Utrecht University

Master's thesis

# Using meta-conversations and user data to improve the naturalness and user experience of DialoGPT

Asun Alsina López (6646271)
Master's in Artificial Intelligence

*Supervisors:*

Dr. Baptist Liefooghe - Utrecht University
Dr. Mathias Funk - Eindhoven University of Technology
Dr. Supraja Sankaran - Eindhoven University of Technology

June 1, 2021

**Abstract**

Negative feelings such as stress and anxiety are common in contemporary society, but not everybody can afford to get help from a professional. Several mental health coaching chatbots have appeared in the last few years, but they lack naturalness in conversation and understanding of context. On the other hand, general-purpose chatbots are not prepared to help users to navigate their negative feelings. Adding scripted conversations about feelings to general-purpose chatbots can create support chatbots for users looking to have a small talk with some friend to vent about their worries. Adding a meta-conversation (talking about the related conversation) about the emotions dialogue will give the users a deeper understanding of the workings of the chatbot and the usage of their data to personalise their experience. Meta-conversations are a great way to deepen the knowledge of a person about the discussed topic and the intention of their conversational partner, improving their relationship.

Prolonged exposure to the chatbot is necessary in order to gather enough data to personalise the user experience to test this chatbot. Therefore, 15 people participated in an experiment that lasted two weeks. After those two weeks, the participants filled out a questionnaire, and six of them participated in a subsequent interview. The results showed that the chatbot being incoherent complicates the interaction between the chatbot and the users.

# Contents

**Abstract** **1**

**1 Introduction** **4**

**2 Literature review** **6**

2.1 Conversational agents . . . 6

2.2 Chatbots . . . 7

2.3 Online therapy . . . 8

2.4 Meta-conversations . . . 10

**3 System architecture** **11**

3.1 Classifier . . . 12

3.2 User profile . . . 13

3.3 Reminder . . . 15

3.4 Response generator . . . 15

3.4.1 Malfunction conversation . . . 15

3.4.2 User data conversation . . . 18

3.4.3 Feelings conversation . . . 22

**4 Experiment** **28**

4.1 Participants . . . 28

4.2 Materials . . . 28

4.3 Procedure . . . 29

4.4 Measurements . . . 30

**5 Results** **31**

5.1 System metrics . . . 31

5.2 Questionnaire . . . 32

5.3 Interview . . . 35

**6 Discussion** **38**

**7 Conclusion and future work** **43**

**References** **44**

**A Questionnaire** **47**

A.1 Statement of informed consent . . . 47

A.2 Survey . . . 48

**B Interview** **50**

# 1 Introduction

Dialogue systems or conversational agents have been around since the early sixties. Their uses range from entertainment (e.g. Cleverbot[1]) to practical purposes, such as testing theories of psychological counselling (e.g. ELIZA [26]). Studying the human mind was one of the applications of this technology that has been present since the beginning of this field. ELIZA was the first chatbot developed for this use.

Online therapy is a way of using the internet to provide mental health services via several channels (messaging, emails, video conferences, etc.). With the development of messaging technologies and the consequent possibility of access to these technologies, more people are reaching therapists and doing so online. But even with these new options, there are still some people that cannot afford online therapy and here is where therapy chatbots come into play (e.g. Woebot[2] or Wysa[3]). These chatbots not only offer therapy to people that cannot afford it, but they also present the possibility of just in time adaptive interventions (JITAIs), designed to provide support at the right moment and adapt to the rapidly changing states of the individual [14].

These chatbots, expressly created for therapy, seem to follow a predetermined script which is impossible to avoid. These scripted conversations diminish the naturalness of the discussion and increase the impression that the chatbot does not comprehend the users and continues the conversation without considering what they are saying, which can lead to people not using it anymore.

On the other hand, there are plenty of chatbots that are not programmed to have a conversation about a specific topic or a scripted conversation, which allows more freedom to the users but does not ensure the naturalness of the dialogue or that the utterances make sense (e.g. DialoGPT [27]). These chatbots, as many artificial intelligence models, suffer from the black-box problem. The black-box analogy arises because the nested non-linear structure of these models does not provide enough information to clarify what is making them predict that outcome [20]. The field of explainable artificial intelligence (XAI) was created to tackle this problem. Their goal is

[1]https://www.cleverbot.com/
[2]https://woebothealth.com/
[3]https://www.wysa.io/

to make AI models understandable to everyone, and for that, the models need transparency, interpretability and explainability [18]. Even if these models reach this goal, they might be too complex for end-users without a background in a technology-related field.

Meta-conversations is the term chosen in this project for those conversations that address the foundations of the conversation. That is the conversation about the conversation. These dialogues are typical in human-human interactions since they are a way to give and obtain feedback about the discussion [6], letting the conversational partners focus on the emotions involved, their relationship, the course of the conversation, etc. All these aspects that meta-conversations provide can be used to analyse the conversational behaviour of the participants and improve the quality of the conversation. Although humans use meta-language extensively, the frequency and conditions of the application of this resource are not known. In [3], they conduct a corpus study in order to get some insight into the use of meta-language, looking towards being able to implement meta-language in natural-language Human-Computer Interaction (HCI) systems.

Currently, therapy and general-purpose chatbots cannot deal with meta-conversations in the same way humans do, thus, losing valuable information. Actively implementing meta-conversations can be a way to provide the users with some explanation about the workings of the chatbot, giving insight into its behaviour. The chatbot would use the data provided by the user or the meta-data gathered during the conversations, such as time or date. These conversations will help users understand the reasons behind the utterance of the chatbot without needing an explicit knowledge of the workings of the model. Therefore, they will provide some transparency, interpretability and explainability to the chatbot as XAI wants to achieve.

In this project, we add meta-conversations about the users' feelings and conversations about feelings to a general-purpose dialogue generator, DialoGPT, to improve the naturalness of the system and improve the overall quality of the conversation. To add these conversations, we implemented a rule-based bot that takes control of the conversation when the users talk about their feelings or ask about how the chatbot uses their data. With this hybrid design, general-purpose plus rule-based response generators, we aim at giving the users more freedom in their interactions while keeping control in critical settings, such as emotion or personal data conversations.

# 2 Literature review

In this chapter, we examined the existing literature to create a clear picture of the current knowledge about the topic at hand.

## 2.1 Conversational agents

Conversational agents are programs designed to communicate with the users through natural language (written, spoken or both). Since these systems use natural language, they are subject to the same features that define human conversations, and developers need to consider them while creating the system since failing to meet these characteristics could ruin the interaction. Some of these features are turn-taking, grounding and conversation control or initiative.

In human-human conversations, people take turns to talk, which can range from a word to several sentences. In both spoken and written conversations (in messaging applications), people can predict when their conversational partner is going to end their turn. Several researchers studied how people anticipate the end of a turn ( [17], [11]). In written conversation, people usually use one message per turn or several of them per turn. While humans have no problem adapting to this, it is not the same for conversational agents that generally can only reply to one message at a time.

Another feature of human dialogues is grounding, the agreements the participants in a conversation reach, meaning the common ground [22].

Usually, in human interactions, the conversation control is shared among the participants, known as mixed-initiative [25]. While the mixed-initiative is quite common among humans, there is also the possibility that one of the participants controls the conversation, a professor giving a lecture has the initiative in that dialogue with the students. In human-system interactions, it is more common that only one side takes the lead on the conversation. These interactions can be user-initiative or system-initiative.

With the evolution of dialogue systems, it has been clear that adding the different features present in human-human dialogues is not an easy task, but in recent years, notable advances have been achieved. These features are not subject to a particular kind of conversational agent both task-oriented systems and chatbots [12] need to consider them. As their name indicates, task-oriented systems focus on solving a specific task (e.g. booking a flight). An example of task-oriented conversational agents is digital assistants, such as Siri or Alexa. The following section will be about the other category of conversational agents, chatbots.

## 2.2 Chatbots

Chatbots are systems designed to have long conversations with users. These dialogues try to mimic human-human discussions by using the unstructured conversational characteristics of these conversations. As opposed to the task-based ones, these systems do not focus on a particular task, but they can focus on a topic. Chatbots, in turn, can be divided into rule-based or corpus-based systems.

Rule-based chatbots have a predefined set of rules that control the conversation, and the input of the users has to conform to these rules. To simplify the user interaction and avoid some of the problems that can arise from formatting the sentences to comply with the predefined set of rules, often the users interact with the chatbot through buttons. One of the issues of these chatbots is that open conversations are either very limited or not possible at all.

There are several well-known rule-based chatbots, such as ELIZA [26] or PARRY [8]. ELIZA, the first chatbot, tried to imitate a Rogerian psychologist, a branch of psychology where the goal is to draw the patient out by reflecting the statement of the patients back at them. Therefore, it is a chatbot that does not need to know anything about the real world since it will reflect the statement of the users back to them. An example of ELIZA's workings would be "Well my boyfriend made me come here", and ELIZA would reply: "Your boyfriend made you come here", these two sentences are part of a sample dialogue included in 34.

Corpus-based chatbots, as their name implies, need a corpus to train on it and generate an appropriate response for the utterance of the users. The most common architectures, in the beginning, were information retrieval (IR) and sequence to sequence models, but nowadays, with the rise of artificial intelligence (AI), most models use a neural network to generate the responses (e.g. the neural language model GPT-2 [16]). The biggest problem of corpus-based systems is that they need a large amount of data on the topic we want to train the chatbot.

A famous example of a chatbot that uses artificial intelligence to create the conversation is Cleverbot[4]. Cleverbot learns from its past interactions with humans and responds to them based on those past conversations. As opposed to ELIZA, Cleverbot does not respond in a fixed way. Instead, it uses fuzzy logic to choose its responses heuristically.

There is a third option where the two previous ones come together to create a hybrid model, like the one used in this project. The hybrid architectures use features of both the rule-based and the corpus-based chatbots. This architecture allows researchers and developers to avoid or lessen the problems of the other two architectures. A rule-based model can be used with a neural network model when the data are not enough to train the neural network, and at the same time, the neural network gives more flexibility to the rigid structure of a rule-based bot.

A state-of-the-art example of a hybrid model is Woebot[5]. A chatbot that using AI and cognitive behavioural therapy (CBT) [5] helps people with their mental health. It is not only for adults but also for teenagers, people with addictions and pregnant women. As they state on their web page, they created this app to help people reduce their stress, discover patterns in their moods, and live happier.

## 2.3 Online therapy

As seen in the previous section, one of the fields where chatbots were implemented is the mental health field. Before chatbots, the first leap in this field was from in-person therapy to online therapy or e-therapy. The accessibility and affordability of this kind of therapy make its advantages evident, but it does not talk about its efficacy. Thus, several research articles that study the effectiveness of this kind of therapy have been written in the past years, showing their results that it is effective [23], [19], [4].

[4]https://www.cleverbot.com/

[5]https://woebothealth.com/

Mental health chatbots present the advantage of just-in-time adaptive interventions (JITAIs) at any time needed. JITAIs are an intervention design to provide people with the right amount and type of support at the right time by adapting to their state at that precise moment [14]. Since this kind of interventions focuses on providing support at the right moment using technology for that, chatbots are a suitable tool to implement them now that they are available the whole day. On some occasions, people might need to talk with a therapist in the early hours of the day, probably the therapist will not be available at that time, but a chatbot will be, which allows the users to get help to improve the mood that made them contact the chatbot. Just-in-time adaptive interventions have been used in different fields such as mental illnesses [7], changing sedentary behaviour [13], or quitting smoking [15].

These chatbots have proved their effectiveness, although their results are different from online therapy or in-person therapy. In [10], they researched the effects of using Woebot to deal with anxiety and depression. In this study, 70 students aged between 18 and 28 had 20 sessions with the chatbot for two weeks. The results showed a significant decrease in the anxiety and depression feelings of the non-control group compared to the control group.

Woebot is not the only therapy chatbot out there. Other examples include Moodnotes[6], which analyses the mood of the users and asks questions if it has a negative valence to determine their pattern of thought, or Wysa[7], which helps the users to manage their emotions and thoughts using different techniques for that, such as CBT, Dialectical Behaviour Therapy (DBT) [9], meditation or yoga. The field of therapy chatbots is a thriving one that will continue to grow with the advances in the field of conversational agents and with the evolution of the different technologies implied, such as natural language understanding (NLU), natural language processing (NLP) or speech recognition, among other things.

[6]http://www.thriveport.com/products/moodnotes/
[7]https://www.wysa.io/

## 2.4 Meta-conversations

Chatbots using deep learning for natural language generation are powerful, but they suffer from the black-box problem. Deep learning models have often been considered black-boxes because their nested non-linear structure does not provide enough information to clarify what is making them predict that outcome [20]. The more complex the deep learning models become, the more difficult it is to figure out how data are treated to produce that prediction. In the case of chatbots, this output is the chatbot sentence, which can come across as strange or peculiar. These strange sentences might create in the user the necessity to know why the chatbot said that.

In the past few years, with the rise of this kind of models, there has also appeared the field of explainable artificial intelligence (XAI) that aims at erasing the term black-box, so all artificial intelligence models are understandable by humans. According to [18], there are three core elements that these models need: transparency, interpretability, and explainability.

Meta-conversations can be a good step forward to clear up the mystery behind the black-box problem. These conversations offer the users information about the usage of their data and the reasons of the system to say what it said in easily understandable terms. Through meta-conversations, end-users receive an explanation about the use of their data. Those explanations can contain data that they have to provide the system and meta-data (such as date or time) that the system has gathered through the conversations with the users.

Although meta-conversations can be advantageous to clarify the black-box problem, many chatbots lack a smooth transition between the related conversation, the discussion that deals with the topic addressed (e.g. emotions, weather, etc.), and the meta-conversation, the conversation about the conversation. These conversations about the conversation put both conversational partners on the same side, facilitating reaching a common ground [22]. Meta-conversations are not only a way to keep a conversation going; they help to improve the relationship of the conversational partners as well because they provide feedback about the discussion [6].

# 3 System architecture

In this project, we developed a chatbot with a focused conversation on the user emotions and a meta-conversation about the focused one. This project intends to check whether these conversations and their personalisation improve the user experience and whether it helps the users with their emotions. The source code of this chatbot can be found in a Github repository[8].

This chatbot runs on Telegram as a Telegram bot allowing the users to access the conversation at any time. This chatbot does not intend to act as a therapist, but as a friend that is there for you at any time you need it which is in line with JITAIs (just-in-time adaptive interventions), in the sense that its pieces of advice and its conversation will act in the precise moment the users need help. The users do not have to wait until daytime to talk to the bot if something is troubling them during the early morning hours when most people are probably still sleeping.

Some Telegram bots work on individual and group conversations. However, this is not the case with this chatbot whose design is specific for private discussions.

To include the meta-conversation and personalise both conversations (related and meta), the system showed in figure 1 was built. This system comprises several modules. There are two modules to generate the response of the chatbot. One of these modules is DialoGPT [27], and the other one is a rule-based bot. The user utterance goes through a classifier to choose the response generator. Besides these modules, there is another one to create the user profile and one for the reminders.

[8]https://github.com/asunalsina/MetaChat

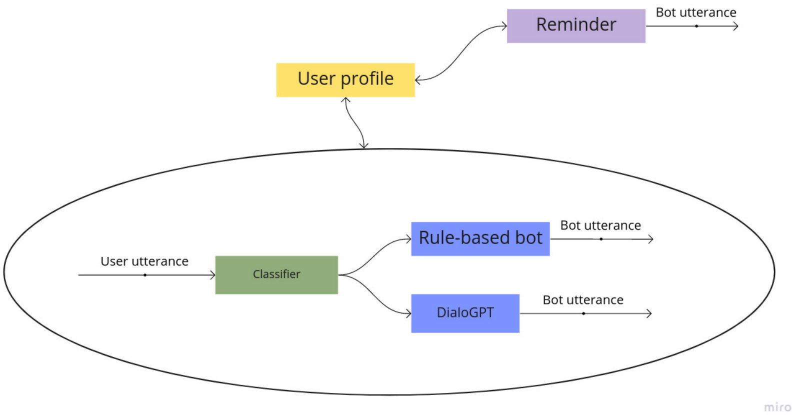



Figure 1: *System architecture*

## 3.1 Classifier

The classification module decides which response generator to call after analysing the user utterance. The sentence is analysed using Facebook's natural language interface Wit.ai[9]. This interface examines the user sentence and returns the category to which it belongs. The different intents are *feelings* and *chatbot workings*, and each one has various entities. The utterances can belong to a third category that indicates that the sentence is out of scope and therefore, the response generator should be DialoGPT.

The chatbot workings conversation has two entities, one of them to report a malfunction and the other to inquire about the user data. The feelings conversation has four entities, one for every quadrant of the activation-valence pair, figure 2.

Most of the data used to train the classifier came from the Dreaddit dataset [24]. This dataset uses Reddit data from different communities and classifies it in five domains (abuse, anxiety, financial, PTSD and social). The data from the anxiety domain is the one used to train the classifier. These data come from the anxiety and stress communities.

Since Reddit is based in a post and comments dynamic, not all the sentences could be used to train the classifier about negative feelings, because these sentences made sense

[9]https://wit.ai/

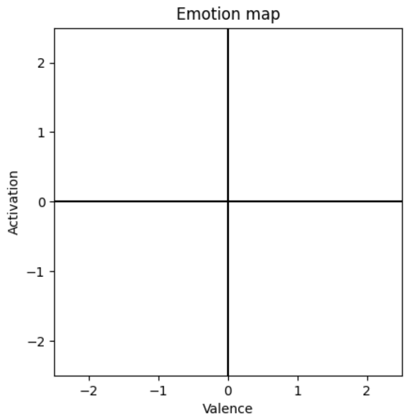



Figure 2: *Activation-valence space*

in the context of the whole post but once they were on their own, either they did not make sense, or they did not express negative feelings anymore. Some of these sentences became part of the "out of scope" utterances.

Besides the Dreaddit dataset, the rest of the sentences come from two other sources. One of them is the sentences inputted while developing the chatbot. And the other one is the Emotions dataset for NLP classification tasks [1]. Almost all the utterances of this dataset were discarded because of the incorrect labelling of several of them.

## 3.2 User profile

The profile builder module registers users data for the rule-based response generator to use. To store this information, the NoSQL database MongoDB is used. This document-oriented database is non-relational, meaning that it is more flexible and adaptable than relational databases.

The user profile contains twelve fields: *chat_id, phase, reminders, conversation, emotion_map, last_conversation, last_reminder, last_quadrant, last_activity, last_utterance, active_conversation and quadrant_time.*

The chat_id field identifies the users, and the phase field shows in which stage are the users. In the reminders field, the users indicate whether they want to have daily reminders or not have reminders. The conversation, both user and bot utterances, is saved in the user profile as well.

The emotion map is a table with five columns (date, day, time, valence, activation, and activity). Table 1 shows an example of an emotion map.The information recorded in the emotion map comes from the feelings conversation. In the second phase, the bot schedules the change of the activities of the emotion map. Since the users can write a whole sentence when the bot asks them about the exercise they feel like doing, it is necessary to clean the sentence and leave only the hobby. After cleaning the sentence of common stop words, the semantic network ConcepNet [21] is used to compare the terms. This knowledge graph connects words and phrases with labelled edges and provides machines with a better understanding of the meaning of the words people use. It also provides the relatedness between two words. Therefore, the relatedness of the user sentence is compared word by word with the words in the hobbies database. The one with the greatest relatedness, above a certain threshold, will replace the exercise in the activity column of the emotion map. A random hobby will be chosen for that row of the emotion map if none of the relatedness values is above the threshold.

The *last* fields record the latest state of several variables. These values are part of the answers to some of the meta-level questions. The *last activity* field includes the time and day of the week besides the exercise. The active conversation field points to the latest feelings conversation activated. It acts as a memory since the previous turn intent, the one triggering the dialogue, is not stored in the system memory.

The bot needs to know which quadrants are predominant in each time frame to send the reminders. The quadrant time field stores that information. It uses the emotion map to see the number of times each quadrant repeats during each time frame and then selects the one with more repetitions. The fields will have the keyword *none* recorded when there are no data or when the maximum repetitions do not include the second or third quadrant.

| Date | Day | Time | Valence | Activation | Activity |
|---|---|---|---|---|---|
| Tuesday | 2020-12-08 | Afternoon | -1 | 1 | Running |
| Tuesday | 2020-12-08 | Morning | -2 | -1 | Dancing |
| Wednesday | 2020-12-09 | Morning | -1 | 1 | Running |

Table 1: Emotion map example

## 3.3 Reminder

The reminder module makes the bot take an active role in the conversation. This module calculates the quadrants that repeat the most during each time frame using the emotion map. Then, the bot uses that information to schedule a message for the user. The sentence is the same for all the users: *Remember to take a break!*.

The chatbot sends the message when the quadrant is the second or third one because those have negative valence. When there is more than one time frame available, the time frame is chosen randomly and converted to hours. These hours are restricted to avoid disturbing the user during the night. The bot can send a reminder between nine in the morning and ten in the evening.

## 3.4 Response generator

The rule-based response generator module takes control of the conversation when the classification module indicates it. The chatbot consists of two phases with several dialogues each, being the difference between the stages the personalisation of the conversation. The first phase is to gather data, and the second one is to recommend some activity based on the users' feelings.

The first step to select the conversation is the intent returned by Wit.ai, and the second step is the entity. In the first phase, there are three possible dialogues, and in the second one, five. The possible intents are the same in both stages, dividing the sentences into *feelings* or *chatbot workings*.

Since it is a rule-based response generator, the users have little freedom to answer the questions, and most of the time, they have to choose the answers from buttons.

### 3.4.1 Malfunction conversation

When a sentence falls into the category chatbot workings, there can be two possible dialogues: malfunction and user data. It is necessary to look at the entity Wit.ai returns after classifying the utterance to choose the appropriate conversation. The malfunction conversation, figure 3, has three stages, and its purpose is to allow the users to complain about a problem they have with the chatbot.

The users choose from a list of possible errors (repetitive, unclear, disrespectful, and illogical) what they want to report. When they select one of these categories, the chatbot will ask them to describe how is the chatbot behaving the way they said. However, if the users think that their complaint does not fit in one of the categories offered, they can choose to report other malfunction, and then the chatbot will ask them to explain what problem they are encountering.

The malfunction conversation is the only one that remains the same during the two phases the chatbot has.

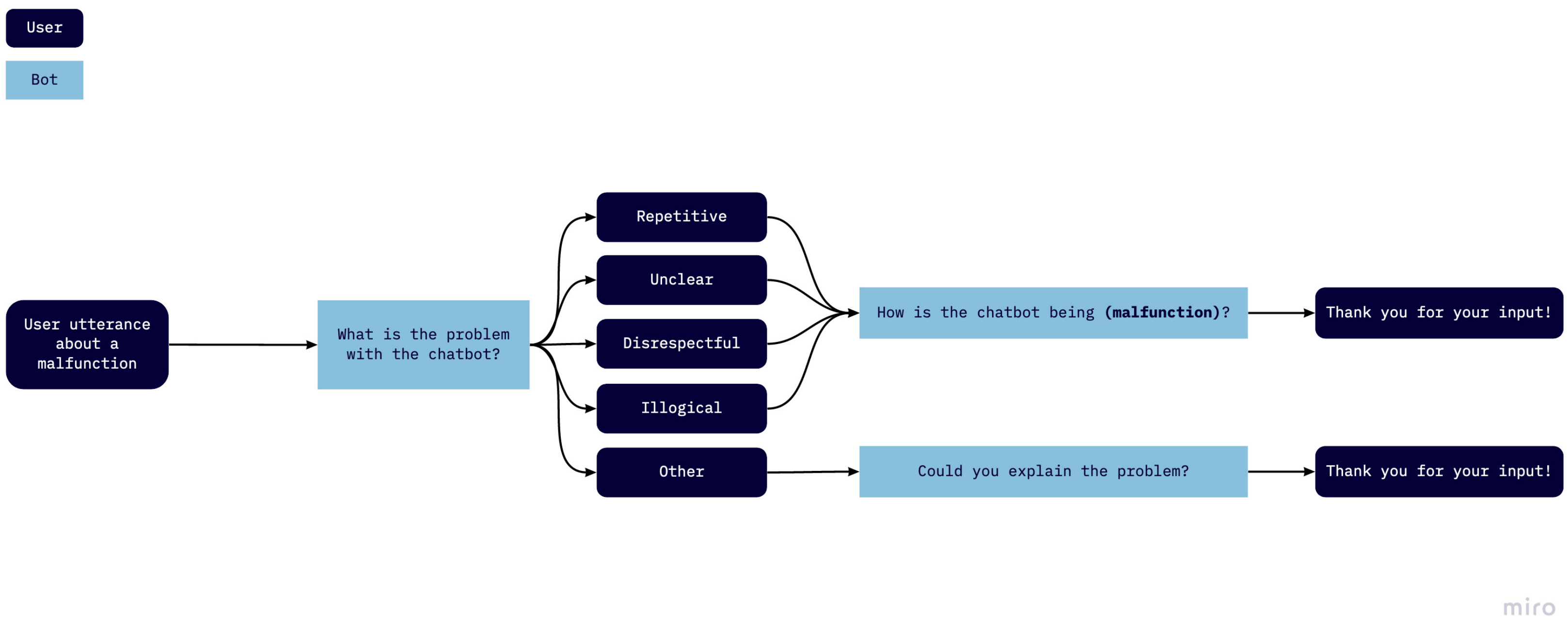



Figure 3: *Conversation flow of the malfunction conversation*

### 3.4.2 User data conversation

The user data conversation like the malfunction conversation activates when the sentence falls into the chatbot workings category, and the entity indicates the wish of the users to talk about their data. In the second phase, the conversation changes to incorporate the data on the user profile. This conversation is a combination of a user profile with a frequently asked questions section.

#### First phase

Figure 4 shows the conversation flow of the user data conversation in the first phase. This conversation has three stages. In the second one, the users can choose between three options; they can delete their data or ask about how the bot uses their data or why it is asking some questions. In this first phase, the responses are generic because the user profile has little or no data.

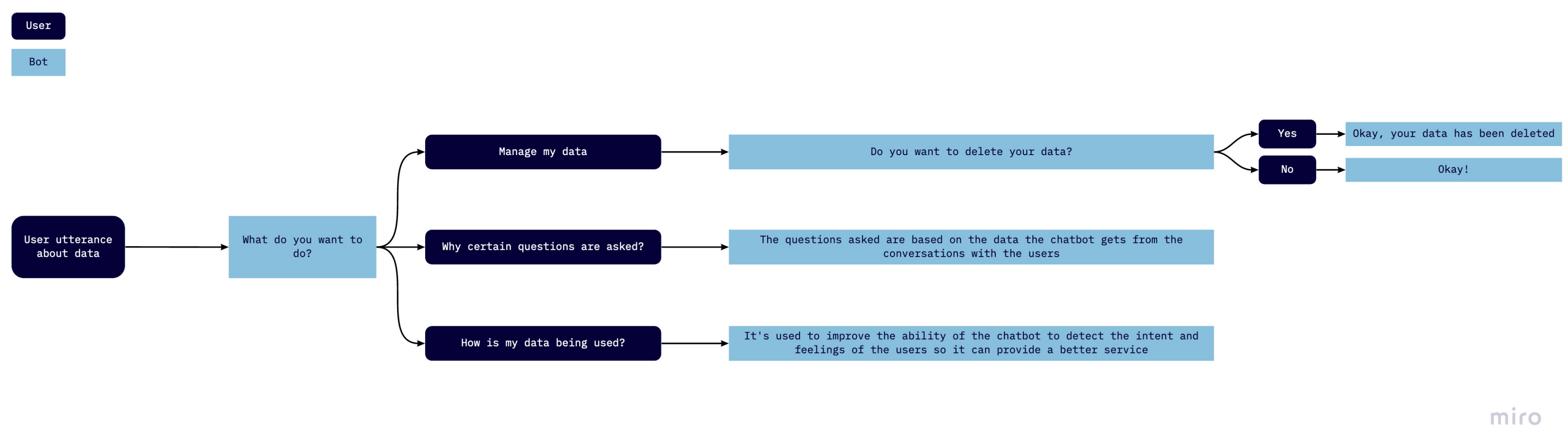



Figure 4: *Conversation flow of the user conversation in the first phase*

### Second phase

In the second phase, the user conversation, figure 5, has now four stages. The user profile part has a new option that allows the users to manage their reminders. They can choose between daily reminders or not receiving any reminders. This option appears in this phase because, in the first phase, the bot does not have enough data to create personalised reminders for the users.

The frequently asked questions part of the conversation is the section that deals with the meta-level since it answers questions about the conversation development. Of the two questions of the previous phase, only one of them remains (*How is my data being used?*) and has the same answer as before. Four more questions are part of the meta-level, and these depend on the user data.

The question *Why did you suggest that activity?* always appears as an option, independently of the last conversation activated. The response to this question depends on whether the user profile has the last activity suggested registered on it. If this piece of information exists, then the reply will include how long has it been since the recommendation, the day of the week, the time of the day, the energy, and the activity. An example response would be: I suggested **running** because **last Tuesday evening** you also had **high** energy and wanted to do that activity, so I thought it could help today.

The other three questions only appear if the required information is on the user profile, given that the response uses data from those conversations. These questions depend on the last feelings conversation the users had and the reminders.

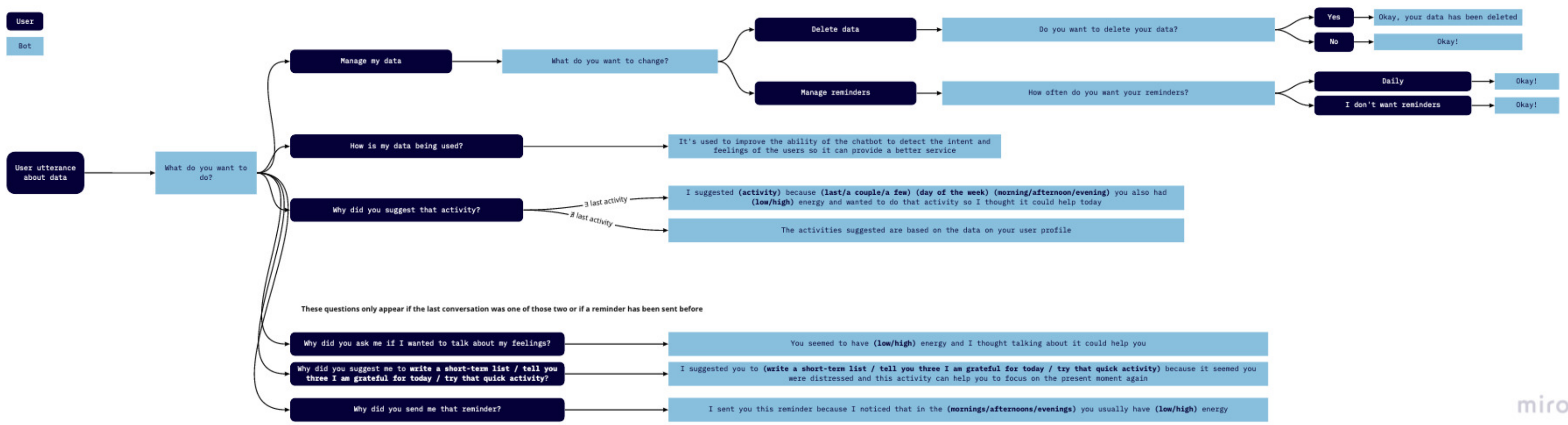



Figure 5: *Conversation flow of the user conversation in the second phase*

### 3.4.3 Feelings conversation

The feelings conversation has one dialogue in the first phase, whereas, the last stage consists of three different conversations.

#### First phase

Figure 6 shows the conversation flow of the dialogue about feelings. This conversation, made up of six stages, activates when the user sentence falls into the feelings category. It is not relevant whether the user sentence expresses positive or negative emotions; both trigger this dialogue.

When this conversation starts, if the users choose to continue with it, they have to specify the valence and the activation of their feelings. After that, they need to tell the chatbot an exercise they like to do when feeling that way. There is a filter in this last step to prevent the chatbot from recommending activities like sleeping in further stages. Therefore, when the users input one of the activities on the filter, they will have to choose an exercise from a list offered by the chatbot. This list comes from the hobbies database.

All the values that the users introduce get saved in the user profile. The valence and the activation form the quadrant in which the emotion fits.

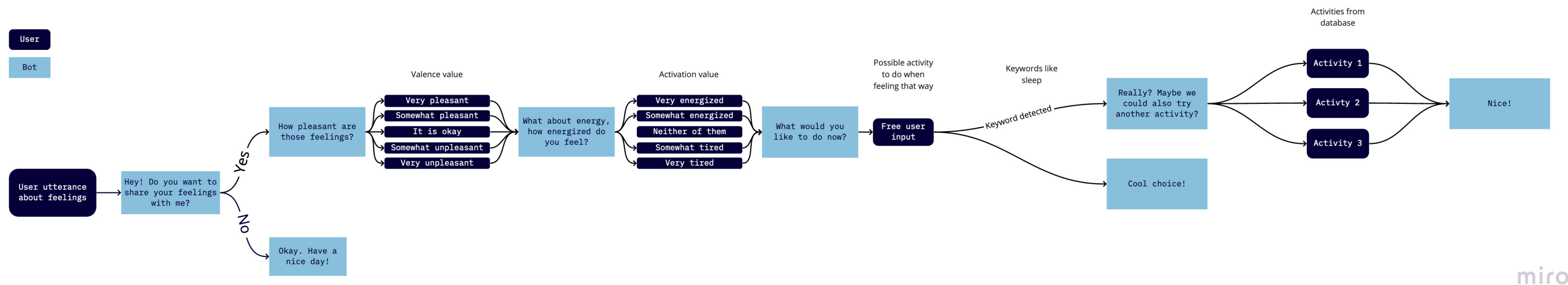



Figure 6: *Conversation flow of the feelings conversation in the first phase*

### Second phase

The feelings conversation splits into three different dialogues in this phase. One of the conversations suggests activities to reconnect with the present, another one asks the users to talk about how they are feeling, and the third one recommends an exercise based on their profile and their emotions in that precise moment.

In the previous phase, the conversation started when the sentence fell into the feelings category regardless of the quadrant. Now, the entities come into play, and the dialogue is only triggered when the user emotion is on the second or third quadrant, figure 2. These quadrants are related to negative emotions since the valence of those quadrants is negative.

The conversation to talk about the feelings has four turns, figure 8. There are at least a couple of reply options for the chatbot in each stage. In each conversation, the chatbot will choose a response randomly except in the third turn, where one of the options is not available when the users use a temporal measure in their previous reply.

The conversation to reconnect with the present has three stages, figure 7. The bot chooses one of the three possible exercises randomly. There is one activity for the users to become aware of their bodies, another to write a short-term list and the last one where the users tell the bot three things for which they are grateful.

The conversation to suggest activities has six turns, figure 9. In the first turn, the bot adapts the sentence to the quadrant in which the users are. If the users say that the detected emotion is erroneous, then instead of getting a recommendation, the bot asks the users to answer the same questions as in the first phase. Making the users answer these questions increases the data in their profiles.

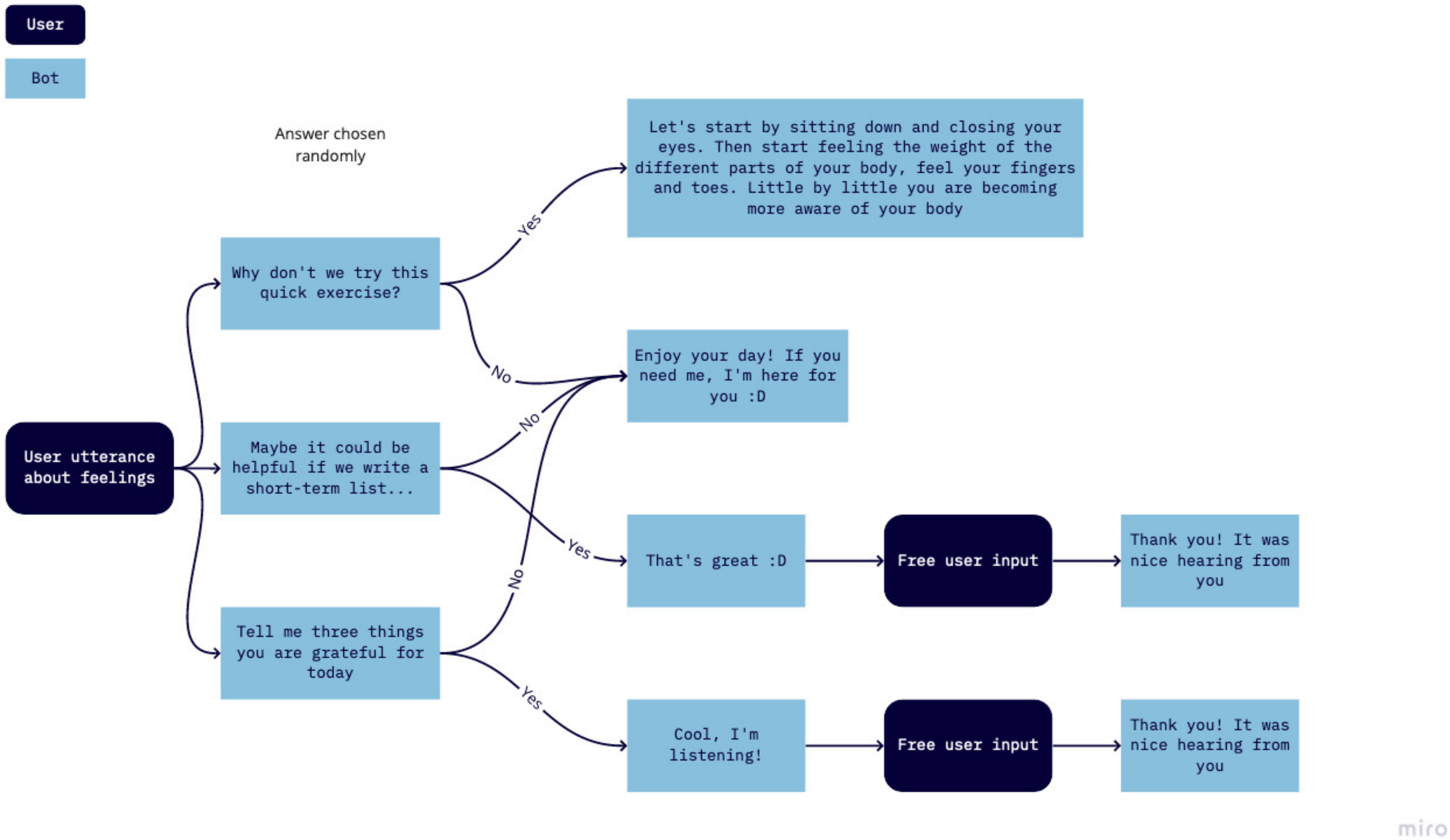



Figure 7: *Conversation flow of the reconnect with the present conversation in the second phase*

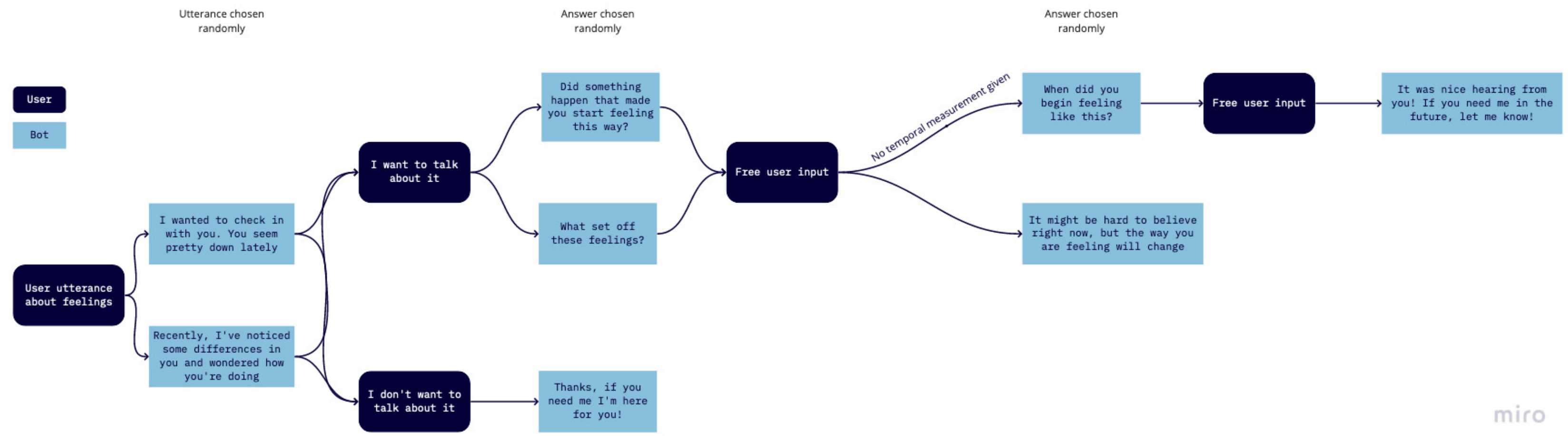



Figure 8: *Conversation flow of the talk about what is happening conversation in the second phase*

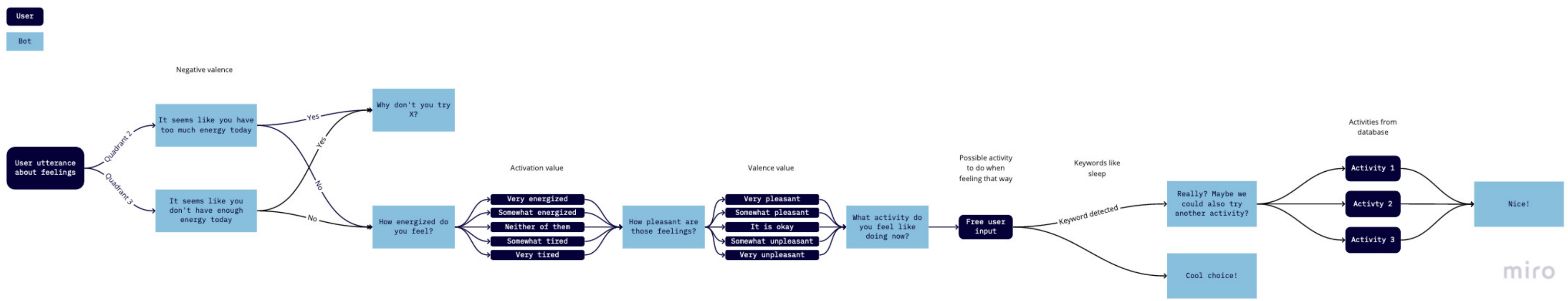



Figure 9: *Conversation flow of the suggest activity conversation in the second phase*

# 4 Experiment

Since the purpose of this project is to address the first stage of a design more than a finalised product, the aim is to gather the perspectives of the users about the direction chosen to develop this design and to check whether this is the right direction to continue or it is better to change this initial design.

## 4.1 Participants

For the current research, 15 participants in a range of age between 23 and 28 years ($M_{age} = 26, SD_{age} = 1.58$) were assessed using an online survey, and six of them participated in a subsequent interview. The type of sampling used to select the participants was convenience sampling. The bias that might be associated with this type of sampling has been considered as a limitation of the experiment.

The participants had to meet the criteria to be deemed suitable to participate in the experiment. They had to be of legal age, have a telegram account and an advanced level of English.

## 4.2 Materials

In this study, participants were asked to interact with the chatbot for two weeks. They had to chat with it for at least a few minutes per day, although they could use it as much as they wanted. The first week they interacted with the chatbot in its first phase, the data-gathering one. After the first week, they continued with the second phase, the recommendation phase, although there is also data gathering during the second phase. Once these two weeks finished, the participants had to fill in a questionnaire developed in Qualtrics, measuring their experience talking with the chatbot. Some of the participants were contacted after filling out the questionnaire to participate in an interview to deepen their responses to some questions.

The survey comprised questions from a standardised questionnaire [2] (e.g. *The conversation with the chatbot is predictable*) and specific questions about the chatbot workings (e.g. *I would follow the chatbot advice*). Both standardised and the other questions use a 5-point Likert scale, with 1 indicating "Strongly disagree" and 5 indicating "Strongly agree". Some of the items were negatively scored to prevent the participants from filling out the survey without focusing on the content of the questions. Besides these questions, there are some open-ended questions after some of the 5-point Likert scale ones so the users can add more information about their response to the Likert scale question. Appendix A and B include the questionnaire and the questions asked during the interview, respectively.

## 4.3 Procedure

Before starting the conversation, the participants had to read and fill out the informed consent, which can be found in Appendix A. Then, they received an introductory text where they were informed on what to expect in this experiment. The instructions they received were the following:

*This experiment utilises a general-purpose chatbot focused on user feelings (managed by a rule-based bot). You can speak with the chatbot about anything you want, although you should talk with the bot about your feelings, you can be as personal as you wish.*
*There are two phases, and I will notify you when you change stages. There are three different conversations in the rule-based bot: one to report a malfunction, one to talk about your emotions, and another one to manage/ask about your data. You need to use the feelings and manage data conversations a few times. The feelings dialogue even once a day if possible.*
*You will have to use the bot for about two weeks and after that answer some questions. Try at least to use the bot once a day at any time, but you can use it more time per day if you want.*
*Your data will be collected, stored and anonymised, but you can delete it at any moment through the chatbot or contact me to delete it. You can withdraw from the experiment at any moment.*
*If the chatbot gets stuck on the conversation, you can use the keyword 'bye' to reset the conversation.*

## 4.4 Measurements

There are two main measurements in this experiment the ones provided by the participants after filling out the survey and answering the questions of the interview, and the precision of the classifier used by the chatbot. In the questionnaire, the first five questions come from the user experience questionnaire (UEQ). The other questions ask about the naturalness of the system, the usability, the personalisation, the reminders and the user satisfaction with the advice the chatbot gives. The precision is calculated as follow:

$$\frac{true\,positive}{true\,positive + false\,positive} \tag{1}$$

Where:

- true positive is the number of times the sentence was classified as feelings and it was correct
- false positive is the number of times the sentence was classified as feelings but it was an incorrect classification

# 5 Results

The results obtained come from the questionnaire filled out (appendix A), the interview questions (appendix B), and the data obtained from the interaction of the users with the chatbot.

## 5.1 System metrics

The precision of the system was calculated with the data obtained from the interactions with the chatbot. Equation 2 shows the formula used to calculate the average precision of the system and the precision for each user.

$$\frac{true\,positive}{true\,positive + false\,positive} \tag{2}$$

The average precision of the system was lower than 0.5 and was 0.36. Figure 10 shows the precision of the classifier for each user. Of the 15 users, only five had a precision greater or equal to 0.5. The maximum precision achieved was 0.71, and the minimum was 0.09. Only four users had a precision that was above chance.

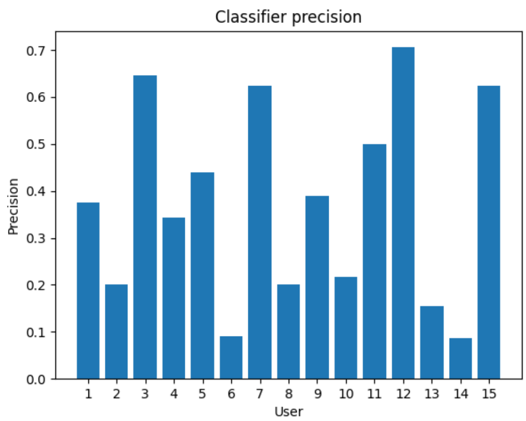
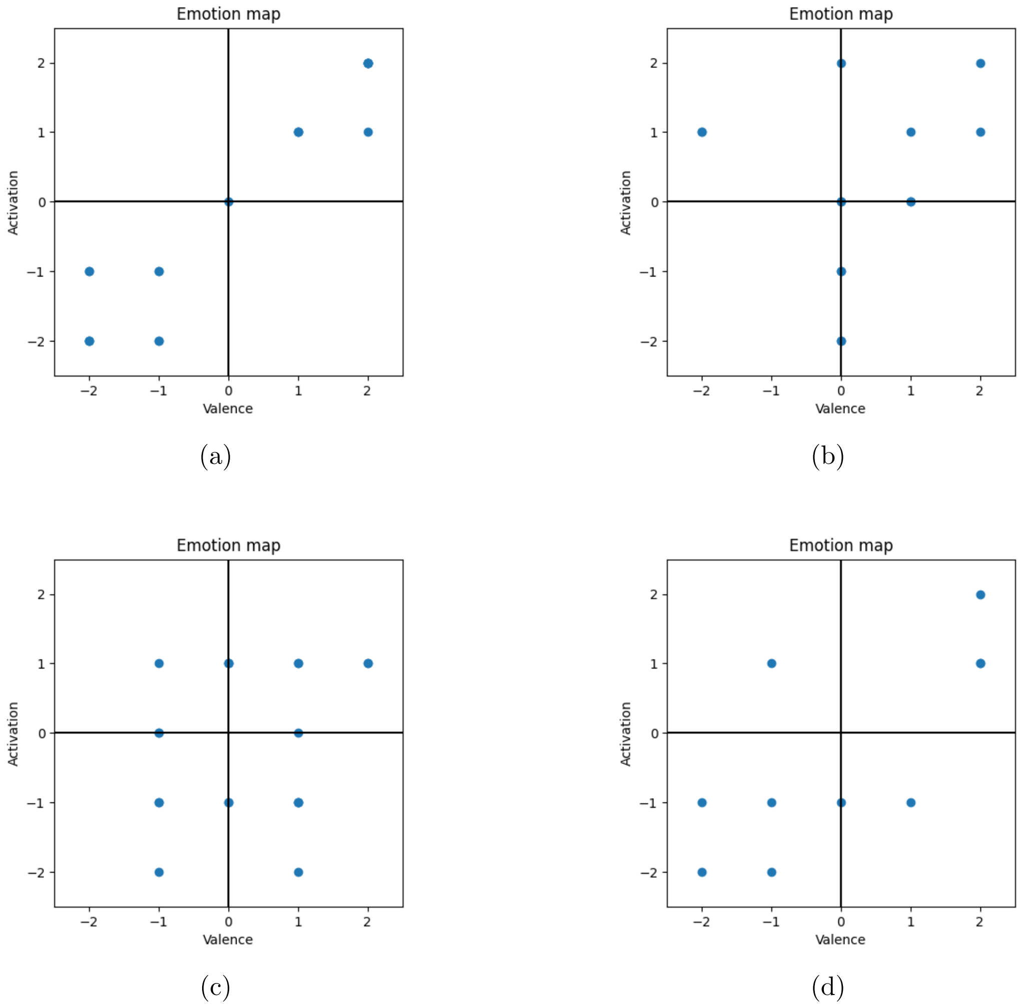



Figure 10: *System precision*

The emotion map of each user was calculated while they talked with the chatbot. Figure 11 shows several emotion maps. Since these graphs are two-dimensional, there can be several points in the same spot because the same activation-valence pair might have happened at different moments (either different time of the day or different day).




(a) (b) (c) (d)

Figure 11: *Emotion maps*

## 5.2 Questionnaire

The survey included a few demographic questions and one about the daily time of use besides the 5-point Likert scale and open-ended questions. The demographic questions were about the age of the participants and their native language. Most participants (14 out of 15) are not natives English speakers. The question about the daily time of use showed that the average usage was of ten minutes per day.

The average value of the 5-point Likert scale questions was calculated, table 2 shows the results. The questions that have an asterisk are questions scored negatively to prevent the participants from filling out the questionnaire without focusing on the questions. A value of 3 in the 5-points Likert scale means neither agree nor disagree. An average value lower or greater than 3 indicates some disagreement or agreement, respectively, with the fact stated in the question. Table 2 shows that almost all values are close to the mean value ("neither agree nor disagree"). A one-sample t-test was conducted for each Likert scale question. The results showed that only the average value of question 7 ($M = 3.67$, $SD = 1.01$) and of question 8 ($M = 3.73$, $SD = 1$) significantly differed from a value of three. Question 7 $t(14) = 2.47$, $p < .05$ and question 8 $t(14) = 2.75$, $p < .05$.

| Question | Average |
|---|---|
| 6. The chatbot is annoying | 3.4* |
| 7. The conversation with the chatbot is predictable | 3.67* |
| 8. The chatbot is friendly | 3.73 |
| 9. This chatbot is unlikable | 3.4* |
| 10. The language and conversations feel natural | 2.8 |
| 11. The buttons are bothersome | 3.47* |
| 12. The chatbot enables me to take regular breaks | 3.4 |
| 13. The reminders given by the chatbot were relevant to my interests | 2.8 |
| 15. The chatbot talked about interesting topics | 3.33 |
| 17. The different conversations with the chatbot were relevant to me | 3.27 |
| 19. The conversations were shallow | 3.53 |
| 21. The reasons given by the chatbot about its suggestions were interesting | 3.2 |
| 23. The chatbot gives practical advice | 3.07 |
| 25. I would use the chatbot in my daily life | 2.47 |
| 26. I followed the chatbot advice | 2.53 |

Table 2: Average value for 5-point Likert scale questions

The responses to the open-ended questions shed some light on the values given by the users to the 5-point Likert scale questions. The first open-ended question was question 14, which was about the reminders. The question showed that the majority of the users were not interested in having reminders. The participants did not find them beneficial because they did not feel like they were overworking, or they got them at times that were inconvenient for them.

Question 16 refers to the topics the chatbot discussed. People found that some of these topics were interesting and others were not. It depended on the subject and the ability of the bot to be coherent. Some users found the topics to be out of context, and also, the bot was not able to follow and develop a logical conversation, so the conversation never reached an intriguing point. Some participants found that the topics were random and funny rather than exciting. A few participants also pointed out that buttons restricted the conversation.

The question about the topics being relevant to the users (question 18) received similar responses as question 16, about those topics being interesting. It depended on the conversation, some of them were relevant, and some of them were not. Some people found that those conversations were catered to their particular interests. They also found that the chatbot took into account their emotions in each moment. But others pointed out that it seemed like the chatbot was not talking to them or repeated itself too much to consider the conversation relevant.

Most people considered the conversations to be shallow for different reasons (question 20), but the main one was that the chatbot struggled to be coherent, jumping from one topic to another, getting confused and replying illogically. A few participants also mentioned that it was difficult for them to engage with an artificial intelligence and even more complicated to go into emotional topics since it lacks empathy.

Question 22 addressed the reason given by the chatbot when the users asked about that reason. The principal response to this question is that they did not remember what the chatbot said when they asked about it, or they did not remember the conversation about their data at all, not only the reasons.

In question 24, the participants expanded their thoughts about the advice given by the chatbot. Some of them indicated that the activities suggested were irrelevant, unjustified, or that it would take too much time to carry them out. Others said that those activities were practical at times.

The last two open-ended questions talk about the changes the users would make to improve the conversations and things they think are lacking from the current ones. The most repeated answer to these questions is about the coherence of the chatbot. They would like the chatbot to be able to keep the conversation going and have a better understanding of the flow of human-human dialogues and what the user is trying to say. Other users mentioned that it sounded too much like a bot, and therefore, they did not find it appealing to talk to it. Participants also indicated that rude responses should be removed, that they want more variety of answers, and that they would appreciate the chatbot trying to cheer them up when they are in a bad or low mood.

## 5.3 Interview

After filling out the questionnaire, six participants were interviewed: three of them in individual interviews and the other three in a group interview. The questions asked can be found in Appendix B. These questions explore different topics that needed more detailed answers after analysing the survey.

The first question deals with the ease of opening up to a machine. The second and fourth questions talk about the management and use of the participants' data. And the third and fifth questions talk about the availability of the reasons behind that management. The sixth question is about the number and depth of the topics the chatbot knows. The last three questions deal with changes the users would like to see in the chatbot to improve their experience. Table 3 lists the codes for the questions. These codes were cross-validated by two independent coders.

To the first question of the interview, participants responded that they did not find it difficult to talk with the chatbot about their feelings. Some of them pointed out that while they did not find it complicated to engage with the chatbot, they did not share their feelings in-depth because they knew the limitations of the chatbot regarding understanding or because they needed more time to become comfortable with someone to share their feelings in that way (this also happens with people).

The following four questions are about how the chatbot uses their data. Some participants answered that they were interested in knowing how the chatbot uses their data, and others were not interested. But everybody agreed that they wanted to have this information available even if they never ask about it. They also pointed out that

| Question | Code |
|---|---|
| Did you find it difficult to share your feelings in a more intimate way with the chatbot? | Relationship human-chatbot |
| Are you interested in knowing how the chatbot uses your data? | Data management |
| Would you want this information to be available to you in case you want to ask about it? | Data availability |
| Do you prefer a general explanation or a tailored one about how the chatbot uses your data? | Data management |
| Would you have wanted an explanation about what the chatbot suggested to you after that suggestion? | Data availability |
| Do you prefer the chatbot being able to talk about any topic but superficially or the chatbot being able to talk about a few of them but deeply? | Number/depth of topics |
| What would you like to have in a chatbot? | Chatbot improvements |
| What would you change about the activities the chatbot suggested? | Chatbot improvements |
| In the beginning, would you like to have an empty user profile or a standard one? | Chatbot improvements |

Table 3: Coding

even if they are not interested in these data, they might ask about it when the bot suggests something that seems weird to them. Most of the interviewed participants preferred tailored explanations rather than general ones, but one of them indicated that a general one is sufficient due to not being that interested in knowing about the usage of the data gathered.

When asked about having the suggestion explanation after that suggestion, they said that it might be interesting to have it there as long as they can choose when to see it. The participants mentioned that having a button to ask about the reasons might trigger their curiosity. Even those that indicated that they were interested said that they might forget that the option to ask about their data exists if they do not see this option.

Regarding the depth of the conversation and the number of topics the chatbot can discuss, there was no consensus since both options have advantages and disadvantages. One of the users pointed out that a chatbot that can talk about any topic but superficially might keep you company and that one that can talk about a few of them but profoundly might give you better advice. Another user thinks that superficial conversations are more draining than helpful. Besides personal preferences, there was a point in which all users agreed on, and this was that if the chatbot is not able to talk coherently, it does not matter how many topics it knows or the depth of its dialogues.

The lasts questions were about changes in the current design and the features the users would like to have. Everybody mentioned coherence and memory to remember what the user just said and use it in a future conversation. They also indicated that gathering more data to tailor more the dialogue. Some user suggested adding interactive activities or jokes to lighten up their mood. The participants agreed that the chatbot should consider more information than the current mood before suggesting an activity. Other data that can consider are location, current situation (e.g. a pandemic), length of the activity, schedule and habits of the user, etc. Regarding having an empty profile or a standard profile, they preferred a standard one, chosen after asking a few questions. That way, they can receive suggestions and recommendations from the very beginning.

# 6 Discussion

The results for the system metrics show that the average precision of the classifier (36%) is lower than above chance. This low precision means that the chatbot started the conversation about the feelings of the users when it should not have done it. Although the average precision was 0.36, it improves during the two weeks that lasted the experiment since we trained the classifier daily with the sentences the participants used.

There are two main groups of incorrect classified utterances: sentences that have nothing to do with feelings and sentences that depend on the context. The utterances that have nothing to do with feelings are easy to detect since the chatbot starting a feelings conversation seems out of place. An example of this kind of interactions (taken from the dialogue of one of the participants) is the following:

**User:** what is your favourite animal?
**Chatbot:** i would say something like a dog
**User:** how about a wolf?
**Chatbot:** hey! do you want to share your feelings with me?

It is clear that the utterance *how about a wolf?* is not related to the user feelings. Nonetheless, it was classified as *feelings conversation*. The misclassification of the other incorrectly labelled sentences is more subtle since these utterances depend on the context, making their classification difficult. While for humans, it is easier to discern what a sentence is about, based on the context of the conversation, for a chatbot is not that easy, and the more general the discussion can be, the more difficult it will be for the chatbot to discern and classify that sentence correctly. The following is an example of a sentence that depends on the context:

**User:** i try my best
**Chatbot:** it seems like you don't have enough energy today

In this situation, *i try my best* is not about the user being in a bad or low mood, so it is a misclassification, but in another context, it can mean that the user is discouraged about something they are doing and it is not giving the expected results.

Usually, people do not have trouble classifying the sentences shown in the examples because they have reached common ground [22] with their conversational partner. Having a common ground allow speakers to have a foundation to deal with ambiguity. If they cannot clarify the ambiguity, they will ask the other person to ensure that they understand the sentence correctly. Although the field of natural language processing (NLP) has improved a lot in recent years, there is still a lot of work to do before chatbots can have conversations similar to human-human discussions. Grounding is a needed feature that will help to deal with ambiguity and add naturalness to conversations with chatbots. Increasing the number of sentences in the database will improve the precision of the classification. But the problem posed by the second kind of sentences is more complex and needs more steps than only increasing the database. Adding a module for context-awareness will help the bot to discern the correct classification in these difficult situations like the one used in this hybrid chatbot[10] that uses contextual information to influence the chatbot responses.

Buttons restrict the user input, and some users tend to avoid using them, which can cause the system to crash in some situations. While getting rid of the buttons might not be an option because it will be more complicated for the chatbot to understand the user utterances; adding implicit and explicit grounding (e.g. "did you say you wanted to do it?") might keep the dialogue active and avoid the system from crashing.

The method chosen for recording the activities was problematic. As figure 6 shows, the users reply to an open-ended question about what they feel like doing at that moment, and here two different problems arise. The first one is that some users start a new conversation instead of answering something that is an activity, and because the filter is not focused on finding this kind of utterances, these sentences are saved in the user profile as the preferred activity when they are not an activity. Here comes into play another feature of human-human conversation that is turn-taking. People are used to adapting to their conversational partner rhythm [11], but machines lack the ability to pick up on all the cues shared by the speakers during the conversation. Therefore, changing the pacing of the discussion suddenly means that the chatbot will need to readjust to the new rhythm. The other problem of this method is when the activity inputted by the user is not registered in the hobbies database. A random hobby from the database was

[10] https://github.com/ubaer/Personalised_context_aware_DialoGPT

selected to solve the problem of having a non-activity in the activity field of the user profile; while this is not the best possible solution, it did not cause any troubles during the experiment.

The filter used for the open-ended question brought up another issue. Many users interacted with the chatbot past midnight, and when the chatbot asked about the activity they wanted to do, they usually said to sleep, which is one of the keywords in the filter due to its implications to low moods, but in this context, it was a normal thing to want to do because it was late and many of them were tired after working the whole day. Adding a contextual information module can solve this problem since it would provide the chatbot with the information needed to assess the validity of the activity.

This problem with the filter shows that chatbots are a good tool for just-in-time adaptive interactions (JITAIs) [14] since they are always available and can talk about what is happening at the exact moment it is happening.

In the following example, the user did not input a valid activity and instead tried to change the topic of the conversation:

**Chatbot:** hey! do you want to share your feelings with me?
**User:** yes
**Chatbot:** how pleasant are those feelings
**User:** it is okay
**Chatbot:** what about energy, how energized do you feel?
**User:** somewhat tired
**Chatbot:** what would you like to do know?
**User:** how are you?
**Chatbot:** cool choice!

The chatbot has some unexpected and rude responses. These responses are not acceptable in a chatbot designed to talk about feelings and should be avoided since they complicate the relationship between the chatbot and the user. One of the things users mentioned during the interviews was that it takes time to become comfortable with someone before being able to talk about deep topics, and they expect the same to happen with a chatbot. Therefore having rude responses will hinder the progress and make it more difficult for the users to trust in the chatbot.

Below there are two examples of rude responses given by the chatbot to two different participants. These examples show the risks of not filtering data before training a natural language generation model.

**Chatbot:** what is your favorite color?
**User:** purple
**Chatbot:** you disgust me
**User:** hahahaha why?
**Chatbot:** because purple?

**User:** i'm happy
**Chatbot:** i'm sad
**User:** why
**Chatbot:** because your mother didn't love you
**User:** mmmm she did
**Chatbot:** you should have kept her

Only two of the 5-point Likert scale questions, table 2, have values significantly different from the average value (three). The open-ended questions and the interview show what the users think in detail. Most participants did not like the reminders because that kind of reminders was ineffective for them. It does not mean that they are against the chatbot taking an active role in the conversations instead of the passive one it usually has. In fact, some users were not too bothered by the low precision of the classifier because they thought it was the chatbot being proactive.

The advice the chatbot provided was not always practical at times. For example, some activities were inconvenient due to the time and the current situation (a pandemic). This topic was explored during the interviews, where participants stated that adding more information such as their location or schedule would improve the suggestions. Also, two weeks might not be enough time for the bot to get enough data and create an accurate user profile. More time of use will allow the chatbot to gather more data, and it will also let the users become familiar with the system.

One of the goals of this project was to create a smooth transition between the related conversation and the meta-conversation. These meta-conversations were related to the dialogue about the users' feelings, and the participants could ask several questions about the usage of their data. The questionnaire results showed that the users did not seem interested in this meta-level. Since it was a core idea of this project, the interviews were conducted to shed some light on this topic. There was no consensus about the users' interest in knowing the usage of their data, but all of them agreed on being interested in having an explanation always available even if they said that they were not interested in knowing how the chatbot used their data. The information obtained from the participants showed that it did not matter whether they were interested or not in knowing the usage of their data; they needed to see that they could ask about it because if not, they would forget about the possibility of asking. They also stated that if that information was available in the same conversations as the one where they share their feelings, it might trigger their curiosity, and they might ask more about it.

These results pose some interesting questions. Users do not seem interested enough in this information to go out of their way to ask about it, but they might get curious if these data are more visible. But it is possible that making the information more visible interferes with the progression of the conversation. In rule-based chatbots, the meta-conversation questions will be limited to a certain number, and these questions will be based on the topic of the related discussion. Therefore, the users' interest in the meta-conversation may vary depending on the subject of the related conversation. Another question is the role meta-conversations play in the explainable AI field since they can explain to the users the reasoning behind the chatbot choice, making the chatbot more transparent [18] for the users. This project tries to shed some light on a topic barely investigated. All these questions that it sets out remain for future investigations.

# 7 Conclusion and future work

From the previous sections, results and discussion, it is clear that the current design can benefit from some changes, such as increasing the database to improve the precision of the classifier or filtering rude responses. Another possible design change would be clustering new users on standard profiles based on their replies to a few questions before they start talking with the chatbot. This change would allow the users to receive suggestions from the beginning and avoid the data gathering phase of the current design. These standard profiles would be tuned using the users' data gathered during their interactions with the chatbot.

Some users suggested during the interviews that increasing the exposure to the chatbot might increase their feelings of familiarity and that they might feel more comfortable sharing their emotions with it. Therefore increasing this exposure will allow testing whether the users feel more comfortable. It would also allow gathering more data to personalise the conversation, which will affect the relationship with the chatbot.

The goal of this project was to smooth the transition between related and meta-conversation in order to improve the user experience and naturalness of DialoGPT [27]. The topic chosen for the related discussion was the feelings of the users.

While the design tested did not improve the user experience or naturalness of the system mainly due to the incoherence of the chatbot, this design shows an intriguing research path to keep investigating. It is left as future research to investigate how the users interact with their data, when they are interested in those data or how to introduce the information while keeping the conversational flow natural. Adding a button after a suggestion with the reasons behind that suggestion is a way to present this information. This way, the users can choose when they want to see it. If they decide not to see those data, this will not affect the conversational flow. Besides these particular challenges, chatbot technology still has to work on including the features of human-human conversations (turn-taking [11], [17], grounding [22], etc.).

In conclusion, it is still unknown when a user will want to access the meta-conversation, but they want to have it available. And also that coherence is a feature of utmost importance in chatbots because users will not engage with the chatbot without it.

# References


[1] Emotions dataset for nlp classification tasks. `https://www.kaggle.com/praveengovi/emotions-dataset-for-nlp`.

[2] User experience questionnaire. `https://www.ueq-online.org/`.

[3] Anderson, M. L., Fister, A., Lee, B., Tardia, L., and Wang, D. J. J. On the types and frequency of metalanguage in conversation : A preliminary report.

[4] Barak, A., Hen, L., Boniel-Nissim, M., and Shapira, N. A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic interventions. *Journal of Technology in Human Services 26*, 2-4 (2008), 109–160.

[5] Beck, J. S. *Cognitive Behavior Therapy, Third Edition: Basics and Beyond*. Guilford Publications, 2020.

[6] Behrens, F., and Kret, M. E. The interplay between face-to-face contact and feedback on cooperation during real-life interactions. *Journal of Nonverbal Behavior 43*, 4 (2019), 513–528.

[7] Ben-Zeev, D., Kaiser, S. M., Brenner, C. J., Begale, M., Duffecy, J., and Mohr, D. C. Development and usability testing of focus: a smartphone system for self-management of schizophrenia, Dec 2013.

[8] Colby, K. M. *Artificial paranoia: a computer simulation of paranoid processes*. Pergamon Pr., 1975.

[9] Dimeff, L., and Linehan, M. M. Dialectical behavior therapy in a nutshell. *The California Psychologist 34*, 3 (2001), 10–13.

[10] Fitzpatrick, K. K., Darcy, A., and Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health 4*, 2 (2017).

[11] Himberg, T., Hirvenkari, L., Mandel, A., and Hari, R. Word-by-word entrainment of speech rhythm during joint story building. *Frontiers in Psychology 6* (2015), 797.

[12] Jurafsky, D., and Martin, J. H. *Chatbots and Dialogue Systems*. 2014.

[13] Muller, A. M., Blandford, A., and Yardley, L. The conceptualization of a just-in-time adaptive intervention (jitai) for the reduction of sedentary behavior in older adults. *mHealth 3*, 9 (2017).

[14] Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine 52*, 6 (12 2017), 446–462.

[15] PhD, W. R., PhD, J. O., and MS, J. J.-M. Internet and mobile phone text messaging intervention for college smokers. *Journal of American College Health 57*, 2 (2008), 245–248. PMID: 18809542.

[16] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners.

[17] Riest, C., Jorschick, A. B., and de Ruiter, J. P. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in Psychology 6* (2015), 89.

[18] Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. Explainable machine learning for scientific insights and discoveries. *IEEE Access 8* (2020), 42200–42216.

[19] Ruwaard, J., Lange, A., Schrieken, B., Dolan, C. V., and Emmelkamp, P. The effectiveness of online cognitive behavioral treatment in routine clinical practice. *PLoS ONE 7*, 7 (2012).

[20] Samek, W., and Müller, K.-R. *Towards Explainable Artificial Intelligence*. Springer International Publishing, Cham, 2019, pp. 5–22.

[21] Speer, R., Chin, J., and Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. *CoRR abs/1612.03975* (2016).

[22] Stalnaker, R. C. *Assertion*. Brill, Leiden, The Netherlands, 20 Dec. 1978, pp. 315 – 332.

[23] Sucala, M., Schnur, J. B., Constantino, M. J., Miller, S. J., Brackman, E. H., and Montgomery, G. H. The therapeutic relationship in e-therapy for mental health: A systematic review. *Journal of Medical Internet Research 14*, 4 (Feb 2012).

[24] Turcan, E., and McKeown, K. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)* (Hong Kong, Nov. 2019), Association for Computational Linguistics, pp. 97–107.

[25] Walker, M., and Whittaker, S. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics* (USA, 1990), ACL '90, Association for Computational Linguistics, p. 70–78.

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

[27] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. Dialogpt: Large-scale generative pre-training for conversational response generation, 2019.

# A Questionnaire

## A.1 Statement of informed consent

**Q1.** This experiment utilises a general-purpose chatbot focused on user feelings (managed by a rule-based bot). You can speak with the chatbot about anything you want, although you should talk with the bot about your feelings, you can be as personal as you wish.

There are two phases, and I will notify you when you change stages. There are three different conversations in the rule-based bot: one to report a malfunction, one to talk about your emotions, and another one to manage/ask about your data. You need to use the feelings and manage data conversations a few times. The feelings dialogue even once a day if possible.

You will have to use the bot for about two weeks and after that answer some questions. Try at least to use the bot once a day at any time, but you can use it more time per day if you want.

Your data will be collected, stored and anonymised, but you can delete it at any moment through the chatbot or contact me to delete it. You can withdraw from the experiment at any moment.

If the chatbot gets stuck on the conversation, you can use the keyword 'bye' to reset the conversation.

**Q2.** You are participating voluntarily in this experiment and can terminate your participation at any moment you want.

The experiment has two parts:

1. Talking with a chatbot for two weeks
2. Completing a questionnaire about the experience

You can be contacted again after the survey for a follow-up interview.

The risks of this study are minimal, but they exist since the automatic generation of the responses. Therefore, some of the messages from the chatbot might come across as offensive.

The survey is a Qualtrics Survey, and the answers are stored in the Qualtrics cloud. Each participant has an ID or number that connects the consent form with their questionnaires. The researchers will be the only people with access to these IDs. In the case of a data breach, it will be more complicated to connect each participant with their survey.

The user profile is stored in the MongoDB cloud, and the personal information that might be disclosed during the conversation will be anonymized.

In case you need to contact the researchers, you can contact Asun Alsina López in the following email address: **m.a.alsinalopez@students.uu.nl**

**Statement of informed consent:** I understand that the responses I provide in this experiment will be stored anonymously. I furthermore understand that I am free to terminate my participation in this experiment at any time. I am 18 years or older. By pressing Yes below, I give my informed consent to store my responses anonymously.

**Q3.** ID

## A.2 Survey

**Q1.** ID

**Q2.** What is you age?

**Q3.** Are you a native English speaker?

**Q4.** On a daily basis, for how long did you use the chatbot? (Approximately)

**Q5.** In the next block you will encounter some questions about your interaction with the chatbot.
Some of these questions have an optional follow-up question where you can extend your response.

**Q6.** The chabtot is annoying

**Q7.** The conversation with the chatbot is predictable

**Q8.** The chatbot is friendly

**Q9.** This chatbot is unlikable

**Q10.** The language and conversations feel natural

**Q11.** The buttons are bothersome

**Q12.** The chatbot enables me to take regular breaks

**Q13.** The reminders given by the chatbot were relevant to my interests

**Q14.** Why were/weren't those reminders relevant?

**Q15.** The chatbot talked about interesting topics

**Q16.** Could you say more about why the chatbot talked/didn't talked about interesting topics?

**Q17.** The different conversations with the chatbot were relevant to me

**Q18.** What made those different conversations relevant/not relevant to you?

**Q19.** The conversations were shallow

**Q20.** Why do you think those conversations were/weren't shallow?

**Q21.** The reasons given by the chatbot about its suggestions were interesting

**Q22.** What makes you say that the reasons were/weren't interesting?

**Q23.** The chatbot gives practical advice

**Q24.** Why was/wasn't the advice practical?

**Q25.** I would use the chatbot in my daily life

**Q26.** I followed the chatbot advice

**Q27.** After talking the chatbot, what would you change/add to improve the dialogues?

**Q28.** What do you think is lacking from the conversations?

**Q29.** Any other remarks?

# B Interview

- Did you find it difficult to share your feelings in a more intimate way with the chatbot?
- Are you interested in knowing how the chatbot uses your data?
- Would you want this information to be available to you in case you want to ask about it?
- Do you prefer a general explanation or a tailored one about how the chatbot uses your data?
- Would you have wanted an explanation about what the chatbot suggested to you after that suggestion?
- Do you prefer the chatbot being able to talk about any topic but superficially or the chatbot being able to talk about a few of them but deeply?
- What would you like to have in a chatbot?
- What would you change about the activities the chatbot suggested?
- In the beginning, would you like to have an empty user profile or a standard one?