Within the next 45-60 minutes You will interact with two dialog systems, creating a playlist of your choice with each dialog system and rating the interaction afterwards. More information on the task is provided below.

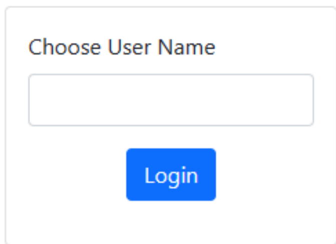
Your Task

Please, make sure to follow each step individually before proceeding to the next one. In case of language barriers, feel free to use a translation tool of your choice.

Step 1: Access the Dialogue System

Please, click on or copy the link to access the study and choose a **username** that cannot be traced back to you:

http://193.196.53.252:8005/

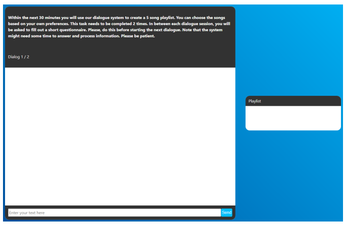


Step 2: Questionnaire

Please, fill out the questionnaire first and agree to the data collection. No personal information will be published or distributed. Your answers will only be used for the study and will be stored anonymously.

Step 3: Create a 5 song playlist

On the right hand side the user interface is displayed. It consists of a chat window and the current status of your playlist.
You can choose **five songs** for your playlists based on your own preferences. The dialogue system will accept your input on `genre`, `artist(s)`, `title`, `album`, `year` and `version`. You can also request a recommendation or ask to change an already picked song.
You will be provided with short sound samples from Spotify. You do not need an account to listen to it.
Please note that the dialog systems might need some time to answer and process information. Please be patient. The dialog systems will provide you with an answer as soon as possible.



Step 4: Survey for previous interaction

Please, fill out the survey questions - you will be asked to rate the conversation and the system based on how you feel about it.
Make sure to only rate your interaction with the current dialogue system. You will then automatically be redirected back to the chat interface to interact with the second dialog system. Here follow the same instructions for Step 3 again. Inbetween dialogues, the system may need a few seconds to load. Please wait until you have received the first message from the system.

A.5 Personality Questionnaire - WOZ Study

Questionnaire

Demographic Information

Please indicate the gender you most identify with.

- ◯ Male
- ◯ Female
- ◯ Other

Please indicate your age

What is your educational background?

Please indicate your native language(s):

Please answer the following questions based on your opinion and/or experience.

Please mark how much you agree with the following statements

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
I already know similar systems (chatbots or dialog systems).	◯	◯	◯	◯	◯

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
I have already used similar systems (chatbots or dialog systems).	◯	◯	◯	◯	◯
One should be careful with unfamiliar automated systems.	◯	◯	◯	◯	◯
I rather trust a system than mistrust it.	◯	◯	◯	◯	◯
Automated systems generally work well.	◯	◯	◯	◯	◯

Please answer the following questions based on your own perception.

Please mark how much you agree with the following statements

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
I prefer variety to routine.	○	○	○	○	○
I like to visit new places.	○	○	○	○	○
I am interested in many things.	○	○	○	○	○
I like to begin new things.	○	○	○	○	○
I prefer to stick with things that I know.	○	○	○	○	○
I dislike changes.	○	○	○	○	○
I don't like the idea of change.	○	○	○	○	○
I am a creature of habit.	○	○	○	○	○
I dislike new foods.	○	○	○	○	○
I am attached to conventional ways.	○	○	○	○	○

Please mark how much you agree with the following statements

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
I am constantly on the lookout for new ways to improve my life.	◯	◯	◯	◯	◯
I feel driven to make a difference in my community, and maybe the world.	◯	◯	◯	◯	◯
I tend to let others take the initiative to start new projects.	◯	◯	◯	◯	◯
Wherever I have been, I have been a powerful force for constructive change.	◯	◯	◯	◯	◯
I enjoy facing and overcoming obstacles to my ideas.	◯	◯	◯	◯	◯
Nothing is more exciting than seeing my ideas turn into reality.	◯	◯	◯	◯	◯
If I see something I don't like, I fix it.	◯	◯	◯	◯	◯

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
No matter what the odds, if I believe in something I will make it happen.	◯	◯	◯	◯	◯
I love being a champion for my ideas, even against other's opposition.	◯	◯	◯	◯	◯
I excel at identifying opportunities.	◯	◯	◯	◯	◯
I am always looking for better ways to do things.	◯	◯	◯	◯	◯
If I believe in an idea, no obstacle will prevent me from making it happen.	◯	◯	◯	◯	◯
I love to challenge the *status quo*.	◯	◯	◯	◯	◯
When I have a problem, I tackle it head-on.	◯	◯	◯	◯	◯
I am great at turning problems into opportunities.	◯	◯	◯	◯	◯

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
I can spot a good opportunity long before others can.	◯	◯	◯	◯	◯
If I see someone in trouble, I help out in any way I can.	◯	◯	◯	◯	◯

Agreement to data collection

By clicking "Submit" you allow us to store and use the information you provided during this study. No personal information will be collected or distributed - all answers will be stored and evaluated anonymously and only for the duration of this study.

Submit

A.6 Questionnaire WOZ Study

Survey

Please, only rate your interaction with the current dialog system.

Feedback to the Dialog System

Please mark how much you agree with the following statements based on your perception

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
The dialog system is capable of interpreting situations correctly.	○	○	○	○	○
The dialog system works reliably.	○	○	○	○	○
A dialog system malfunction is likely.	○	○	○	○	○
The dialog system is capable of taking over complicated tasks.	○	○	○	○	○
The dialog system might make	○	○	○	○	○

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
sporadic errors.					
I am confident about the dialog system's capabilities.	◯	◯	◯	◯	◯
The dialog system state was always clear to me.	◯	◯	◯	◯	◯
The dialog system reacts unpredictably.	◯	◯	◯	◯	◯
I was able to understand why things happened.	◯	◯	◯	◯	◯
It's difficult to identify what the dialog system will do next.	◯	◯	◯	◯	◯
I trust the dialog system.	◯	◯	◯	◯	◯
I can rely on the dialog system.	◯	◯	◯	◯	◯

Feedback to the User Experience

Please rate how much you agree with the following statements about the dialog system's usability

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
I think that I would like to use this dialog system frequently.					
I found the dialog system unnecessarily complex.					
I thought the dialog system was easy to use.					
I think that I would need support of a technical person to be able to use this dialog system.					
I found the various functions in this dialog system were well integrated.					
I thought there was too much inconsistency					

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
in this dialog system.					
I would imagine that most people would learn to use this dialog system very quickly.	◯	◯	◯	◯	◯
I found the dialog system very cumbersome to use.	◯	◯	◯	◯	◯
I felt very confident using the the dialog system.	◯	◯	◯	◯	◯
I needed to learn a lot of things before I could get going with this dialog system.	◯	◯	◯	◯	◯

Please rate your overall impression of the dialog system

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
Overall, I liked using this dialog system.	◯	◯	◯	◯	◯
Overall, I enjoyed the	◯	◯	◯	◯	◯

	Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
way the dialog system communicated with me.					
Overall, I was satisfied with the playlist I created using this dialog system.	○	○	○	○	○

Please rate your impression of the dialog system's likability

	1	2	3	4	5	
Dislike	○	○	○	○	○	Like
Unfriendly	○	○	○	○	○	Friendly
Unkind	○	○	○	○	○	Kind
Unpleasant	○	○	○	○	○	Pleasant
Awful	○	○	○	○	○	Nice

What did you like about the interaction with this dialog system?

What did you dislike about the interaction with this dialog system?

What would you have changed about the system's behavior to make the interaction better for you?

General feedback

Is there any other feedback you'd like to provide? Were there any problems with the dialog system?

Submit

A.7 Evaluation values for WOZ study

Policy	Condition	Mean	Standard Deviation	Standard Error
HPHO	Match	4.27619	0.658714	0.143743
	Opposite	4.35	0.574456	0.287228
HPLO	*Match	5	0	0
	Opposite	3.72	0.756307	0.338231
LPHO	Match	3.72	0.672309	0.300666
	*Opposite	4.9	0.141421	0.100000
LPLO	Match	4.5	0.600000	0.300000
	Opposite	4.371429	0.684940	0.149466

Table 5: Mean, Standard Deviation and Standard Error for Godspeed Likability across policies and conditions. Rows marked with a * on their condition are considered to be not representative due to a low number of participants.

Policy	Condition	Mean	Standard Deviation	Standard Error
HPHO	Match	3.484762	0.614806	0.134162
	Opposite	3.2525	0.500558	0.250279
HPLO	*Match	3.585	0.120208	0.085000
	Opposite	3.2	0.891319	0.398610
LPHO	Match	3.434	0.344572	0.154097
	*Opposite	3.665	0.233345	0.165000
LPLO	Match	3.4575	0.799891	0.399945
	Opposite	3.904762	0.495909	0.108216

Table 6: Mean, Standard Deviation and Standard Error for Reliability/ Competence across policies and conditions. Rows marked with a * on their condition are considered to be not representative due to a low number of participants.

Policy	Condition	Mean	Standard Deviation	Standard Error
HPHO	Match	3.97619	0.762007	0.166284
	Opposite	4.1875	0.125000	0.062500
HPLO	*Match	3.875	0.176777	0.125000
	Opposite	4.05	0.480885	0.215058
LPHO	Match	4.3	0.570088	0.254951
	*Opposite	3.875	0.176777	0.125000
LPLO	Match	4.25	0.353553	0.176777
	Opposite	4.452381	0.458387	0.100028

Table 7: Mean, Standard Deviation and Standard Error for Understanding/ Predictability across policies and conditions. Rows marked with a * on their condition are considered to be not representative due to a low number of participants.

Policy	Condition	Mean	Standard Deviation	Standard Error
HPHO	Match	3.52381	0.782243	0.170700
	Opposite	3.5	1.00	0.50
HPLO	*Match	3.75	0.353553	0.25
	Opposite	3.4	0.961769	0.430116
LPHO	Match	4.1	0.223607	0.10
	*Opposite	3.5	0	0
LPLO	Match	3.5	1.0	0.50
	Opposite	3.809524	0.732738	0.159897

Table 8: Mean, Standard Deviation and Standard Error for Trust in Automation across policies and conditions. Rows marked with a * on their condition are considered to be not representative due to a low number of participants.

Policy	Condition	Mean	Standard Deviation	Standard Error
HPHO	Match	81.07143	11.224177	2.449316
	Opposite	78.75	6.614378	3.307189
HPLO	*Match	73.75	5.303301	3.750
	Opposite	74.5	11.096171	4.962358
LPHO	Match	81.0	5.755432	2.573908
	*Opposite	77.5	7.071068	5.00
LPLO	Match	81.25	10.307764	5.153882
	Opposite	84.88095	8.820296	1.924746

Table 9: Mean, Standard Deviation and Standard Error for System Usability across policies and conditions. Rows marked with a * on their condition are considered to be not representative due to a low number of participants.

Policy	Condition	Mean	Standard Deviation	Standard Error
HPHO	Match	4.047619	0.669043	0.145997
	Opposite	3.5	1.000000	0.500000
HPLO	*Match	4.5	0.707107	0.500000
	Opposite	3.8	1.095445	0.489898
LPHO	Match	4.0	0.707107	0.316228
	*Opposite	3.5	0.707107	0.500000
LPLO	Match	3.5	1.914854	0.957427
	Opposite	4.190476	0.813575	0.177537

Table 10: Mean, Standard Deviation and Standard Error for Overall liking the Dialogue System across policies and conditions. Rows marked with a * on their condition are considered to be not representative due to a low number of participants.

Policy	Condition	Mean	Standard Deviation	Standard Error
HPHO	Match	4.142857	1.014185	0.221313
	Opposite	4.25	0.500000	0.250000
HPLO	*Match	4	0	0
	Opposite	3.4	0.547723	0.244949
LPHO	Match	3.6	0.547723	0.244949
	*Opposite	4.0	0	0
LPLO	Match	4.25	0.500000	0.250000
	Opposite	4.095238	0.943650	0.205921

Table 11: Mean, Standard Deviation and Standard Error for Overall enjoying the interaction with the Dialogue System across policies and conditions. Rows marked with a * on their condition are considered to be not representative due to a low number of participants.

Policy	Condition	Mean	Standard Deviation	Standard Error
HPHO	Match	4.142857	1.152637	0.251526
	Opposite	4.75	0.500000	0.250000
HPLO	*Match	4.5	0.707107	0.500000
	Opposite	4.8	0.447214	0.200000
LPHO	Match	4.4	0.547723	0.244949
	*Opposite	4.5	0.707107	0.500000
LPLO	Match	4.75	0.500000	0.250000
	Opposite	4.666667	0.577350	0.125988

Table 12: Mean, Standard Deviation and Standard Error for Enjoying the Playlist created with the Dialogue System across policies and conditions. Rows marked with a * on their condition are considered to be not representative due to a low number of participants.

Value 1	Value 2	Pearson	p-value
Like DS	Like Playlist	0.267	0.033
Like DS	Enjoy DS	0.319	0.010
Enjoy DS	Like Playlist	0.297	0.017

Table 13: Pearson's correlation coefficient ρ (rho) with p-value across all submitted answers.

Condition	Value 1	Value 2	Pearson	p-value
Match	Like DS	Like Playlist	0.259	0.152
	Like DS	Enjoy DS	0.460	0.008
	Enjoy DS	Like Playlist	0.498	0.004
Opposite	Like DS	Like Playlist	0.343	0.055
	Like DS	Enjoy DS	0.174	0.341
	Enjoy DS	Like Playlist	1.388e-17	1.0

Table 14: Pearson's correlation coefficient ρ (rho) with p-value with respect to Match and Opposite conditions.

Policy	**Value 1**	**Value 2**	**Pearson**	**p-value**
HPHO	Like DS	Like Playlist	0.324	0.114
	Like DS	Enjoy DS	0.430	0.032
	Enjoy DS	Like Playlist	0.568	0.003
HPLO	Like DS	Like Playlist	0	1.0
	Like DS	Enjoy DS	0	1.0
	Enjoy DS	Like Playlist	0.091	0.846
LPHO	Like DS	Like Playlist	-0.258	0.576
	Like DS	Enjoy DS	0.353	0.437
	Enjoy DS	Like Playlist	-0.730	0.062
LPLO	Like DS	Like Playlist	0.407	0.407
	Like DS	Enjoy DS	0.308	0.134
	Enjoy DS	Like Playlist	0.081	0.698

Table 15: Pearson's correlation coefficient ρ (rho) with p-value with respect to Match and Opposite conditions.

A.8 Content Analysis Examples

Label	Sum	HPHO-M	HPHO-O	HPLO-M	HPLO-O	LPHO-M	LPHO-O	LPLO-M	LPLO-O
clear	7	3	0	0	1	1	1	1	0
competent	21	5	0	1	3	3	1	0	8
direct	1	1	0	0	0	0	0	0	0
easy	7	3	1	0	1	1	0	0	1
fast	5	3	1	0	1	0	0	0	0
friendly	1	0	0	0	0	0	0	0	1
good amount of options	1	0	0	0	0	0	0	0	1
good options	1	0	0	0	0	1	0	0	0
good questions	4	0	0	0	1	0	0	1	2
good recommendations	19	9	1	1	1	3	0	2	2
good song library	1	0	0	0	0	0	0	0	1
impolite	1	0	1	0	0	0	0	0	0
natural	5	1	0	0	0	1	0	0	3
natural	5	1	0	0	0	1	0	0	3
preemptive	1	0	0	0	1	0	0	0	0
robotic	6	2	0	0	2	0	0	0	2
simple	2	1	0	0	0	0	0	0	1
ambigious questions	1	1	0	0	0	0	0	0	0
bad amount of options	2	0	0	0	0	0	0	0	2
bad questions	1	1	0	0	0	0	0	0	0
bad recommendations	4	2	0	1	0	0	0	0	1
bad song library	3	0	1	0	0	1	0	1	0
biased	1	0	0	0	0	0	0	0	1
boring	1	0	0	0	1	0	0	0	0
controlling	1	1	0	0	0	0	0	0	0
hasty	1	0	0	0	0	1	0	0	0
impatient	1	1	0	0	0	0	0	0	0
incompetent	3	0	0	0	1	2	0	0	0
incompetent regarding complex queries	2	0	0	0	1	1	0	0	0
incomprehensible	1	1	0	0	0	0	0	0	0
inquisitive	2	1	0	0	0	0	0	0	1
inquisitive	2	1	0	0	0	0	0	0	1
lost information	3	1	0	0	0	0	0	0	2
missunderstood	1	1	0	0	0	0	0	0	0
no undo	1	1	0	0	0	0	0	0	0
premature desicions	1	1	0	0	0	0	0	0	0
random	2	2	0	0	0	0	0	0	0
repetitive	4	1	0	0	0	2	0	1	0
single-tasking	1	0	0	0	0	0	0	0	1
slow	19	4	2	0	0	3	1	1	8
unambitious	1	0	0	0	1	0	0	0	0
uninformative about functions	1	0	0	0	1	0	0	0	0

A 1	I liked that the system offered me various possibilities to discover new things. (**good recommendations, HPHO-Match**)
A 2	Text (**Label, HPHO-Opposite**)
A 3	It was easy and clear (**easy, clear, LPHO-Match**)
A 4	Same as before. The dialogic situation makes it feel natural (**competent, natural, LPLO-Opposite**)
B 1	it had repetitive answers (**repetitive, HPHO-Match**)
B 2	same answers, kinda slow (**repetitive, slow, LPLO-Match**)
B 3	If there were 2 artists to chose from and you stated to like both, the system started with one artist and then asked to change the artist or genre first. It would be more natural to ask if you want to continue with the second artist. (**lost information, LPLO-Opposite**)
B 4	It was a bit boring and I wasn't sure if there were any functions that I didn't know about. (**boring, uninformative about functions, HPLO-Opposite**)
C 1	It would be helpful to get the possibility to answer a few general questions in the beginning, e.g. prefered music genres or what the playlist is going to be used for. (**HPHO-Match**)
C 2	Most likely not possible but maybe some future bot can also search songs requests through youtube idk (**HPHO-Opposite**)
C 3	Suggest more than 2 songs. Suggest more new songs; if I am telling it to put songs I already know in a playlist I don't really see the point. When suggesting new songs, pick songs that are more like the ones I like. Make it faster. (**LPLO-Opposite**)
C 4	asking for more hints or information on what I like/want instead of coming immediately with song recommendations (**LPHO-Match**)
D 1	Other than being able to "chat", none. (**HPHO-Match**)
D 2	It would need to understand things like "last song from an album" and "replace" (**HPLO-Opposite**)
D 3	It was easiser to use system 2, because it was clearer how the system works after the experience I had from system 1. (**HPLO-Opposite**)
D 4	The system once misinterpreted my answer, even if I like a song, it doesn't necessarily have to be in the list (**HPHO-Match**)

Table 16: Selection of free feedback with labels for A and B, and information on policy version and condition: A = Positive Feedback, B = Negative Feedback, C = Changes user would make, D = General/ Other Feedback.