Imperial College London

MEng Individual Project

Imperial College London

Department of Computing

Evaluation of a Virtual Agent in Guiding Users from the Non-Clinical Population in Self-Attachment Intervention

Author:
Ali Ghachem

Supervisors:
Professor Abbas Edalat
Adrian Millea

Second Marker:
Professor Lucia Specia

June 13, 2021

Abstract

Acknowledgements

I would like to thank Prof. Abbas Edalat for his tremendous support and guidance in developing my understanding of Self-Attachment and coordinating this project and the corresponding trial, which has greatly assisted the development of this project. His contributions have also greatly helped with the design of the rule-based model.

I would also like to thank Prof. Lucia Specia for her suggestions during the early stages of this project, which provided various research directions to consider to reach a final project.

I would like to thank Adrian Millea and Neophytos Polydorou for their help with researching the project and developing some of the initial codebase for the project platform.

I would like to thank Lisa Alazraki for her help and numerous discussions throughout the latter stages of this project, as well as her assistance with the construction and handling of the survey distributed to Amazon Mechanical Turk.

I would like to thank James Tavernor, Huilun Tao, and Fengzheng Shi for their discussions throughout this project that have helped me to form this project.

I would like to thank Farnaz Edalat for her help in reaching out to participants for the trial. I would also like to thank all participants who examined the platform as the feedback has been incredibly useful.

I would also like to thank Dr Kami Saedi, Dr Frank Ryan and Dr Dasha Nicholls for their evaluation of the platform.

Finally, I would like to give special thanks to Prof. Foaad Khosmood (California Polytechnic State University) for his fantastic assistance with the development and planning of the rule-based model in the mid to late stages of the project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mental health is a growing issue worldwide, with the COVID-19 pandemic contributing to increasing mental health cases. 10 million people require additional mental health support as a result of COVID-19 [10]. Additionally, people diagnosed with COVID-19 in late 2020 and early 2021 were found to be more likely to suffer from anxiety, depression, mood disorders, or other psychological or neurological conditions [11]. In recent years, chatbots such as Woebot have demonstrated the ability to assist in delivering types of therapy, such as Cognitive Behavioural Therapy [12], to assist with people suffering from anxiety and depression. Furthermore, at least 264 million people worldwide currently experience depression, which has led to the development of applications such as the Flow Depression App [13]. The app uses a chatbot to support users taking electrical treatment, which stimulates parts of the brain that are less active for people suffering from depression. Users are then provided with exercises to perform alongside the treatment and videos on mindfulness to support their mental state. The chatbot is also able to encourage users to continue progressing with sessions and encourage positive actions from the user while recording the user's progress with the treatment so this can be displayed to the user.

Other approaches to improve certain mental health conditions include video games such as Neuroracer and Meditrain, which were primarily developed to improve the attention abilities of patients [14]. Another video game used in this area is Endeavour RX, which became the first FDA-cleared video game for any medical condition following a one-month trial on children with ADHD, which led to improved real-world functionality in patients. Finally, a more passive treatment was developed by Sensync in the form of a Sensory Immersion Vessel, which a user can enter to experience and become immersed in a natural setting, such as a forest, by engaging with the environment through all senses.

Self-Attachment is a mental health treatment designed to mentally retrain a patient to more readily control their own emotions [15]. A set of Self-Attachment protocols were proposed by Edalat [15] and subsequently developed by the Algorithmic Human Development group to assist patients suffering from chronic depression or anxiety. Treatment aims to provide a patient with unconditional love that may not have been provided to them by a carer during childhood. This is achieved through developing an imaginary bond with the user's childhood self, where the user undertakes both the role of parent and child to create and maintain this bond and subsequently be more able to regulate their emotional state.

The Algorithmic Human Development group have developed a Virtual Reality (VR) platform to scale the delivery of Self-Attachment. To assist in scaling the delivery of Self-Attachment, we have developed a rule-based model for a chatbot, SATbot, that will support the delivery of the Self-Attachment protocols. This has been achieved by communicating with users through a text interface and inferring a user's emotion and mindset, using this information to assist the model's responses to assist users using the platform. SATbot has been integrated into a web interface to use within a trial on the non-clinical population to evaluate its effectiveness at delivering Self-Attachment protocols. We also gathered results from a survey of 100 people who rewrote base prompts of the chatbot to be more empathetic, which provides the chatbot with a collection of responses to select from at each stage. This survey has been gathered and will be integrated into the chatbot in future work. We received ethical approval from Imperial College to perform the survey and trial.

This report will discuss the development of SATbot and the survey and trial. The various approaches considered, existing chatbots and rule-based models, and an overview of Self-Attachment are described in Chapter 2. The design choices are discussed in Chapter 3, and the technical implementation of the model is described in Chapter 4. The trial is discussed in Chapter 5. The trial for the chatbot is evaluated in Chapter 6. The survey which we gathered is discussed in Chapter 7, alongside future work it may contribute to, and Chapter 8 will discuss findings and future iterations of the rule-based model.

Chapter 2

Background

2.1 Self-Attachment

Self-Attachment aims to assist an individual in receiving unconditional love and attention that may have been deprived during childhood due to not forming a "secure attachment" [16, p. 6] with their parent, which is hypothesised to result in mental health problems for an individual in later stages of life [15]. Self-Attachment is based on attachment theory, which considers secure and insecure attachments between young children and their caregivers. Secure attachments are encouraged as through sufficient and appropriate attention from the caregiver, the child will develop a more secure mental state and internalise the caregiver's ability to regulate their emotions [16, 17, 15]. By contrast, the development of insecure attachments can lead to the development of mental health problems through dysfunctional thoughts and functions emerging within the child [16]. As early development of the brain is shaped by these attachments, Self-Attachment is designed to retrain the brain to allow the individual to become more capable of regulating their own emotions [16, 15].

Self-Attachment therapy consists of four steps [15]:

1. Volunteers are introduced to the concept of Self-Attachment therapy and attachment theory to understand what the therapy will require and the aim of the therapy. As the protocol requires a large amount of dedication from volunteers, the requirements are described to motivate volunteers towards continuing with the therapy.

2. Volunteers are encouraged to connect with their childhood self by looking at photos of their childhood self and recalling their childhood environment. Volunteers are asked to provide happy and sad photos of themselves and think about happy and sad moments from their childhood. Exercises are performed to connect to the childhood self in different emotional states, where the goal is for volunteers to feel compassionate towards their childhood self. A volunteer should be able to conceptualise the child and form empathy and compassion towards it, which is instrumental in being able to mimic the relationship between parent and child [17].

3. Volunteers are tasked to "fall in love" with their childhood self, which will form the affectionate bond required and lead to being able to regulate their emotions more effectively [15]. This is hypothesised to be achievable through "primary narcissism" [15, p. 6], which is present among all humans. The volunteer is tasked to imaginatively re-adopt the childhood self and vow to re-parent it, using a favourite love song to emphasise the emotional attachment. As a result, the bond-making should improve a patient's willingness to proceed with the remaining sessions through negative emotions reducing as a result of this process, which motivates patients to continue as they will anticipate further emotional development.

4. Volunteers then re-parent their childhood self [15, 17]. Their childhood self is interacted with in such a way to minimise negative emotions in the volunteer and maximise their positive emotions. By adopting the role of the parent, the volunteer can develop optimal neural functions and patterns to replace dysfunctional patterns that developed during childhood. A secure attachment

object is formed either in the patient's mind or on paper, such as a picture of a bright house, which symbolises the secure attachment that the volunteers now have. Exercises presented throughout this stage are intended to be repeated gradually into everyday activities to be able to retain the connection with their childhood self and the subsequent emotional regulation.

2.1.1 Scaling Self-Attachment Delivery

To enhance the delivery of the Self-Attachment protocols and make it scalable, a Virtual Reality (VR) platform has been developed by members of the Algorithmic Human Development team to deliver sessions to patients. The platform consists of a virtual agent and a child avatar. The virtual agent explains the steps of each protocol and interacts with the user to determine their current emotion through their speech, and provides subsequent Self-Attachment sub-protocol suggestions, whereas the child avatar is used to represent a user's childhood self. Applying the user's childhood photo to the avatar has been suggested to be more effective than using a childhood photo without the avatar as the object to connect to their childhood self. The child avatar can be interacted with through speech or touch, and the user's hand gestures are used to track physical interaction such as embracing the child. The emotion recognition model used within the platform has been extended to analyse both text and speech [18], which removes the need to pre-process data and improves the accuracy of emotion recognition for happy, sad, neutral and angry emotions. The model was additionally applied to assist predictions of depression and anxiety.

Following from this, the aim of this project is to develop a chatbot which will be able to infer a person's emotion and mindset and further assist patients undertaking Self-Attachment therapy. The chatbot can then be integrated into the VR platform in future work.

2.1.2 Protocols

In previous trials of Self-Attachment, volunteers have been taught 20 protocols over 8 weeks that they can utilise to maintain their emotional state. These protocols primarily focus on reinforcing a positive mindset through furthering the relationship with the childhood self, comforting the childhood self as if the volunteer is the parent to be able to overcome emotions that arise from recalling previous events, and being able to change perspectives about a particular topic so a volunteer can better manage their emotional state.

In the model that we are developing, participants will have already been introduced to these protocols in previous Self-Attachment workshops, and will be suggested a protocol from this set depending on the situation and emotion described when conversing with the model.

A list of protocols with short descriptions is provided in Appendix B.

2.2 Definitions

Some definitions that appear in later research are described in this section.

2.2.1 Perplexity

Perplexity was first defined as "the reciprocal of the geometric average of the probabilities of the predictions in S" [19, p. 3] for a sample text S, but has also been defined for a sequence of words w_1, ..., w_t as $1/P(w_t|w_1^{t-1})$ where $P(w_t|w_1^{t-1})$ is the probability of generating the t-th word w_t given the words w_1, ..., w_{t-1}, and $1/P(w_t|w_1^{t-1}) = f(w_t, \cdots, w_{t-n+1})$ where f is the function that the model intends to learn [20]. It is also defined as the exponential of the average negative log-likelihood [20], which was used by Serban et al. [21]. A lower perplexity indicates a higher probability of selecting a given token, which means the model is more certain about a prediction [19]. Serban et al. [21] additionally discuss that perplexity measures the model's ability to account for the syntactic structure of both the dialogue and each utterance. Overall, perplexity is a suitable measure for a probabilistic model and is used as a metric in automated evaluation when comparing dialogue generation.

2.2.2 F1 Score

The F1 score is defined as

$$\frac{2 * (Precision * Recall)}{Precision + Recall}$$

where Precision is the number of true positives divided by the sum of the number of true positives and the number of false positives, and Recall is the number of true positives divided by the sum of the number of true positives and the number of false negatives.

2.2.3 N-grams

N-grams are a type of language model where two histories are equivalent if they both end with the same n - 1 words, but the nth word can be different [19]. N-gram models are described as models that construct conditional probabilities for the nth word over possible combinations of the last n - 1 words, also known as contexts [20].

2.3 Neural Network Models

Some models that appear in later research are described in this section.

2.3.1 Long Short Term Memory

To overcome the inability of the Recurrent Neural Network (RNN) to retain information over long periods, alongside the vanishing or exploding gradient issues that may occur in RNNs, the Long Short Term Memory (LSTM) model was developed [22, 23]. The LSTM is divided into modules which each have three gates: the input, output and forget gates [22, 24]. These gates regulate how much information is allowed through to update the state, where the input gate regulates how much of the newly computed state is used, the forget gate regulates how much of the previous state is used, and the output gate regulates how much of the internal state is passed to the rest of the network [22, 24]. The information will then update a hidden state in each cell similarly to an RNN [22, 24].

2.3.2 Sequence-to-Sequence

The sequence-to-sequence model (Seq2Seq) uses two LSTMs: one for the input sequence and one for the output sequence [25]. The first LSTM maps the input sequence to a fixed representation, and the second LSTM decodes the representation into a target sequence.

2.3.3 Memory Network

Memory Networks consist of an array of objects that form the memory, and four components: an input feature map that converts the input to the feature representation; a generalisation component that updates old memories given the new input; an output feature map that produces the new output given the input and current memory state; and a response component that converts the output into the desired response format [26].

The architecture has since been recreated so that it can be trained end-to-end and requires less supervision during training, so it is more applicable to realistic settings [27]. The new architecture has similar performance to the original Memory Network and outperforms trained RNNs and LSTMs for language modelling, so it is suitable for dialogue generation. This model was used to identify relevant knowledge from dialogue to be able to return more informed responses, particularly when trained on the Wizard of Wikipedia dataset [1].

2.3.4 Transformer

The Transformer network is based on the concept of attention mechanisms [28]. An attention function takes a query vector and a set of key-value pairs, where keys and values are vectors, as inputs and maps them to an output vector, where the output is defined as a weighted sum of the values passed as parameters to the attention function. The weight for each value is determined by a compatibility function to determine how compatible the query is with the corresponding key.

The Transformer network utilises attention to determine dependencies between the input and output and determine which values are more important by providing these with larger weights for the final softmax layer, which outputs a probability distribution of the words from the key-value pairs [28].

2.3.5 BERT and BioBERT

BERT stands for Bidirectional Encoder Representations for Transformers, and uses Masked Language Modelling to train the network by masking random words, or tokens, from the input, and attempts to predict these tokens given the remaining unmasked input [29]. BERT models aim to learn a language representation from inputs it is trained on that can then be applied to various tasks when fine-tuning on smaller datasets. By pre-training using Masked Language Modelling, the network only requires an additional output layer to fine-tune to different tasks. For this reason, BERT was considered as a potential model to use for modelling dialogue for SATbot.

BioBERT stands for Bidirectional Encoder Representations from Transformers for Biomedical Text Mining, which is a language representation model that is domain-specific to biomedical data [30]. The model was trained on biomedical corpora and greatly outperformed BERT and other state-of-the-art models on text mining tasks when these models were pre-trained on biomedical corpora. While these corpora include some references to therapy and mental health, many tokens within the corpora focus on other areas in medicine that are less relevant for training. This model was considered to use for the chatbot to provide a basis of relevant terms for language modelling.

2.4 Analysis of Therapy Sessions

Previous work has investigated dialogue from sources such as therapy sessions to learn more about how to analyse the results of psychotherapy sessions. This section explores these investigations to evaluate whether these findings will be useful for producing dialogue for SATbot for conversations with users.

2.4.1 Analysing IECBT Sessions

Due to a lack of objective methods for measuring psychotherapy, an investigation of Internet-Enabled Cognitive Behavioural Therapy (IECBT) sessions was undertaken to automate analysis of therapy sessions using deep learning and determine whether there was an association between different aspects of therapy sessions and the clinical outcome of patients [31]. Cognitive Behavioural Therapy is described as a class of therapy that motivates changing away from thoughts, behaviours or functions that are maintaining mental disorders.

To determine whether an association existed, IECBT sessions between June 2012 and March 2018 were used as a dataset to train a deep learning model [31]. Within the model, words are represented as vectors called word embeddings which are computed with Word2Vec and preprocessed by tokenising on whitespace and punctuation and lowercasing all tokens, including punctuation. These were fed into a bidirectional LSTM to allow context and word order to be encoded. Max pooling was then applied to take the most likely token from each word and combine them to form the representation. Another LSTM takes the utterance representation and encodes it using context from the entire transcript. The utterance representation is then converted into an utterance-in-context representation and translated into the class that the utterance falls into.

The model was trained on data from more than 14,000 patients, resulting in over 90,000 therapy sessions being used for training [31]. The model was used to classify therapist utterances into one of 24 categories based on the role they had during the therapy session, such as setting homework or asking the patient for feedback. The model was then used to classify data for 14,899 patients so the data could be analysed. The data was analysed on reliable improvement, which was measured by a decrease in score within the Patient Health Questionnaire (PHQ-9) and Generalized Anxiety Disorder 7-item scale questionnaires, and engagement with treatment, which was referred to as IAPT-engagement, to determine whether there were associations between: features of all sessions that patients took part in and their subsequent clinical outcome; features from the first session that a patient took part in and their corresponding engagement with treatment; and features from the first session that a patient took part in and their subsequent clinical outcome.

The findings illustrate a focus on setting the next session, setting homework, saying thanks, changing methods, giving feedback and determining the disorder, which all favoured improvement in the patient if displayed in a first session [31]. From examining all sessions we see praise, planning for the future, perceptions of change, changing methods, setting an agenda, eliciting and giving feedback, and reviewing homework all favour improvement.

Another study was performed to analyse patient utterances to determine whether a patient that acts towards change will have a better outcome, as expected from CBT [32]. The dataset was extended to consist of 34,000 patients, resulting in 188,000 hours of therapy within the dataset. 340 transcripts were manually annotated to be able to categorise patient utterances into one of 5 categories, such as Change-Talk Active, which indicated the degree of change of a patient displayed by their utterances. The model retained the ability to classify therapist utterances; a role mask is applied to the deep learning model to ensure therapist utterances can only receive therapist classes and patient utterances can only receive patient classes. The model was optimised to jointly classify patient and therapist utterances. Predictions were notably worse for the classes where patients spoke against change or about problems that would prevent change as a result of having less data for these classes, but the model demonstrated predictive power for all classes.

As with the first study [31], clinical outcomes were measured on reliable improvement and their engagement with treatment [32]. Analysis of data was performed to determine: whether there is an association between patient utterances and their corresponding clinical outcomes; whether patient utterances within the first treatment session were predictive of reliable improvement at the end of treatment sessions; and whether patient utterances in the first treatment session were predictive of their engagement.

As hypothesised there was a better clinical outcome for patients saying and acting towards changing behaviours, and a worse clinical outcome when a patient was speaking against change or showing no movement[32]. Similarly, within the first session, if a patient indicated more readily about a desire to change, their final engagement with treatment and reliable improvement was better. SATbot could utilise coding demonstrated in this deep learning model to better understand utterances within a therapy dataset. The dataset will similarly be useful for training the chatbot.

This dataset was requested to use to train the model but ultimately could not be shared due to privacy restrictions on the data.

2.4.2 Deep Learning for Language Understanding

Rojas-Barahona et al. [33] aimed to understand CBT concepts through deep learning, and defines a mental health ontology consisting of thinking errors, emotions and situations, such as work and relationships. Thinking errors are ways of thinking that are counterproductive, such as jumping to negative conclusions.

A Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) model were developed to extract features corresponding to thinking errors, emotions or solutions [33]. GloVe word vectors were used for the CNN model to represent words as vectors to use as word embeddings for these models, whereas skip-thought vectors were used to represent entire utterances as vectors for the GRU model to process. To train the model, a corpus consisting of approximately 500,000 utterances from itskoko.com was used. The corpus consisted of peer-to-peer therapy: a person would post their problem and the most negative take on that problem that they can describe, and other users will post suggestions to attempt to resolve those problems and give a positive perspective. 4,035 posts were labelled with thinking errors, situations and emotions that were used as features that the model was trained to identify.

The model was evaluated on both their average F1 score in terms of correctly classifying features and their average F1 score weighted with the frequency of each CBT label, and compared to other models [33]. Before weighting the F1 score, the CNN model had the highest average F1 score, and after weighting the GRU model had the highest average F1 score.

The dataset would be useful to train SATbot with, and the ontology defined by Rojas-Barahona et al. [33] could be adapted to identify features within utterances to use in the response. However, when designing the chatbot, we determined that it was more important for the chatbot to be able to provide appropriate, empathetic and engaging dialogue, and future sections detail the research undertaken to explore these areas for neural models.

This dataset was requested to use to train the model but ultimately could not be shared as the original authors did not have access to the data and the company working with the authors who owned the data was acquired and was not interested in further university collaborations.

2.4.3 Large Scale Analysis

To determine what components lead to a good counselling conversation, Althoff, Clark and Leskovec [34] evaluate a counselling dataset which uses SMS to communicate between the patient and therapist. The conversations were anonymised and provided by a not-for-profit organisation. The conversations were between only two people, which would make them suitable for SATbot to use for training to understand the relationship between patient and therapist. 3.2 million messages across over 80,000 conversations were analysed as part of this study, where 408 counsellors provided advice across the dataset. System messages and survey responses which would indicate ground truths about the opinion of counselling were filtered out.

As a result of analysing the dataset, five components were identified from conversations [34]:

1. Adaptability - a good therapist should adapt to the needs of the conversation. A vector of word occurrences was constructed per conversation, and the difference between how a counsellor interacts in positive and negative conversations was evaluated, and from this, it was deduced that better counsellors adapted when the conversation became negative to improve the situation.

2. Ambiguity - the less ambiguous a therapist is, the better the conversation went. This was measured by examining the start of the conversation to determine whether a therapist asks for information and whether a therapist provides enough information when a question is asked. Generally, longer responses indicated a lower ambiguity and a better result.

3. Creativity - conversation results were better when counsellors used more varied sentences as opposed to templated responses.

4. Making progress in the conversation - a therapist that made progress in the conversation quickly led to better results. More successful counsellors would begin with a shorter introduction and quickly move to discuss the problem with the patient.

5. Perspective change - over time texters start to write more about the future, which is correlated with signs of progress in the patient and feeling better. Counsellors prompting these discussions were able to lead patients towards thinking about the future and moving past their problem. Another indicator of change is how texters change from talking about themselves to about others. Notably, a lack of change near the end is usually due to the person thanking the counsellor even if no improvement was detected.

Therefore, SATbot should incorporate these factors into its dialogue. The model used for the chatbot should generate adaptive and creative dialogue, and clarify any concerns raised by a user. Making progress is also important to reach the problem of the patient. Perspective change may also be useful to encourage patients towards forming further attachments with their childhood self to move past the lack of attention placed on them during childhood. As previously mentioned, it is important that the chatbot produces safe dialogue as well as engaging dialogue, and this dataset would have provided the capacity for the model to learn how to respond more appropriately.

This dataset was requested to use to train the model but ultimately could not be shared in order to protect the privacy of texters who took part.

2.5 Current chatbot approaches

Chatbots exist across multiple domains, including therapy, and can assist individuals suffering from conditions such as anxiety and depression. These chatbots are capable of communicating with humans and sustaining conversation over a long period. Our chatbot intends to similarly communicate with humans to be able to deliver appropriate protocols based on their emotion and the input provided. We evaluate existing methods to determine features that we can utilise.

2.5.1 Woebot

Fitzpatrick et al. [12] trialed Woebot, a chatbot that was designed to deliver Cognitive Behavioural Therapy (CBT) through conversations with a human and tracking their mood. Woebot is used within an Instant Messaging (IM) application and can be used on either desktop or mobile platforms. A trial of 70 individuals aged 18 to 28 was used to evaluate whether it was feasible to deliver CBT through the use of a chatbot. Participants were divided into two groups, where one group used Woebot for 2 weeks and one group only used the National Institute of Mental Health e-book to act as a control group that was only provided information. These groups were randomly allocated, where 34 people used Woebot and 36 used only the e-book.

Participants interacted with Woebot on average 12 times over the 2 weeks [12]. In each interaction, Woebot would begin by asking about current events and how the person feels so the chatbot can track their mood. After this, the participants were shown core CBT concepts through word games or videos. Within the first interaction, Woebot also briefly explains CBT and how the bot is more like a "choose your own adventure self-help book" [12, p. 3] than a psychotherapist. A decision tree was used to determine how the chatbot responded during the conversation, which means the available dialogue options of the chatbot were limited to the options within the tree.

The chatbot responded in an empathetic way based on the participant's given mood and sent specific content based on their mood, such as an event on being anxious [12]. Participants were asked if they had a goal they hoped to achieve over the 2 week period, and Woebot would expect regular check-ins from participants where it could follow up on previous activities. The chatbot sent one personalised message a day to check up on participants. The bot also attempted to encourage completion of tasks through the use of emojis and animated gifs containing messages.

The Patient-Health Questionnaire-9 (PHQ-9) and Generalised Anxiety Disorder-7 questionnaire were used to determine the frequency and severity of symptoms of depression and anxiety respectively as a result of the trial, and were used alongside a 20 item measure of positive and negative emotions called Positive and Negative Affect Schedule (PANAS), and a set of mixed format questions called Acceptability and Usability [12]. From the trial, both groups experienced significantly reduced symptoms of anxiety and the score from PHQ-9 was significantly lower only for the group that had spoken with Woebot and not for the control group, which means there were significantly fewer symptoms of depression following the trial for the group that spoke with Woebot.

This highlights that Woebot had potential to deliver CBT to a larger group of people and could be more effective than reading information to deliver self-help. In addition, these questionnaires are useful metrics to utilise when evaluating the success of the model in terms of providing effective therapy. These questionnaires were ultimately not used as we tested on the non-clinical population, though we screened participants using a different questionnaire to ensure no participant suffered from severe depression or anxiety.

2.5.2 ESM and EMMA

The Experience Sampling Method (ESM) is an emotion-aware chatbot trialed by Ghandeharioun et al. [35] that was designed to perform experience sampling while also displaying empathy by responding appropriately based on the emotion inputted by the user of the chatbot.

Experience sampling is described as the process of recording the user's feelings, which is preferred to retrospectively providing these feelings as this avoids memory bias towards inaccurate emotions [35]. To achieve this, an interface was created where a user would input their position on a 2-dimensional grid known as the Russell Circumplex model, where sample emotions were displayed to assist them. The grid represented both a person's valence, which indicated how pleased or displeased they were, and arousal, indicating whether they had high or low energy. After the user inputs their emotion, the chatbot responds by randomly selecting a prompt from one of six prompts created per possible dialogue interaction, which may include emojis to support the messages. However, the user was not able to additionally communicate with the chatbot using natural language input through the interface that was provided.

A week-long experiment was conducted with 39 participants to determine whether the positive effects of a user reflecting could be enhanced through appropriate responses from the chatbot [35]. This was achieved by dividing the participants into two groups, where one received an empathetic response upon inputting an emotion, and one received a neutral response regardless of the inputted emotion, which unlike some empathetic responses would not include emojis. A rule-based decision tree determined how the chatbot would respond to the given input from the user. The Depression Anxiety Stress Scales (DASS) questionnaire was used to assess symptoms of depression and anxiety from participants before starting and after the week. Other questionnaires were used to assess personality, current emotion (using PANAS) and whether the user liked or disliked aspects of the chatbot such as appropriateness of the tone of messages and how likeable the chatbot was. The chatbot would probe the user to perform experience sampling 5 times a day.

There was a significant difference in positive emotions as a result of the empathetic responses generated by the chatbot compared to the neutral responses for the control group [35]. Also, extroverts liked the empathetic responses more than introverts. User feedback gathered described self-reflection as a useful task, which could be applied to SATbot's conversations.

The chatbot was later extended by Ghandeharioun et al.[36] to become EMMA (EMotion-Aware mHealth Agent) through introducing the detection of a user's mood through smartphone data as opposed to user inputs. EMMA was designed to provide mental wellness through micro-activities and demonstrate that emotions could be inferred through sensor data collected from a smartphone. To achieve this, a two-week experiment was performed with 39 participants where the first week was similar to the trial performed with ESM, where both groups would submit their current emotion through the Russell Circumplex model, but one group would receive empathetic responses whereas the other only received neutral responses. The chatbot also provided small individual or social activities to address a person's mood, which are known as interventions, based on a submitted emotion. The control group similarly received interventions based on the submitted emotion.

Within the second week of the trial, EMMA would utilise data such as the average distance of a user from their workplace and current distance from home to determine a person's emotion [36]. The emotion was predicted by a Machine Learning regression model which was designed to model personal patterns of valence and arousal over the continuous scale. This emotion prediction would then be used for determining how the model would provide interventions, and also determined how the chatbot responded after a user submitted their emotion through experience sampling, which was collected the same number of times as before and was only used to validate the predicted emotions from location data.

The model was compared to picking the most frequent class from the training set and was found to perform as well as these predictions for valence, and significantly improves on predictions for arousal [36]. User feedback noted that the short activities were helpful, but sometimes they were prompted too often to take part in activities or perform an activity that they could not perform as a result of not being near work colleagues. Participants did not want to be interrupted while feeling positive and have to do subsequent exercises while they are happy, and also noted that as the empathetic tone of the model was expected after a few inputs it had less of an effect.

Overall, these trials demonstrate chatbots can utilise alternative methods to determine and react to a person's emotion that does not involve natural language input. Also, self-reflection is highlighted as an important feature from ESM and EMMA, and therefore could be incorporated within SATbot to support any discussions held between the patient and therapist.

2.5.3 #MeTooMaastricht

The #MeTooMaastricht chatbot was developed by Bauer et al. [37] to assist victims of sexual harassment within the city of Maastricht. To achieve this, the chatbot undertakes three tasks: classifying the type of harassment that the person is describing, extracting spatio-temporal information about the event, and performing dialogue with users. The chatbot was trained to classify harassment types by using data from the SafeCity reports, which consist of approximately 12,000 texts describing different types of sexual harassment. To extract spatio-temporal information, the BERT model was fine-tuned on Named Entity Recognition (NER) datasets to be able to extract named entities from texts through NER.

The chatbot asks the user about specific information related to the harassment, and provide relevant websites to support the user based on the identified type of harassment. The user is then asked if they have reported the incident to the police and the chatbot provides relevant information to do so if this is not the case, before asking for consent to keep the data anonymously for further use. Slot filling based dialogue modelling was used to construct dialogue for the model, where the model attempts to determine information from what the user says, such as where the harassment took place, and use it to determine the next line of dialogue.

The chatbot uses a rule-based approach as opposed to a deep learning model to determine future dialogue. When it was not possible to determine what a user said, the user was prompted to provide the relevant information. At the end of the conversation, the user is asked to give feedback on whether the chatbot was useful. No trials were performed but various conversations were held to ensure that the rules of dialogue generation were correct and that the chatbot would prompt for further information when the user did not provide this in the initial greeting.

This chatbot demonstrates how a rule-based approach is used to generate dialogue and shows how the dialogue is interpreted through the use of slots to extract relevant information that can be used to further the conversation. This approach could be useful for SATbot to ensure we have a fixed, controllable set of dialogue that we know will be safe and provide all the relevant information about which Self-Attachment protocols that the user should attempt.

2.5.4 Summary Of Approaches

The above approaches show that current approaches often utilise a rule-based approach to determine how to respond or utilise decision trees to return answers from a fixed set of prepared utterances. The chatbots are effective at distributing assistance but can only communicate with users through a limited set of utterances that were prepared for the model to use in very specific cases, and not all the chatbots accepted language input from the users. The chatbot may have to incorporate rules to guide it in certain circumstances as demonstrated by the #MeTooMaastricht chatbot, and can develop its conversational abilities through discussing how a user is feeling to encourage self-reflection as shown with inputting emotions for ESM and EMMA. Finally, the empathetic messages of Woebot, ESM and EMMA were effective during conversations despite a limited expressiveness as a result of few options for utterances to respond to participants. These approaches influenced the development of SATbot through highlighting the effectiveness of a rule-based approach in providing empathetic messages to users in a controlled manner and achieving the aim of that chatbot.

2.6 Previous Work On Developing Dialogue

SATbot will need to be able to communicate with patients to answer their queries and assist them with the Self-Attachment Protocol. Different approaches have been explored to allow a model to produce believable dialogue, such as using knowledge or the emotion of the user. This section will evaluate some methods to determine which may be useful for the chatbot to incorporate.

2.6.1 Combining Retrieval and Generative Models

Song et al. [2] explore the combination of retrieval and generative models to produce responses to queries. To achieve this, the relevant response is first retrieved from a knowledge base, where relevance is determined by numerous features such as the word overlap ratio and cosine similarity of word embedding vectors for both the query and response from the knowledge base. When the query is passed into the knowledge base, up to 1000 pairs are returned, and the response that is most relevant is returned from the retrieval model. A BiSeq2Seq model, which is a variant of the Seq2Seq that takes in two inputs, then takes the query and retrieved response as inputs to generate a new response. A reranking system is then applied to the generated response and the original retrieved response, using the same criteria to determine which is more relevant as with responses that were retrieved from the knowledge base by the retrieval model, which was aimed to remove less relevant responses or short responses that are generated.

The retrieval system was trained with 7 million samples collected from online forums and question-answering communities, the scorer used for relevance was trained on 50,000 samples, where 10,000 of these were human-human utterances and the remainder were constructed by negative sampling, and the BiSeq2Seq model was trained on 1,500,000 samples, which were again gathered from public websites. The model was then evaluated by humans, who gave a score for responses evaluated between 0 and 2, where 0 is the lowest and 2 is the highest). Responses were also evaluated using the BLEU metric to evaluate individual words in the response, or unigrams (BLEU-1), up to 4-grams (BLEU-4). BLEU was used due to slight correlation with human dialogue in a similar domain. However, Liu et al. [38] noted that the BLEU score is inappropriate to use as a metric for open-domain dialogue generation as it greatly differs from human evaluation. Additionally, Caglayan, Madhyastha and Specia [39] determined that the BLEU score favoured texts produced by systems as opposed to human texts, which would lead to the difference between BLEU scores and human evaluation as described by Liu et al., and that single sentences used as outputs for entire test sets can receive high BLEU scores from models despite rarely being selected in practice, so therefore it is consequently less suitable for ranking dialogue that should be close to human dialogue. Therefore, comparisons will focus on human evaluation. The combined system received the highest human evaluation and BLEU-1 to BLEU-3 scores, and was only marginally worse at BLEU-4 than the retrieval model alone. Responses generated by the BiSeq2Seq model were also marginally more random and slightly longer than those generated by the Seq2Seq model.

This process was considered to use for dialogue generation for SATbot as being able to retrieve the most appropriate response for a situation and using this as a baseline for generation would give us more control over the dialogue that is produced. This is discussed further in Chapter 3.

2.6.2 Hashcode Modelling

Garg et al. [3] explore an alternative to using neural network models for developing dialogue by converting text to hashcodes and then determining the hashcode for the next response before converting this back to text. The model aims to maximise mutual information between the patient's utterance and the model's response through this method. The model was trained on three datasets: a set of therapy and counselling sessions discussing depression, consisting of 4,000 transcripts of therapy sessions; transcripts from Larry King TV interviews, which had 8,200 sessions consisting of interviews between 2000 and 2011; and a dataset of 1.3 million Twitter conversations.

To convert utterances into hashcodes, either a kernel-based method or LSTM is used to classify into hashcodes [3]. Then a Random Forest classifier is used to determine the hashcode of the corresponding therapist response. The hashcode is then converted into text either by selecting from responses from the training set or by generating using N-grams. The model was evaluated using embedding metrics described by Liu et al. [38], where it is discussed that the use of embedding metrics weakly correlate with human judgement and so are unsuitable for comparisons with other models. The developed models were evaluated by humans on how appropriate and diverse the responses were and were more appropriate and diverse than neural network models which were used as baselines [3].

While the results indicate that the model is better when trained on smaller datasets than neural network models, this does not indicate the model is better than neural network models trained on larger datasets. In addition, the hashcode modelling for dialogue does not utilise more than the last sentence to construct a therapist response, and therefore does not take the entire conversation into consideration. Therefore, this method is less suitable to apply to dialogue generation for SATbot. However, the therapy datasets may be useful as supporting data to train the chatbot. Unfortunately, we were unable to request the therapy datasets to use as training data.

2.6.3 Applying Knowledge

Dinan et al. [1] explored how to make models more knowledgeable through training them to utilise knowledge in dialogue. Wizard of Wikipedia attempts to make a model more knowledgeable through training a model to use knowledge from articles on Wikipedia. The model is trained on a dataset of dialogue between two humans, where one assumes the role of the wizard, who informs an eager apprentice about a topic that they ask about. The wizard has access to a set of Wikipedia articles that may be relevant, where the top 7 articles from this set that are related to the last two turns of dialogue between the wizard and the apprentice are provided to the wizard alongside an article relevant to the current topic, which can be used to structure their response. However, the wizard is instructed not to simply repeat information given in the articles provided. The apprentice is also tasked with keeping the conversation engaging, which results in an engaging and informative dataset that the model can train with.

The architecture for the model used is a combination of the Transformer network and end-to-end Memory Network architectures, which results in a Transformer Memory Network [1]. Two separate variants are created: a Two-stage Transformer Memory Network which uses a separate encoder after adding knowledge to the dialogue, and an End-to-End Transformer Memory Network, which uses the same encoder throughout the network.

When running the model, the Information Retrieval (IR) system is the same as what was used for the wizard [1]. The IR system extracts useful knowledge from the context of dialogue viewed so far and encodes this using a Transformer network. The most important knowledge is identified using an Attention mechanism to highlight keywords, and the most important knowledge is appended to the dialogue. The amended dialogue is then encoded by the relevant encoder depending on the architecture, and then decoded by the decoder to produce the next response from the model.

This model was compared to other models, including a variant on the Memory Network called the Bag-of-Words Memory Network, through examining which models were best at retrieving the best response from the training set, and which models were best at generating the next response [1].

In terms of automated metrics, the generative models were evaluated on their perplexity and F1 score, where the F1 score refers to how closely the response overlaps the given "gold knowledge" [1]. Retrieval models measure how often the gold response is selected out of 100 responses through measuring the recall for the top response. For the retrieval model, the Transformer Memory Network had the highest recall both for seen and unseen topics, and performed better when the gold knowledge was provided. For the generative models, both the Two-Stage and End-to-End models outperformed the other models in terms of perplexity and F1 score, where the Two-Stage model performed better when gold knowledge was not provided, and the End-to-End model performed better when gold knowledge was provided. The models were also evaluated by humans, who evaluated how engaging the models were, where a higher rating meant the model was more engaging, and also evaluated how knowledgeable the models were through the measuring of the Wiki F1 score, which measured how much the response overlapped with the first 10 sentences of the relevant Wikipedia article that the model would have access to; a higher Wiki F1 score meant the model was more knowledgeable. Only the Two-Stage model was compared due to its effectiveness when not provided with gold knowledge and outperformed the other models in terms of engagingness and Wiki F1, showing it was more knowledgeable and engaging than other models.

Using this mechanism may allow SATbot to incorporate knowledge about Self-Attachment into its conversations with humans, but this may require a similar IR system to be established.

2.6.4 Comparison and Evaluation

Figure 2.1 [1, p. 7] compares the perplexity of each model. The Two-stage Transformer Memory Network [1] has the lowest perplexity, though these models are hard to compare to the entropy metric used by Song et al [2]. The datasets used in each method greatly differ in size, with Wizard of Wikipedia using 201,999 turns of dialogue, whereas the BiSeq2Seq model was trained on 1,500,000 samples [1, 2]. However, the BiSeq2Seq model was trained to produce an utterance given the previous one, whereas Wizard of Wikipedia was trained to produce an utterance given all previous utterances, which results in a larger variance of output utterances [1, 2]. Therefore, a combination of these approaches may be most appropriate to ensure a knowledgeable but controllable response.

Figure 2.2 [2, p. 6] shows the human evaluations for the retrieval and BiSeq2Seq models, both with and without the rerank system, and Figure 2.3 [1, p. 7] shows human evaluations for the Wizard of Wikipedia models. The BLEU scores are not shown due to their inappropriateness for measuring open-domain dialogue, as described by Liu et al. [38] and Caglayan, Madhyastha and Specia [39]. Figure 2.2 also shows the percentage of generated responses that are selected over retrieval responses. The combined BiSeq2Seq model outperforms the others in human evaluation, but only marginally. However, 10% more responses are selected by the reranker from the dialogue generations with BiSeq2Seq as opposed to using a Seq2Seq model, which uses only the query to produce a response, so this approach may be suitable to use for dialogue generation while conditioning it on suitable information. Additionally, Song et al. do not describe what the model was evaluated on, so the scores are subjective and hard to compare to other papers. Furthermore, the Two-Stage Generative Transformer Memory Network, which outperformed the End-to-End Transformer Memory Network, only received an engagingness rating of 2.92. It was remarked that the Wizard of Wikipedia models that were provided knowledge would overcome issues with repetition by reusing large fragments of knowledge from the article within their responses, and would provide detailed responses without inviting further responses [1]. A psychotherapist may wish to provide information and then ask why the patient has asked for this, or ask for supporting information to determine how to proceed, but this was not the case in this study. To overcome this, a variety of datasets would need to be used to follow the methods described by Dinan et al., or combine the method with another. Finally, Figure 2.4 [3, p. 6] shows the human evaluation for modelling hashcode dialogue. When generating new responses, only 33% of responses were deemed appropriate, which is lower than the percentage of generated responses chosen when using the BiSeq2Seq model, demonstrating that this is less effective.

Model	Perplexity
Two-stage Transformer Memory Network, [1]	**46.5** for topics covered (seen), **84.8** for new topics (unseen)
End-to-End Transformer Memory Network (with knowledge dropout, [1])	63.5 for topics covered (seen), 97.3 for new topics (unseen)

Figure 2.1: Comparison of perplexity between models. Knowledge dropout is applied to the End-to-End Transformer Memory Network by preventing the model from applying attention to knowledge a fraction of the time. This was more effective than when knowledge dropout was not applied.[1, p. 7]

Model	Human Score	Generated Responses Chosen (%)
Retrieval	0.996	N/A
BiSeq2Seq	0.966	N/A
Combined Retrieval and Seq2Seq model, with response reranking	1.030	45.35
Combined Retrieval and BiSeq2Seq model, with response reranking	1.131	55.23

Figure 2.2: Comparison of human evaluation of dialogue, where a number between 0 and 2 is given (2 is best). An average is then taken from the scores given by volunteers. [2, p. 6]

2.7 Previous Work On Making Dialogue Empathetic

The dialogue generated by the model must be empathetic towards the patient as supportive dialogue will be useful in ensuring the patient continues with Self-Attachment Therapy. This section explores methods looking to make dialogue more empathetic and expand further on using knowledge to detect implicit emotions within dialogue.

2.7.1 Using Empathetic Data

One method explored by Raskin et al. [4] adjusts how models are trained through the creation of the Empathetic Dialogues dataset, which is provided to models during training so the model displays more empathetic dialogue. By fine-tuning existing models with this dataset, which means additionally

Method	Engagingness (Seen)	Wiki F1 (Seen)	Engagingness (Unseen)	Wiki F1 (Unseen)
Human Performance, [1]	4.13	11.1	4.34	10.6
Wizard Performance, [1]	N/A	43.3	N/A	43.1
Generative Transformer (no knowledge), [1]	2.11	15.3	2.54	10.1
Retrieval Transformer Memory Network, [1]	3.43	23.4	3.14	16.3
Two-Stage Generative Transformer Memory Network, [1]	2.92	30.0	2.93	26.2

Figure 2.3: Comparison of engagingness and Wiki F1 rankings, where engagingness is between 1 and 5 (5 is best) [1, p. 7]. Mean rankings are displayed in this table. The Retrieval Transformer Memory Network, where responses were retrieved from the training set, is included as a comparison.

Model	Appropriate(%)	Diverse (%)
HRED, [3]	12.2	4.0
Selection-based Hashing Model,[3]	36.8	80
Generative Hashing Model, [3]	33.8	12.0

Figure 2.4: Human evaluation of appropriateness and how diverse responses are from various models generating hashcodes of responses [3, p. 6]. Selection-based hashing model is included as a comparison.

training the model with a separate dataset, the model should adjust to new dialogue and respond to different emotions more empathetically by considering the user's emotion. The dataset consists of approximately 25,000 conversations between two people about a situation and how the situation made them feel based on a given emotion label, where neither person can see what the other's label is. The conversations are labelled with an emotion label so the model can learn about each emotion.

To evaluate its effectiveness, the outputs of Transformer models trained using different methods were compared to determine which produced the most empathetic dialogue with the most appropriate content [4]. A Transformer network was pre-trained on a dataset of 1.7 billion Reddit messages, and to compare the difference, this was compared to fine-tuning on the Empathetic Dialogues dataset. These models were also compared against separate networks that were similarly fine-tuned on the Empathetic Dialogues dataset: the EmoPrepend-1 model, which prepends the relevant emotion to each utterance and trains with this data, and the TopicPrepend-1 model, which prepends the relevant topic to each utterance and trains with that data.

The models were evaluated using automated metrics and human evaluation on their ability to retrieve appropriate responses from a set or generate new responses; the results here focus on the generative comparison, though retrieval models are later compared. Retrieval models were pre-trained on BERT to return the best response from a set based on which response y maximises the dot product with the query x, $\boldsymbol{h_x} \cdot \boldsymbol{h_y}$ for embeddings h_x and h_y corresponding to x and y, respectively [4]. Perplexity and BLEU scores were used as automated metrics, where the BLEU score is a ranking for the generation from the model. However, as previously mentioned, Liu et al. [38] and Caglayan, Madhyastha and Specia [39] both noted that the BLEU score is inappropriate to use as a metric for open-domain dialogue generation, so comparisons will focus on perplexity and human evaluation. Human evaluation consisted of evaluating the models on three areas: how empathetic the responses were; whether the content of the response was appropriate; and whether the response was coherent [4]. The Transformer model that was fine-tuned on Empathetic Dialogues had the lowest perplexity and displayed the most empathy and relevance. However, TopicPrepend-1 was ranked as more fluent. All models fine-tuned on Empathetic Dialogues outperformed the pre-trained model in all human evaluated categories and perplexity, demonstrating the dataset's effectiveness at making dialogue more effective and empathetic. This dataset will be useful when developing dialogue for the chatbot.

2.7.2 Using Multi-Type Knowledge

Li et al. [5] extend on this area through the introduction of the "multi-type knowledge aware empathetic dialogue generation framework" [5, p. 2] (MK-EDG), which aimed to utilise knowledge to determine implicit emotions in conversations so a model could provide an appropriate response. The framework consisted of three steps. First, an emotional context graph is created from the dialogue history and relevant knowledge. Then, the Transformer encoder would encode data from the graph and determine the relevant "emotion signal". This is then passed to Transformer decoder layers which use an alternative attention mechanism called an Emotional Cross-Attention Mechanism, which determines emotional dependencies in the utterance to use while generating the knowledge.

The framework was trained on the Empathetic Dialogues dataset and compared against other models, including the Transformer network and EmoPrepend-1, by evaluating the responses generated from these frameworks [5]. Automated metrics that were evaluated were the emotion accuracy, which was whether the emotion of the response matched the emotion label given to each utterance in training, and perplexity. Similarly to the evaluation of Empathetic Dialogues [4], human evaluation consisted of three categories: how empathetic the model was (known as empathy), how appropriate the content was (known as relevance) and how coherent the response was (known as fluency) [5]. The MK-EDG outperformed all other models in all areas except fluency, where EmoPrepend-1 had the highest fluency.

Notably, the results of this paper differ from the first evaluation of Empathetic Dialogues due to a different number of parameters used: this paper only uses 31 million parameters whereas the first paper uses 85 million. Additionally, the first paper fine-tuned on the dataset whereas this paper pre-trains on it. Using more parameters may have resulted in greater results within this experiment. Regardless, this paper demonstrates the framework's effectiveness, which could be incorporated into the chatbot.

2.7.3 Evaluation

Figure 2.5 [4, p. 8][5, p. 6] shows a comparison between the two papers. The perplexity of all models is very low, so a combination of this and previous papers such as Wizard of Wikipedia would prove to produce a more well-rounded model. Furthermore, human evaluation produced high results for empathy, relevance and fluency when fine-tuning on the Empathetic Dialogues dataset. Additionally, a Transformer Network was used in these papers, so this architecture may be useful to utilise. Also, the first paper had models that used approximately 85 million parameters, whereas the second paper only used up to 31 million parameters, which may have impacted the results [4, 5]. Finally, the retrieval model pre-trained on BERT and prepending with the topic or emotion outperforms all generative models, suggesting a retrieval model may be a suitable alternative for dialogue generation to retain high metrics while also producing controllable dialogue. As we would also control the set of responses, this would prevent any unsafe responses being produced by SATbot.

2.8 Previous Work On Constraining Output

As with human psychotherapists, it is important for the chatbot to produce dialogue that is constructive and does not offend the patient, which would result in the conversation deteriorating and the patient becoming disconnected with the treatment. Various approaches have been considered, and this section will discuss these approaches.

Model	Perplexity	Empathy	Relevance	Fluency
Retrieval w/BERT, EmoPrepend-1 (Prepending with topic, fine-tuned with ED, [4])	N/A	3.93 +/- 0.12	3.96 +/- 0.13	4.54 +/- 0.09
Retrieval w/BERT, TopicPrepend-1 (Prepending with topic, fine-tuned with ED, [4])	N/A	**4.03 +/- 0.10**	**3.98 +/- 0.11**	**4.65 +/- 0.07**
Generative, pre-trained on Reddit dataset (no fine-tuning on Empathetic Dialogues (ED)), [4]	27.96	2.31 +/- 0.12	2.21 +/- 0.11	3.89 +/- 0.12
Generative, pre-trained on Reddit dataset, fine-tuned on ED, [4]	**21.24**	**3.25 +/- 0.12**	**3.33 +/- 0.12**	4.30 +/- 0.09
Generative, EmoPrepend-1 (Prepending with emotion, fine-tuned with ED, [4])	24.30	3.16 +/- 0.12	3.19 +/- 0.13	4.36 +/- 0.09
Generative, TopicPrepend-1 (Prepending with topic, fine-tuned with ED, [4])	25.40	3.09 +/- 0.13	3.12 +/- 0.13	**4.41 +/- 0.08**
Generative, EmoPrepend-1 (Prepending with topic, trained with ED, [5])	38.30	3.23	3.51	**3.74**
MoEL (Transformer Network that probabilistically uses responses from different decoders, trained with ED, [5])	38.04	3.37	3.78	3.64
MK-EDG (trained with ED, [5])	**34.85**	**3.49**	**3.91**	**3.65**

Figure 2.5: Perplexity and human evaluation of models on how empathetic they are, fluent they are and whether the content produced is suitable/relevant. Models are divided by section. [4, p. 8][5, p. 6]

2.8.1 GeDi

One approach developed by Krause et al. [6] to constrain output uses generative discriminators, known as GeDis, to control the output of a model. Discriminators are used to determine whether the output of a model is good or bad based on how they are trained and have previously been used to assist dialogue generation of models [40]. GeDis were designed to be more efficient than previous methods, and uses Class-conditional language models (CC-LMs) as generative discriminators for other models [6]. CC-LMs aim to control text generation by conditioning on a designated control code, which is an attribute that represents a class. CC-LMs predict a probability distribution to describe a specific attribute of the text, where sequences in training are paired with their corresponding attribute.

Krause et al. [6] focus on detoxifying outputs produced by a model. This is achieved by fine-tuning the GPT2-medium model to use as a CC-LM with GeDi training which is used to guide generation from the GPT2-XL model. The GeDis were trained as toxicity classifiers by using the Jigsaw Toxic Comment Classification Challenge Dataset to be able to identify whether content is toxic based on whether it is labelled as toxic. The model was compared against the base GPT2-XL model on toxicity, which was ranked from 1 to 3 and a lower toxicity ranking results in less toxicity, and how linguistically acceptable the output was, which is ranked from 1 to 4 and a higher result is more acceptable. Through using a GeDi-trained CC-LM to detoxify outputs, the toxicity rating was lower and the outputs were also considered to be more linguistically acceptable. Therefore the use of a CC-LM where GeDi training has been applied could prove useful for SATbot to produce outputs that are suitable for conversations with patients.

2.8.2 Plug-and-Play Methods

Another approach by Madotto et al. [7] involves the addition of plug-and-play methods to existing models to be able to adjust the output distributions of the decoder without adjusting the parameters of the model. This would allow dialogue generation to be constrained without restricting the model. Plug-And-Play methods are added through adding Residual Adapter modules, which can be trained to steer the output distribution towards the desired output by training parameters of these modules, to the end of existing Transformer networks [7, 41].

In experiments, the DialoGPT model was used as a base architecture, which is a "large pre-trained model trained on 147 Million multi-turn dialogues from Reddit" [7, p. 5]. The DialoGPT model was extended through adding a Residual Adapter per style that would be considered, such as making the response positive or negative. The addition of Residual Adapters per style was compared to the DialoGPT model without augmentations or with other methods such as the PPLM method, which was another method used to control generated text. The responses were scored by classifiers for each style that was being considered, and their perplexity was measured. While the new method did not achieve the lowest perplexity, it received the highest score across all styles and from the external classifier. The responses were also evaluated by humans on their humanness, which reflected how fluent and diverse the response was, and their attribute consistency, which reflected how closely the dialogue reflected the style that was the intended focus. No method consistently outperforms another model in terms of humanness, but the new method is best at outperforming other models in terms of attribute consistency. Therefore, the use of Residual Adapters could be utilised to support a model for SATbot which was pre-trained.

2.8.3 Safety Recipes

Another approach by Xu et al. [8] evaluates how to best constrain the output of a model, dividing possible methods across four categories:

- Unsafe Utterance Detection, which is the process of detecting unsafe utterances in dialogue by training classifiers to identify these,
- Safe Utterance Generation, which is the process of generating only safe utterances,
- Sensitive Topic Avoidance, which is the process of avoiding sensitive topics such as politics or religion, and
- Gender Bias Mitigation, which is the process of using gender-neutral language within dialogue from the model.

Within Unsafe Utterance Detection, the Bot-Adversarial Dialogue Safety (BAD) method is proposed, which involves training a model and then have a human engage in conversation with the model, who attempts to provoke unsafe utterances from the model [8]. This data is then used to fine-tune the model to ensure that the model would not be susceptible to the same provocations at test time. Another method is the use of classifiers to classify utterances as safe or not safe, which first used data from the Wikipedia Toxic Conversations dataset and was extended to collect data using the "Build-it, Break-it, Fix it" method, which is described as collecting training examples where a human engages in conversation with a model and succeeds in making the model classify an unsafe response as safe. A semi-supervised approach was also considered where gold knowledge was provided during training. Another method is the use of two-stage models, which detect whether an utterance is unsafe and if so either attempt to change the subject using a "non-sequitur" or use a generic safe response. BAD can be used within a Two-Stage model for extra safety.

Within Safe Utterance Generation, Baking-in the Safety layer is proposed, which involves labelling training data as either safe or unsafe so the model becomes aware of what is unsafe and only generates safe responses or changes the subject, which is known as a non-sequitur [8]. This would be preferable to using a method such as preprocessing data to remove unsafe utterances as by preprocessing, the model would not know how to react when faced with an unsafe utterance. Other methods that are considered include using methods to condition on safety or style of text to only generate either safe responses or responses in a certain style such as positivity, which was shown to be more effective than simply generating either safe or unsafe responses, and adjusting the search for the next response when decoding to only return safe responses by comparing against an unsafe word or n-gram list approach.

For Sensitive Topic Avoidance, specific topics can be chosen to be avoided [8]. A multi-class classifier was used to predict the topic label for an utterance, and if a sensitive topic was encountered then a safe response or non-sequitur would be generated if using this within a two-stage model. For Gender Bias mitigation, the model can be trained using gold knowledge to divide utterances based on whether they are biased towards either female or male, or biased towards both or neither. The model then aims to generate utterances biased towards neither gender.

Within experiments, a Seq2Seq Transformer architecture is utilised as the model, which is compared to the DialoGPT model and the GPT2 (Large) model [8]. Models were evaluated on their perplexity and F1 overlap with the provided gold knowledge to determine the quality of the dialogue produced, and classifiers from the ParlAI dialogue platform [42] were applied to determine the safety of dialogue produced. Human evaluation was also performed, where the responses of each model were generated and humans were asked to say which they preferred out of a pair presented to them [8]. Models were compared to the BST 2.7B model for this metric. Humans were also asked to judge the safety of responses from the BAD method. This section will focus on the results for the BAD method, including applying the method to a Two-stage model, and Baking-in the Safety layer.

The classifier trained on the BAD dataset produced similar results to classifiers trained on alternative methods when tested on datasets such as the Wikipedia Toxic Comments dataset but outperformed other classifiers when tested against the BAD dataset. When applied to a Two-stage model, engagingness is similar to results produced from the BST 2.7B model, but the safety classifier received a higher percentage of OK results compared to other safety classifiers [8]. When the BST 2.7B model was fine-tuned with a dataset where safe responses are baked-in, which means unsafe responses were replaced either with safe responses or non-sequiturs, the model was similar in terms of engagingness to the baseline BST 2.7B model, and safer than other models when comparing how often words from an unsafe word list were used, how often a classifier classified a result as unsafe and how often safe responses were triggered on a dataset gathered from Reddit and the ConvAI2 dataset [8]. However, human evaluation of these models resulted in a far lower percentage of OK results compared to applying BAD (94.4% in BAD vs 68.3% with non-sequitur responses from a fine-tuned BST) [8].

These methods could be combined to suit the needs of SATbot. In particular, the BAD safety method could be used to prevent cases where a patient attempts to force SATbot to not provide a suitable response.

2.8.4 Evaluation

Figure 2.6 [6, p. 10] shows the toxicity and linguistic acceptability of dialogue produced by various models, including one guided by a GeDi. This demonstrates how a GeDi guide reduces toxicity. Figure 2.7[7, p. 5] shows how control of style is greatly improved when adding a Residual Adapter per style that can be controlled on and demonstrates a lower perplexity as well. Furthermore, Figure 2.8[8, p. 17] shows how the percentage of OK responses greatly improves when applying BAD or Baking-In the Safety Layer. In addition, Figure 2.9 [8, p. 18,20] shows that the perplexity of models trained on the ConvAI2 dataset is far lower than other models considered, which also implies the BST 2.7B model fine-tuned with BAD would also have a very low perplexity if trained on the ConvAI2 dataset. Therefore, the use of BAD, the ConvAI2 dataset and BST 2.7B model will likely produce the best results for the chatbot, but to control specific styles, the use of Residual Adapters may be best to ensure we always receive a certain style, such as positive responses. Notably, another model was tested to control style alongside the BST 2.7B models, but the percentage of OK responses was only 60%, which implies Residual Adapters may be more useful for controlling style while retaining safe dialogue. This would need to be tested on a larger model such as BST 2.7B to determine the best result.

Model	Toxicity	Linguistic Acceptability
GPT2-XL	1.45	3.23
GeDi-trained guide	1.17	3.44
Generatively trained guide	1.13	3.25

Figure 2.6: Human evaluation of models on how toxic their output was (between 1-3, 1 is best) and how linguistically acceptable their output was (1-4, 4 is best). [6, p. 10]

Model	Perplexity	Positive	Negative	Business	Sci/Tech	Sport
DialoGPT	**39.60**	65.67	19.40	17.41	91.04	27.86
DialoGPT with Residual Adapter per style	41.57	**93.03**	**73.13**	**68.66**	**99.00**	**83.06**

Figure 2.7: Automatic evaluation of models on how effective they are at producing an expected style. A lower perplexity is better; for all other categories a higher score is better. [7, p. 5]

Model	OK Responses (%)
DialoGPT	52.8
BST 2.7B	55.0
BST 2.7B + BAD	94.4
BST 2.7B + BAD + Topic Classifier	96.6
BST 2.7B + Baking-In Safety Layer (Non-sequitur responses, fine-tuned)	68.3
Controlling Style (Calm, 400 million parameters)	60.0

Figure 2.8: Human evaluation of models on percentage of responses deemed safe [8, p. 17].

Model	Perplexity
400 million parameters (Non-Sequitur)	18.2
BST 2.7B	8.8

Figure 2.9: Perplexity of models, trained on the ConvAI2 dataset. Non-Sequitur means the model is trained to produce non-sequitur responses if it deems a response to be unsafe[8, p. 18,20].

Chapter 3

Design Choices

Throughout this project, we investigated numerous approaches to determine what would be most appropriate for SATbot to utilise. Ultimately, we determined that, as a result of the limitations of current technology, a rule-based model would be best for the chatbot to ensure responses were appropriate both in tone and content for the context of delivering Self-Attachment protocols. We later created a dataset from the results of a survey we produced so the chatbot would be able to choose from a set of responses at each point, and this is described in Chapter 7. This chapter describes what we required for SATbot and the design choices that we considered to reach this stage.

3.0.1 Limitations of Current Chatbots

We briefly discuss some limitations of current approaches that we aimed to overcome during this project. Both EMMA and Woebot only provide activities based on the emotion provided by the user, and it is harder to choose an emotion for EMMA as participants must select from a two-dimensional grid of emotions. The #MeTooMaastricht chatbot additionally is very rigid in the dialogue produced in that it does not adjust its dialogue based on the user's emotion.

3.0.2 Requirements

For SATbot to be suitable for patients suffering from depression or anxiety, any response given needed to be safe and non-toxic. This ensures that the chatbot would not say anything that could negatively affect the user through encouraging negative thoughts or processes, such as suicide. While SATbot was only tested on the non-clinical population in the trial we undertook, future trials would test participants suffering from depression or anxiety. Therefore, it was crucial that from an early stage, the dialogue produced was safe. We also aimed for the chatbot to produce empathetic and engaging responses as by being more supportive, the patient will be encouraged to continue with Self-Attachment therapy. This would also allow the user to engage more with the chatbot. However, this was secondary to producing safe responses.

From the analysis of therapy sessions discussed in Chapter 2.4, we identified that responses from the chatbot should be engaging and not repetitive. This ensures that the user is more engaged with the conversation. The chatbot should also be adaptable to the needs of the conversation and quickly assist the user by suggesting appropriate Self-Attachment protocols. Encouraging change within the user was also important to shift the user away from negative emotions and towards positive emotions. Receiving feedback was also discussed to improve a user's emotion by reflecting on what they have achieved.

To overcome the limitations of the current chatbots we discussed, we aimed for SATbot to have various rules that suggest different Self-Attachment protocols so it is more fine-grained than current approaches, which only provide differing protocols per emotion. We would only differentiate between positive and negative emotions when deciding the next rule as we would determine different Self-Attachment protocols regardless of the emotion specified due to the relevance of these protocols for users experiencing various negative emotions. We would also produce dialogue that is less rigid than dialogue from the #MeTooMaastricht chatbot. We also aimed to make emotion selection easy by allowing users to pick from 6 emotions compared to the two-dimensional grid used by EMMA. We added a simple emotion classifier as an extension to infer this through keywords from the user's input to further ease this process, and we discuss this in Chapter 4.

We also wanted the model to be easily integrable so it could be utilised within an existing VR platform. For this project, a separate platform was developed with a text interface that would allow users to communicate with the model before choosing options to answer questions from the chatbot to receive appropriate suggestions. This could be adapted for the VR platform in future work.

We considered various approaches which utilised neural networks, described within Chapter 2, and these will be discussed.

3.0.3 Models Examined

From an early stage, we intended to utilise a rule-based model to encode the rules of Self-Attachment protocols and enhance the expected response through a Machine Learning component, but this later changed to solely be a rule-based model due to the limitations discussed below.

Chapter 2.6 describes methods that were researched and considered for generating open-domain dialogue, either using an expected response and the query as an input or only using the query as an input to generate an appropriate response. We first examined whether the Empathetic Dialogues dataset and method described by Raskin et al. [4] could be utilised through using a pre-trained variant of BERT and fine-tuning on the Empathetic Dialogues dataset to generate appropriate and empathetic responses. We chose the approach by Raskin et al. over the approach by Li et al. [5] as the results of Raskin et al. outperformed the use of the MK-EDG framework. We aimed to extend this to make responses more knowledgeable with the Wizard of Wikipedia dataset and approach described by Dinan et al. [1]. A large neural network such as BERT was needed as a model to have a sufficient understanding of the language so it could be fine-tuned to produce coherent and appropriate dialogue. This ruled out the hashcode modelling method proposed by Garg et al. [3], which did not have the depth of understanding required, and we focused on neural approaches for this problem.

We determined through preliminary experiments that without encoding a means for the model to correctly respond through a large dataset of conversations involving Self-Attachment, which does not exist, a model fine-tuned on Empathetic Dialogues alone would be insufficient for the approach we intend. We also considered using BioBERT as a pre-trained model due to its greater understanding of potentially relevant domain-specific terms compared to BERT but determined that while the model's understanding of relevant terms would be greater, without a relevant dataset of specific cases, using this model would still be insufficient. In addition, we were unable to acquire related therapy datasets relating to mental health due to the owners' inability to share private patient data outside of their usage. The data could have been used to constrain the model to produce domain-related dialogue, even if the model would not possess relevant knowledge of Self-Attachment, which may have been easier to fine-tune to Self-Attachment concepts. Therefore, we considered developing a survey to gather a small dataset to have a selection of responses to use for the model, either to assist generation or to retrieve from.

Krause et al. [6], Madotto et al. [7] and Xu et al. [8] all developed methods to constrain outputs through adding additional components to the network or otherwise designing the network or data to reduce the number of unsafe responses. However, we determined that even if we used these methods, open-domain dialogue generation would be unsuitable for SATbot due to the remaining possibility of creating unsafe responses that could potentially harm participants, particularly if they suffer from depression or anxiety. We also wanted to ensure that responses were informative and that the chatbot operated similarly to a psychotherapist, providing the information necessary for a participant to take part and maintaining this role throughout the conversation.

To overcome the problems of generating dialogue, we considered the generative method described by Song et al. [2], combined with the retrieval approach described with the Empathetic Dialogues dataset by Raskin et al. [4], which would involve selecting from a set of responses and then generating an improvement, using the query and retrieved response as inputs. However, this again introduced uncertainty and may produce unsafe results, so this was not used. We then explored using only a retrieval network with the results of the survey we have gathered. A retrieval network was ranked marginally worse by humans than the combined model utilised by Song et al. when examining responses and outperformed generative approaches when using a Transformer network for retrieving appropriate responses, as Raskin et al. described. Using a retrieval network required an appropriate query leading to the response that we intended to use, even if we defined the base prompts manually. Therefore, we determined that, due to current limitations of data and technology, even with the survey we had gathered, a rule-based approach would be most appropriate for the chatbot.

3.0.4 Chatbot Interface Design

Once we determined that we were using a rule-based model, we then considered the use of Rasa, an open-source library that was first introduced by Bocklisch et al. [43]. Upon receiving a new message, Rasa interprets the input message with a module for Natural Language Understanding (NLU), which can determine and classify the intent of the user as well as user-defined entities that correspond to specific sections of sentences that the user can train on. A Tracker maintaining the conversation state will then be notified of the new message from the user. The Policy then receives the current state of the tracker and uses this to determine the next action, based on features such as previous actions and the user's intent. The action is executed, which updates the Tracker state by logging the action, and returns a message to the user. Rasa could provide appropriate actions for user inputs and allow us to determine the user's intent, but this required a large training set to train the NLU module. Handcrafting the rules ourselves would prove to be clearer and achieve the same aim, so we decided not to use Rasa, and instead used the react-chatbot-kit library, which will be discussed further in Chapter 4.

Chapter 4

Rule-Based Model

In this section we describe the rule-based model that determines how SATbot responds to users.

4.1 Conversation Flow

Figure 4.1 describes how a conversation with the user goes. First, the user is prompted to provide some information about how they are feeling, including relevant situations that led to the emotion the user currently feels. Then, the model will utilise a keyword classifier to determine which emotion is most likely from the user out of 6 possible emotions: Happy, Neutral, Sad, Angry, Anxious and Scared. If the model determines that the user is experiencing any of these, it will ask the user if they are experiencing this. Otherwise, the user will be prompted to choose from the 6 emotions, which appear as buttons.

If the user picks either Happy or Neutral and the user has not yet attempted 10 protocols, SATbot will provide the user with suggested protocols randomly. If the user has attempted 10 protocols then SATbot will suggest protocols that have been attempted least frequently so the user is more well-rounded in terms of growth and development, and becomes more able to regulate their emotional state.

If the user picks a different emotion, SATbot will suggest that the user attempts one of several protocols intended to contain negative emotions. If the user chooses to do this, they will choose one of the protocols to attempt. Otherwise, a series of questions will be asked to determine whether the user was affected by an event and whether this event was recent or distant. If the user says their emotion is affected by a distant or recent event, the model would suggest a protocol to assist with developing a relationship with the childhood self while recalling details of the recent or distant event. Further questions would then be asked to determine whether specific Self-Attachment protocols addressing certain cases are also relevant for the user. These questions are presented in the order of priority determined by a second keyword classifier, which identifies relevant words from the user input corresponding to each question and prioritises questions in order of which has the most associated words so that a user discussing a problem with family members is immediately asked whether they are undergoing a personal crisis. Questions are asked in order until the user says Yes to one of these questions, at which point the protocols corresponding to that question are added to the list of suggestions and the chatbot will move to present these suggestions. This allows for the question that is most relevant to be asked first. If the user says no to all of these suggestions, we provide Protocols 13 and 14 by default, which are suited to changing perspectives to mitigate a negative emotion and maintain a positive emotion to overcome the "psychological abyss" that the user finds themselves in.

Once the protocols are determined, the chatbot will present these as a list of suggestions for the user to attempt. The user will attempt one of these and then give feedback on whether it improved their mood, made it worse or had no effect. Then, the user will be prompted to restart questioning, choose from follow-up suggestions that include protocols that follow other protocols or the remaining suggestions that have not been chosen or stop.

Example conversations can be found in Appendix D.

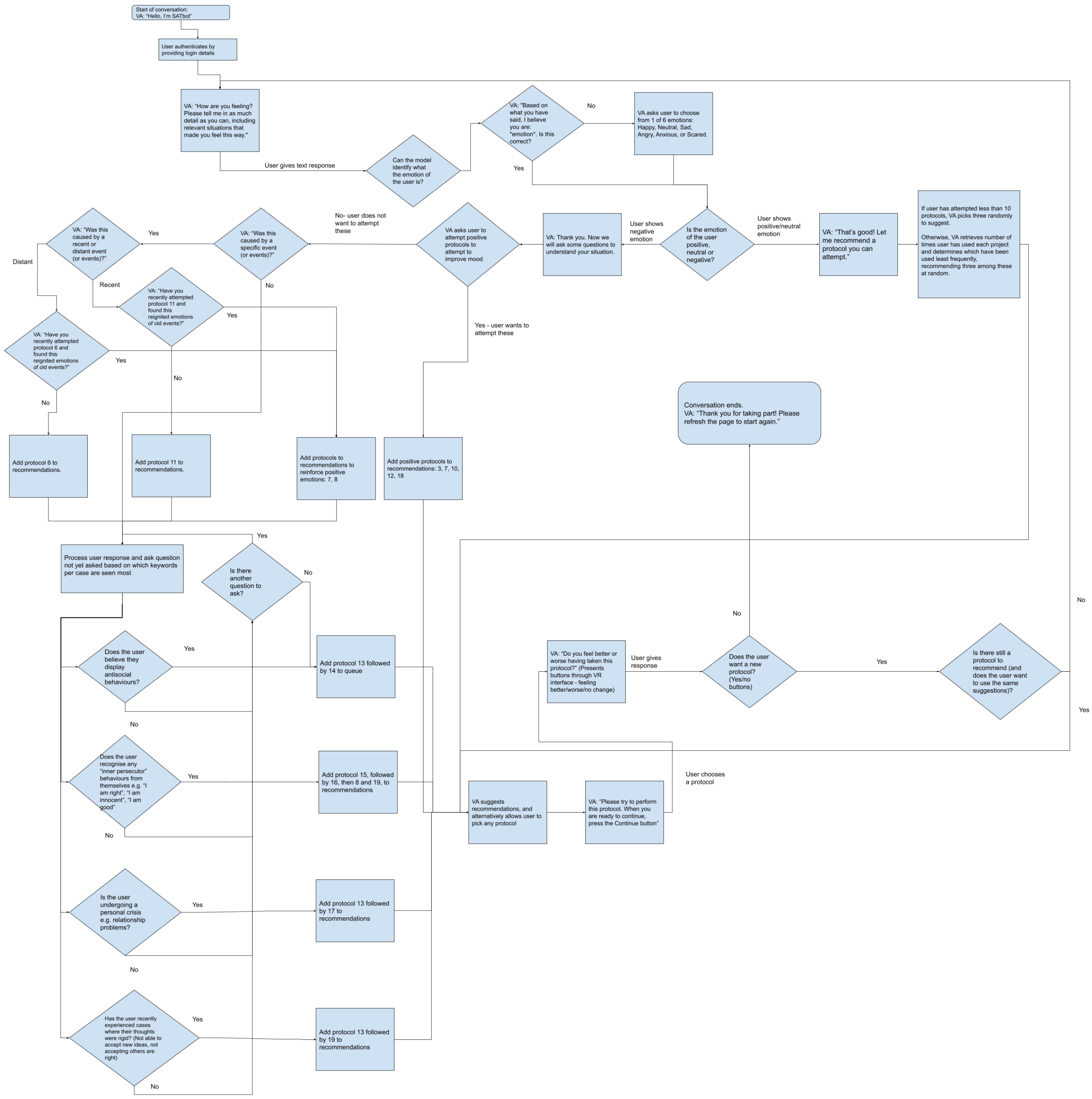

Figure 4.1: Flow chart detailing the process SATbot follows in conversations with users.

4.2 Technology Used

As part of this project, we developed a platform consisting of a web interface using JavaScript with the React framework, an API consisting of Python with the Flask framework, and the rule-based model, which was coded in Python.

For the rule-based model, we chose to use Python (v3.8) due to two reasons: not only does Python contain many libraries such as NLTK which make it suitable for NLP tasks, or PyTorch for ML tasks, but Python is also well suited for web development as it is supported by frameworks such as Flask and Django, which make this easy to embed the model in a web server. These two benefits mean that not only will Python be suitable for this project, but it will be easily extensible to future NLP and ML tasks due to the variety of libraries available and can easily be integrated into a web server in future projects which can be contacted by other platforms that use the model, such as a mobile platform or the existing VR platform.

For the platform itself, we chose to use AWS Elastic Compute Cloud for a cloud instance to host the model and web interface during the project. We used a single instance to host the web interface, model and embedded database. We used SQLite as the database as we did not need a larger database instance on AWS to contain all of the information from participants using the platform during the trial. This allowed everything to be hosted on a single instance. On the instance, we used an Nginx server to host the web interface and act as a proxy to the Gunicorn server hosting the backend content.

Finally, for the web interface, we used JavaScript with the React framework as the framework supported many libraries, including react-chatbot-kit, which we utilised to create the interface for the chatbot. As the framework utilises a component-based design to structure web pages, which made it easy to define buttons to use as options and link these to the processes tracking the current state of the user.

4.2.1 Interface

The web interface was constructed using the react-chatbot-kit library [44]. The interface allows the user to enter a text input or select from a series of options presented by the model. Figure 4.2 shows part of a conversation through the interface.

Within the interface code, when the user provides their text input, this is parsed to determine how to respond based on the current state. Certain conditions will cause the interface to trigger an API request to ask the rule-based model how to proceed. An action provider handles the API requests and the corresponding response and updates the interface by adding the next message to the chatbot's state, which causes it to render on the user's screen.

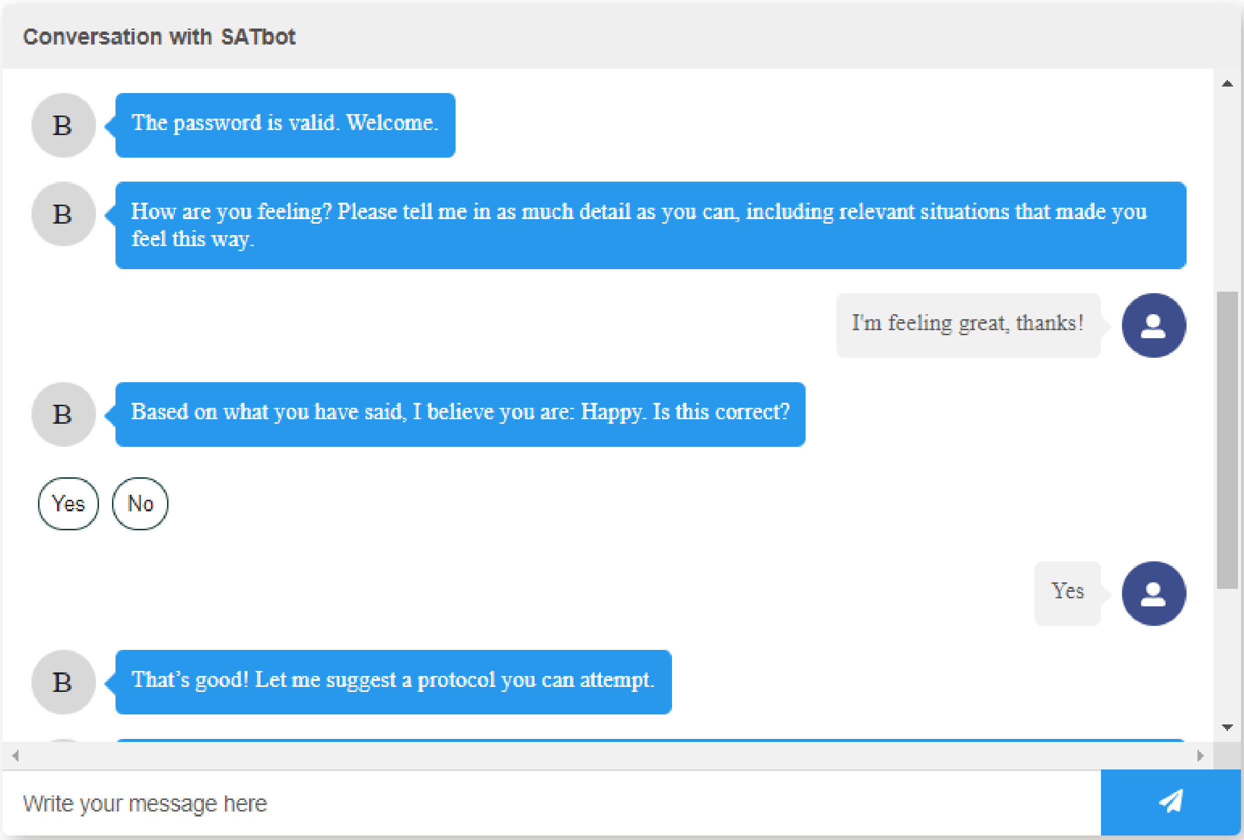

Figure 4.2: Example conversation shown through interface. The user will be able to enter their own text or choose from options presented (which appear as their text).

4.2.2 Interface to Model

We briefly describe the process from interface to model to understand the model in the context of the interface.

When the user selects an option or provides a text input, a request is sent to the API to determine what the next message is for the interface. The API checks which user the request corresponds to, where users are provided with their user ID after logging in, which for this trial was also performed through the chatbot. The current choice is saved in the database, and information for the next choice is determined and returned to the user. Some content such as the current choice is saved in memory so the user does not need to access the database to retrieve the current state.

4.2.3 Rule-Based Model Structure

The rule-based model is defined in Python as a dictionary with 24 keys. The keys represent the different positions of the model that the user may find themselves in, such as being asked to say how they feel or being provided suggestions. Each key maps to a nested dictionary, which contains three keys: "model_prompt", which represents the prompt for the choice corresponding to the key, "choices", which itself maps to another dictionary where each key maps to the next stage of the model that the choice would lead you to, and "protocols", which maps to another dictionary where each key represents the choice made by the user and the corresponding value represents the protocols that are added to the list of suggested protocols that will be presented when the user reaches the stage of the conversation where the model suggests appropriate protocols. Suggestions are stored in a list of queues, where each queue consists of a set of protocols to be undertaken in sequence, and only the first item of each queue is presented when querying for suggestions. When an option is selected, it is removed from the queue so the following item can be suggested.

4.2.4 Classifiers

Two keyword classifiers are used as part of this model. One classifier is used to determine the most likely emotion of the user out of the 6 options we present: Happy, Neutral, Sad, Angry, Anxious, and Scared. The other is used to determine how to order questions to determine specific cases are relevant to the user based on the event they describe. Both classifiers operate by using a list of for and against keywords. These lists are augmented with synonyms for each word in the list, which are determined by downloading the wordnet corpus using the nltk library. Each synonym for words in the list of keywords supporting the emotion is added to the list of supporting keywords, and each antonym is added to the list of against keywords. The number of for keywords in the user input is subtracted from the number of against keywords and the result is the score given to the specific emotion to choose or the specific question to present first. The emotion with the highest score is presented as a suggestion to the user, though in the case of ties, the earliest emotion is chosen (where emotion scores are calculated in the order of emotions: Happy, Neutral, Sad, Angry, Anxious, and Scared). For the classifier of questions, the question with the highest score is presented to the user first, and then the question with the next highest score is presented if the user says No to the previous question until all questions have been asked or until the user says Yes.

Chapter 5

Non-Clinical Trial

5.1 Ethical Discussion

As part of this project, we performed a non-clinical trial with humans using SATbot to assess its effectiveness in assisting users to undertake Self-Attachment protocols. We wanted to determine how the model affects the delivery of Self-Attachment, and we also wanted to evaluate the empathy of the model. Participants consisted of individuals in Canada who had previously undertaken a trial involving the evaluation of Self-Attachment protocols who had given consent to be contacted again for future trials and were therefore familiar with the Self-Attachment protocols that they practised.

We collected personal data, such as the person's name, age and gender to analyse the conversations per demographic. Furthermore, we asked volunteers to complete questionnaires asking about depression and anxiety, where those displaying symptoms of either did not take part in this trial. The focus of this trial was to examine the effectiveness of the model on the normal population before it could be evaluated on individuals that have depression or anxiety.

We also acquired ethical approval from Imperial College to undertake the trial. While the trial of the chatbot itself was not therapeutic, we needed to determine the health of any participants through the information described to ensure that we were testing on the non-clinical population and that no participants were suffering from severe depression or anxiety. We asked participants to complete a questionnaire answering a series of questions to determine whether they were suffering from depression or anxiety. All who volunteered to take part were eligible for the trial. We requested that participants would interact with the chatbot through the web interface for 30 minutes each day for 5 days. Due to time constraints resulting from requiring significant time to receive ethical approval, we asked participants to interact with the chatbot for at least 2 days within a 5 day period.

Data was stored within a database in the UK. Access to data was restricted to specific members of the Algorithmic Human Development Group at Imperial College London. If a participant chose to withdraw from the trial, their data was immediately deleted from the database. Data provided by participants will only be held for analysis and will be deleted by the end of September 2021, following the completion of a subsequent MSc project. Conversational data was similarly held within a database in the UK to analyse the effectiveness of the model throughout conversations, but this will be restricted similarly to other data provided by participants. Participants were not monitored in real-time during the conversation with the model as conversations took place remotely. We collected approval from all participants to use their personal data, and have anonymised any data used in the report.

We expected to require approval, ethical or otherwise, for any datasets we intended to use that were related to counselling or other forms of therapy such as Cognitive Behavioural Therapy. However, the datasets which we requested were unavailable due to the terms and conditions of these datasets, which meant no such dataset was used.

As part of this project, we also gathered a small dataset using Amazon Mechanical Turk. The dataset aimed to guide the development of the model in future iterations by providing the model with a set of empathetic prompts to choose from in conversations with users. Survey responses that were suitably empathetic from this dataset can then be utilised in conversations to ask users about cases that can be used to determine the most appropriate Self-Attachment protocols. The task was advertised on the Amazon Mechanical Turk website, and users of Amazon Mechanical Turk could choose to take part in this survey, for which they received \$5 as financial reimbursement. We asked 100 users to take part in this survey and did not gather information from these users so they remained anonymous, other than their Amazon Mechanical Turk worker ID so we could confirm that a worker had taken part in the survey and provided meaningful results. Nonsensical or incomplete results from workers were discarded, and workers were not reimbursed, per the terms of Amazon Mechanical Turk. We received ethical approval from Imperial College to gather and use this data, and the survey included the forms that users were required to agree to progress with the survey.

Chapter 6

Evaluation

6.1 Comparison with Prior Work

We evaluate how effective SATbot was at delivering Self-Attachment protocols, compared to how effective other chatbots were at performing their roles to assist in delivering therapy.

6.1.1 Woebot

We begin by discussing Woebot, the chatbot trialled by Fitzpatrick et al. [12] that delivered CBT within an IM application to participants. Both Woebot and SATbot tested participants of the non-clinical population. Both Woebot and SATbot provided specific content to users based on emotions but where Woebot provided content for users experiencing specific emotions, SATbot provided suggestions to users based on specific events, such as whether the participant recently disagreed with a friend or family member. SATbot provides questions to users based on these criteria if the user indicates through text or selection that they are experiencing a negative emotion, such as anger or sadness, but does not differentiate between negative emotions when asking questions. However, SATbot will use the keyword classifier to prioritise which questions to ask first. Woebot responded empathetically when providing appropriate suggestions, but our prompts for SATbot did not directly address the emotion that the user felt. Therefore, SATbot displayed less empathy, despite slight adjustments in dialogue for when the user presented positive and negative emotions, which allowed the chatbot to be more supportive towards users with negative emotions.

In terms of content, Woebot provided links to videos describing core concepts of CBT and would also suggest word games to teach participants about more complex concepts. Additionally, Woebot provided an onboarding session to new users to briefly describe CBT and introduce them to the chatbot. While SATbot did not directly provide participants with links to content, participants were provided with a description of workshops and protocols separately, as shown in Appendix B, and these descriptions could be integrated into the chatbot in the future. However, SATbot was designed for participants who have already undertaken the 8-week workshop and are familiar with Self-Attachment protocols, so the chatbot would not need to provide more than short descriptions to remind participants about protocols to assist them in undertaking these. Furthermore, as participants were already familiar with Self-Attachment protocols and the process was described through the trial documents, an onboarding process was not required, though an onboarding process could be added to introduce new users.

Woebot provided reminders for participants to check in regularly and reminded participants to start conversations with the chatbot by sending personalised messages. SATbot did not need to remind users to complete tasks as participants were asked to undertake protocols immediately and then say whether they felt better or worse having attempted the protocol. It was both possible and necessary with Woebot to remind users due to users interacting with the chatbot through a mobile application. SATbot was designed to be integrated within a VR platform where reminders will be unnecessary. However, if SATbot was integrated into a mobile application, a similar reminder system could be established. Finally, both applications had a simple text interface to communicate with the chatbot, though SATbot additionally allowed the user to pick from a selection of options throughout conversations. Both approaches reach a solution quickly to deliver appropriate content, but due to addressing a larger number of cases than Woebot, SATbot must ask more questions to the user to determine which Self-Attachment protocols are most appropriate. However, based on the interactions with users we discuss below, we intend to introduce certain protocols sooner to provide more immediate assistance and then provide follow-up suggestions if these do not aid the user.

6.1.2 Other chatbots

We briefly compare SATbot to the other chatbots discussed in Chapter 2.5. EMMA, which extended ESM, gathered feedback from users about their current emotion by asking them to select a location on Russell's two-dimensional model of emotion [35] and subsequently delivered relevant micro-interventions to improve a participant's wellness [36]. By contrast, SATbot attempts to infer the emotion if text input is provided or will ask the user to select from 6 specific emotions, which is easier to choose from. SATbot also provides suggested Self-Attachment protocols instead of micro-interventions. Both interfaces are easy to use, though the user is only required to input text or choose from options for SATbot. ESM and EMMA both utilised empathetic responses for specific quadrants of Russell's model of emotion, whereas SATbot largely only differentiates between positive and negative emotions to produce different questions for the user. Finally, EMMA asks users to provide their emotions several times a day to provide meaningful micro-interventions, whereas SATbot only needs to ask once at the beginning of questioning.

The #MeTooMaastricht chatbot, as developed by Bauer et al. [37], did not alter responses based on a user's emotion and provided the same prompts every time a user interacted with the chatbot, whereas SATbot adjusts its response based on what the user says and the user's identified emotion. Additionally, the model will repeatedly ask users to provide further information if this cannot be identified from earlier user inputs, whereas SATbot will default to asking questions in a specified order or ask users to select from the emotions listed instead of attempting to infer relevant information repeatedly. Therefore, this makes the #MeTooMaastricht chatbot less engaging and less human-like than SATbot, though given that the chatbot aims to assist victims of sexual harassment to get proper assistance, this is secondary to achieving its goal, which the chatbot does despite its repetitive nature.

6.2 Conversations with Participants and Participant Choices

We evaluate the conversations that our 10 participants had with the model, and the choices they made while they used the platform. We also consider the effectiveness of the keyword classifiers utilised to determine emotion and the order of questions to ask participants displaying negative emotions. The latter of these is only used twice for interactions from this trial. Data is from conversations and choices made within the 5-day period of the trial, where participants took part for at least 2 days.

Figure 6.1 shows the protocols attempted by participants during the trial. We see that protocols 3, 7, 10, 12 and 18 are attempted most, where all protocols were suggested for users displaying negative emotions to reinforce positive emotions before further questions were provided to suggest other Self-Attachment protocols. Very few attempt protocols intended to address negative emotions, such as protocol 6 or 11.

Figure 6.2 shows the breakdowns of protocols that made users feel better, worse or had no effect on their emotion. The protocols that were most frequently selected (3, 7, 10, 12, 18) resulted in the large majority of users undertaking those feeling better. Additionally, protocol 11 largely resulted in users feeling better despite the protocol being used to confront negative emotions from current events. Protocol 3 is also the largest protocol that causes participants to say they experienced no change or felt worse. This protocol is proposed when users display negative emotions, or when the user displays a positive emotion. It is also first to appear as a suggested protocol for users with negative emotions.

Therefore, it is presented more frequently, which may result in some users experiencing no change or feeling worse. Additionally, the only other time a user states they feel worse is when they attempt protocol 6, which is somewhat expected as this protocol requires users to recall events that may upset them so they may strengthen the bond with their childhood self. However,the choice to not attempt the positive protocols should have been made clearer to users through adjusting the wording of the choices or the question.

Notably, the two instances of a user feeling worse as a result of protocol 3 happen one after another when the user displayed a negative emotion, which should be addressed by instead suggesting a protocol to directly address negative emotions by forcing the user to address specific events, where Protocols 6 and 11 would address this. We believe that this was an oversight in the code and the model should have immediately asked the user to attempt protocol 6 or 11.

When classifying emotions, the keyword classifiers were 66% accurate at determining the correct emotion. The model most incorrectly classified a user as happy and rarely suggested other emotions incorrectly. This could be because it received the same score as another emotion, so the happy emotion was suggested by default as it is considered earlier by the classifier. Alternatively, the conversation's tone was not recognised by the classifier. An example of a misclassified conversation is: "I feel calm right now. but at the same time very worry about projects i [sic] need to do and I feel I am so behind", which the classifier equally classified as happy and anxious, but defaulted to suggesting the user was happy, which the user was not. Another example of a misclassified conversation is: "I am moody, I feel sorrow and after 5 min I am happy. I am crying for any emotional movies or event. My father was diagnosed with ALS about 4 months ago and I feel so desperate at the bottom of my heart." This was similarly classified as happy. The model did not detect the tone of the conversation in this instance. This could be improved with a better emotion classifier, and this is discussed in Chapter 7.

Based on conversations attempted by users displaying negative emotions, there are only 3 instances where a user says they do not want to attempt the suggested Self-Attachment protocols to reinforce positive emotions through reciting a jolly song or happy phrase. Only two of these users reach the stage where the keyword classifier for question reordering begins analysing the sentence, and in both of these cases, the model correctly identifies that a personal crisis is most likely and asks this first, and in both instances, the user answers the question with yes. Of those who chose to attempt positive protocols and give feedback, there were 18 instances of a user saying they felt better, 5 instances of a user saying they felt no change and 2 instances of a user saying they felt worse. The instances where a user felt worse occurred when the user chose to attempt protocol 3 twice and felt worse both times. The user chose to restart questioning after the first protocol and attempt protocol 3 a second time before trying another protocol, but the user could have been suggested another protocol sooner by the model. To improve this for future trials, the model would suggest a protocol to directly address negative emotions by forcing the user to address specific events, such as Protocols 6 and 11, and follow this with further suggestions if this failed to improve the user's mood. Additionally, these protocols should be prioritised before suggesting further protocols.

6.3 Post-Trial Questionnaire Results

We discuss the results of the 7 participants who completed the post-trial questionnaire, out of the 10 who took part in the trial. The complete results are included in Appendix E.

We observe that over half of the participants disagreed or strongly agreed that the model displayed empathy throughout conversations, which is understandable given that the emotion classifier was imperfect and specific responses were not provided for certain negative emotions. More than half of the participants thought the interface was easy to use and useful, which indicates the platform was largely easy to interact with, though to further increase ease of use, onboarding instructions could be provided to new users and the interface could be adjusted to make it clearer when to provide text input and when to select from an option. More than half thought the model provided suitable protocols as

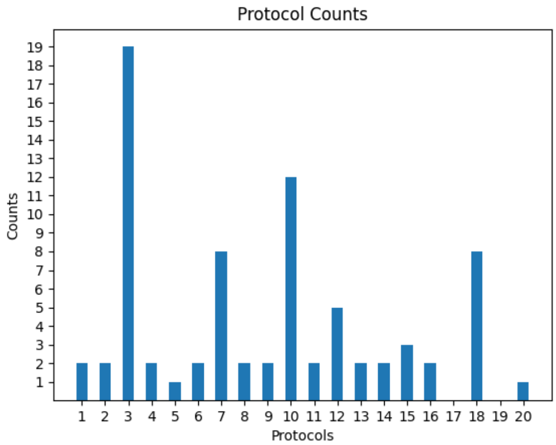

Figure 6.1: Counts of protocols attempted by participants (not divided on feedback).

suggestions for negative emotions, whereas more than half thought the model did not provide suitable protocols as suggestions for positive or neutral emotions. This is understandable as we designed the model to randomly provide protocols for users with positive or neutral emotions until the user has attempted at least 10, at which point we begin to suggest protocols that the user has not attempted as frequently to make them more well-rounded to ultimately be able to better maintain their emotional state when they feel negative emotions. Therefore, a longer trial would allow us to understand how to further improve this area. Similarly to empathy, more than half thought the model was not engaging during conversations, and this could be improved by using a range of prompts for each response the model gives as opposed to using the same prompt. While SATbot does utilise more cases than the #MeTooMaastricht chatbot, based on different events and emotional states, it does not differentiate between different emotions in conversations, which leads to a loss of empathy that can be resolved with appropriately selected responses. Only 2 of the 7 participants thought the model was more useful at providing protocol suggestions than them choosing themselves from workshop materials. When Woebot was trialled, the group using Woebot had a greater reduction in depression than the group using information alone, and while these results are not directly comparable to SATbot, we expected that the majority of users would find the model to be more useful at providing suggestions, similar to how Woebot outperformed reading the information. We did not utilise a control group of participants using the workshops alone as a result of a small number of participants taking part, so we cannot directly compare the impact of Woebot and SATbot here, but these preliminary results indicate that SATbot should be extended to utilise further cases for both the positive/neutral and negative cases to make this more useful for participants than choosing information themselves. Finally, only 3 out of 7 participants thought the chatbot provided useful suggestions when they chose to recall negative events before using the platform, whereas 3 neither agreed nor disagreed and the remaining 1 disagreed. This supports the need for SATbot to utilise more cases to be able to provide more useful suggestions. However, we previously witnessed that few participants reached the greater range of questions for specific negative cases. If the trial was held for longer, we may have seen more useful suggestions being provided for participants. Immediately suggesting negative protocols sooner if the individual feels worse after attempting positive protocols may also move participants towards feeling the chatbot is more useful.

We discuss some written feedback for the last 3 questions in the questionnaire, which ask the user to describe the overall emotional impact of their experience and provide further suggestions or comments. Some users commented on the lack of empathy and engagingness of the platform and how responses can be repetitive after a while, and this can be resolved through adding further responses that the model can choose from at each stage. Others enjoyed using the platform and found it useful as if SATbot was a friend that provided all the solutions. Other users suggested a wider range of emotions should be used, with some saying some emotions are not reflected in the options provided, or that sentences are further interpreted to determine an emotion through further input instead of asking a simpler

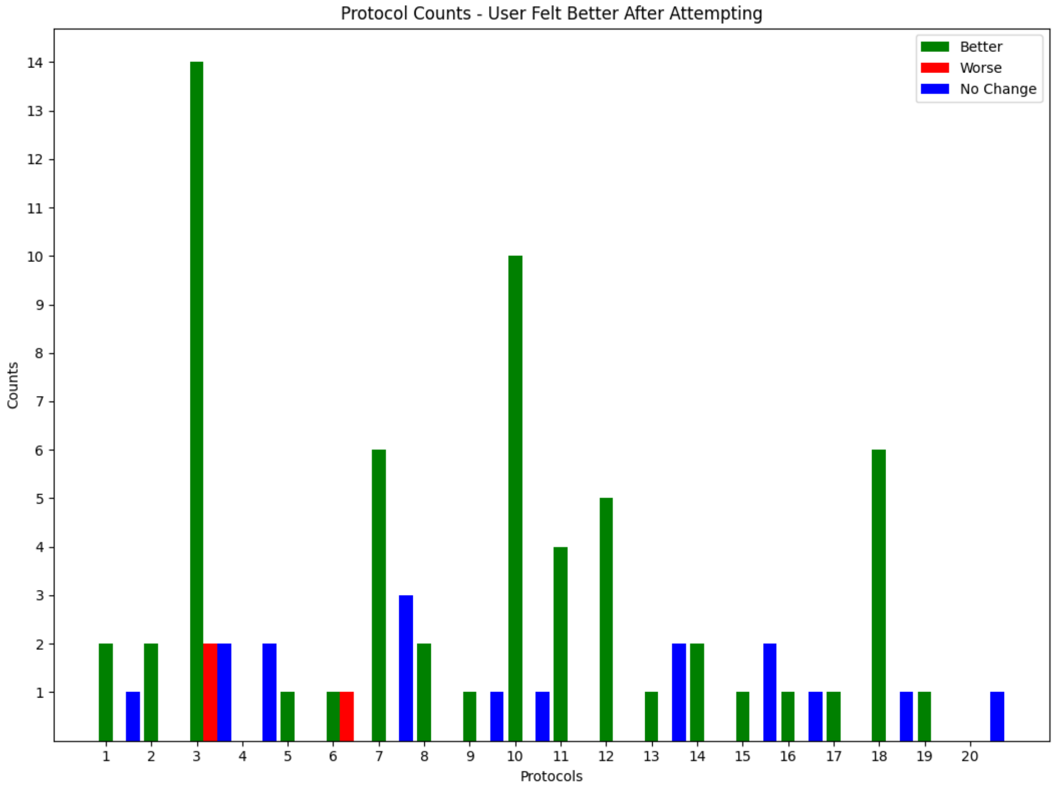

Figure 6.2: Counts of protocols where users stated it made them feel better, had no effect on their emotion or made them feel worse.

question to determine the relevant emotion. Other participants also mentioned how there should be more distinction between different emotions, and this can be improved through responding differently based on the emotion identified. One participant noted that there could be more to being neutral as opposed to suggesting the same as with positive protocols. We could also determine which protocols are most useful for specific emotions and provide these more frequently. Some also commented on how the platform did not recommend many protocols for further growth, and while these were provided for users showing explicit negative emotions that failed to improve following positive protocols, this could be improved by recommending these more for users with positive emotions or introducing a new path to allow users to choose to focus on personal growth and development instead of identifying which is best based on the user's emotion. One user also suggested the use of names to personify the chatbot, and this is an idea we intended to develop as future work. This will be discussed in Chapter 7. One participant said the model should understand concepts instead of the words and vocabularies, so the classifiers could be modified to learn this. Participants also mentioned that integrating the chatbot with information about each protocol so the user can more readily ask for how to undertake each protocol could be provided.

6.4 Evaluation from Clinicians

We discuss feedback from 3 clinicians who evaluated this platform. Unlike the participants, the clinicians did not have prior knowledge of Self-Attachment protocols, though they were provided with the same materials as participants, which included a short video of the platform, instructions on how to use the platform and a short description of protocols as provided in Appendix B. The clinicians were asked to engage with the model for at least 2 days for at least 30 minutes each day, as with the participants. The complete feedback is provided in Appendix E.

We discuss the feedback for the multiple-choice questions. Feedback from the clinicians is consistent with participants on the lack of empathy as 2 clinicians agree that the model did not display empathy, with the remaining clinician saying the model was empathetic. Feedback is divided on the engagingness of the platform. All three clinicians said the interface was easy to use, but only two agreed that it was useful, with the other neither agreeing nor disagreeing. 2 clinicians agreed or strongly agreed that the model provided suitable protocols for any emotion. The clinicians were divided on whether the platform was more useful than using the information alone. Neither clinician that chose to recall a negative event agreed nor disagreed that the chatbot provided useful suggestions when they chose to recall negative events.

We discuss the feedback for the written questions. One clinician describes how they attempted protocols 16, 17 and 19, where their interactions with the platform revealed that they experienced positive emotions which led to these protocols being suggested to them randomly. However, they were unprepared for protocol 16, which is expected as earlier protocols are intended to prepare you for later protocols, which the clinician did not attempt. The explanation for exercise 17 was deemed unclear and could be improved for future trials, and exercise 19 positively empowered the clinician. Another clinician tried some positive protocols during their interactions which occur earlier in the process, which may have led to the lack of evoked emotions. The final clinician found it interesting and found some protocols were easier than others, though they later comment that the information for each protocol should be integrated into the chatbot instead of provided as a separate document and that the text did not seem to influence the choices made, which again highlights that additional cases could be added to suggest specific protocols for both positive and negative emotions so that it is evident that a user's input affects their suggested protocol. They also noted that the range of emotions presented was limited. One clinician suggested expanding the suggestions after asking one question, which suggests we could provide further suggestions in the case where the user experiences a positive emotion to match the suggestions we provide for negative emotions. Another clinician highlights that if someone needs immediate support, the chatbot should highlight how they can contact someone for assistance, similar to how the #MeTooMaastricht chatbot operates. The clinician also notes that depressed people may become irritated or frustrated when presented with the task of imagining happy memories or activities, so the process could be adjusted when considering depressed participants by first asking the participants to complete a questionnaire and bypassing this step if their depression score from the questionnaire is sufficiently high.

Chapter 7

Future Work

Chapter 8

Conclusion

Bibliography

[1] Dinan E, Roller S, Shuster K, Fan A, Auli M, and Weston Jason. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 4, 11, 19, 20, 21, 22, 30

[2] Song Y, Yan R, Li X, Zhao D, and Zhang M. Two are better than one: An ensemble of retrieval- and generation-based dialog systems. *ArXiv*, abs/1610.07149, 2016. 4, 18, 20, 21, 31

[3] Garg S, Rish I, Cecchi G, Goyal P, Ghazarian S, Gao S, Steeg G, and Galstyan A. Modeling Psychotherapy Dialogues with Kernelized Hashcode Representations: A Nonparametric Information-Theoretic Approach, 2019. *arXiv preprint arXiv: 1804.10188*. 4, 19, 21, 22, 30

[4] Rashkin H, Smith E, Li M, and Boureau Y. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset, 2019. *arXiv preprint arXiv: 1811.00207*. 4, 21, 22, 23, 24, 30, 31

[5] Li Q, Li P, Chen Z, and Ren Z. Towards Empathetic Dialogue Generation over Multi-type Knowledge, 2020. *arXiv preprint arXiv: 2009.09708*. 4, 23, 24, 30

[6] Krause B, Gotmare A, McCann B, Keskar N, Joty S, Socher R, and Rajani N. GeDi: Generative Discriminator Guided Sequence Generation, 2020. *arXiv preprint arXiv: 2009.06367*. 4, 24, 26, 27, 30

[7] Madotto A, Ishii E, Lin Z, Dathathri S, and Fung P. Plug-and-Play Conversational Models, 2020. *arXiv preprint arXiv: 2010.04344*. 4, 25, 26, 27, 30

[8] Xu J, Ju D, Li M, Boureau Y, Weston J, and Dinan E. Recipes for Safety in Open-domain Chatbots, 2020. *arXiv preprint arXiv: 2010.07079*. 4, 25, 26, 27, 28, 30

[9] Schultz J. *Gestalt therapy explained: History, definition and examples [Internet].*, 2020. URL https://positivepsychology.com/gestalt-therapy/. [Accessed 8th June 2021]. 4, 53

[10] Sample I. *Covid poses "greatest threat to mental health since second world war". Guardian[Internet]*, 2020. URL http://www.theguardian.com/society/2020/dec/27/covid-poses-greatest-threat-to-mental-health-since-second-world-war. [Accessed 17th January 2021]. 7

[11] Schraer R. *Covid-19 linked to depression and dementia, study suggests. BBC[Internet]*, 2021. URL https://www.bbc.co.uk/news/health-56650125. [Accessed 9th April 2021]. 7

[12] Fitzpatrick KK, 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 Ment Health*, 4(2):e19, 2017. doi: https://doi.org/10.2196/mental.7785. 7, 15, 16, 38

[13] BBC. *Click - 07/11/2020* [Internet], 7 November 2020. URL https://www.bbc.co.uk/iplayer/episode/m000pc4w/click-07112020. 7

[14] Gazzaley A. *A New Era of Experiential Medicine - The Future of Technology as Cognitive Treatments.* [Presentation] Imperial College London, 11 November 2020. URL https://imperialcollegelondon.app.box.com/s/6gcjglz3pidptxld8cik8a0yg39g2hri. [Accessed 17th January 2021]. 7

[15] Edalat A. Introduction to Self-Attachment and its Neural Basis. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2015. URL http://humandevelopment.doc.ic.ac.uk/papers/ijcnn-2015-a.pdf. 7, 9

[16] Edalat A. Self-Attachment: A New and Integrative Psychotherapy. [Presentation] Institute of Psychiatry, King's College London, 2013. URL http://humandevelopment.doc.ic.ac.uk/papers/iop-talk.pdf. 9

[17] Edalat A. Self-Attachment: A Holistic Approach to Computational Psychiatry. In *Computational Neurology and Psychiatry*, 2016. URL http://humandevelopment.doc.ic.ac.uk/papers/cnp-2016.pdf. 9

[18] Tavernor J. Cross-corpus speech and textual emotion learning for psychotherapy, 2020. URL https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1920-ug-projects/distinguished-projects/Cross-corpus-Speech-and-Textual-Emotion-Learning-for-Psychotherapy.pdf. 10

[19] Brown P, Della Pietra V, deSouza P, Lai, J and Mercer, R. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–480, 1992. URL https://www.aclweb.org/anthology/J92-4003. 10, 11

[20] Bengio Y, Ducharme R, Vincent P, and Jauvin C. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155, 2003. URL https://jmlr.csail.mit.edu/papers/v3/bengio03a.html. 10, 11

[21] Serban I, Sordoni A, Bengio Y, Courville A, and Pineau J. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models, 2016. *arXiv preprint arXiv: 1507.04808*. 10

[22] Olah C. *1. Understanding LSTM Networks* [Internet]. URL http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed 17th January 2021]. 11

[23] Hochreiter S and Schmidhuber J. Long short-term memory. *Neural Comput.*, 9(8):1735–80, Nov 1997. doi: https://doi.org/10.1162/neco.1997.9.8.1735/. 11

[24] Britz D. *Recurrent Neural Network Tutorial, Part 4 – Implementing a GRU/LSTM RNN with Python and Theano* [Internet], 27th October 2015. URL http://www.wildml.com/2015/10/recurrent-neural-network-tutorial-part-4-implementing-a-grulstm-rnn-with-python-and-theano/ [Accessed 17th January 2021]. 11

[25] Sutskever I, Vinyals O, and Le Q. Sequence to Sequence Learning with Neural Networks, 2014. *arXiv preprint arXiv: 1409.3215*. 11

[26] Weston J, Chopra S, and Bordes A. Memory Networks, 2015. *arXiv preprint arXiv: 1410.3916*. 11

[27] Sukhbaatar S, Szlam A, Weston J, and Fergus R. End-To-End Memory Networks, 2015. *arXiv preprint arXiv: 1503.08895*. 11

[28] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, and Polosukhin I. Attention Is All You Need, 2017. *arXiv preprint arXiv: 1706.03762*. 12

[29] Devlin J, Chang M-W, Lee K, and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. *arXiv preprint arXiv: 1810.04805*. 12

[30] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, and Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Sep 2019. ISSN 1460-2059. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682. 12

[31] Ewbank M, Cummins R, Tablan V, Catarino A, Bateup S, Martin A, and Blackwell A. Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. *JAMA Psychiatry*, 77(1):35–43, 2020. doi: https://doi.org/10.1001/jamapsychiatry.2019.2664. 12, 13

[32] Ewbank M, Cummins R, Tablan V, Catarino A, Buchholz S, and Blackwell A. Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, pages 33–51, 2020. doi: https://doi.org/10.1080/10503307.2020.1788740. 13

[33] Rojas-Barahona L, Tseng B, Dai Y, Mansfield C, Ramadan O, Ultes S, Crawford M, and Gasic M. Deep learning for language understanding of mental health concepts derived from Cognitive Behavioural Therapy, 2018. *arXiv preprint arXiv: 1809.00640*. 13, 14

[34] Althoff T, Clark K, and Leskovec J. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health, 2016. *arXiv preprint arXiv: 1605.04462*. 14

[35] Ghandeharioun A, McDuff D, Czerwinski M, and Rowan K. Towards Understanding Emotional Intelligence for Behavior Change Chatbots, 2019. *arXiv preprint arXiv: 1907.10664*. 16, 39

[36] Ghandeharioun A, McDuff D, Czerwinski M, and Rowan K. EMMA: An Emotion-Aware Wellbeing Chatbot. In *International Conference on Affective Computing and Intelligent Interaction*, September 2019. URL https://www.microsoft.com/en-us/research/publication/emma-an-emotion-aware-wellbeing-chatbot/. 16, 17, 39

[37] Bauer T, Devrim E, Glazunov M, Jaramillo W, Mohan B, and Spanakis G. #MeTooMaastricht: Building a chatbot to assist survivors of sexual harassment, 2019. *arXiv preprint arXiv: 1909.02809*. 17, 39

[38] Liu C, Lowe R, Serban I, Noseworthy M, Charlin L, and Pineau J. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation, 2017. *arXiv preprint arXiv: 1603.08023*. 18, 19, 20, 22

[39] Caglayan O, Madhyastha P, and Specia L. Curious Case of Language Generation Evaluation Metrics: A Cautionary Tale, 2020. *arXiv preprint arXiv: 2010.13588*. 18, 20, 22

[40] Holtzman A, Buys J, Forbes M, Bosselut A, Golub D, and Choi Y. Learning to Write with Cooperative Discriminators, 2018. *arXiv preprint arXiv: 1805.06087*. 24

[41] Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, Laroussilhe Q, Gesmundo A, Attariyan M, and Gelly S. Parameter-Efficient Transfer Learning for NLP, 2019. *arXiv preprint arXiv: 1902.00751*. 25

[42] Miller A, Feng W, Fisch A, Lu J, Batra D, Bordes A, Parikh D, and Weston J. ParlAI: A Dialog Research Software Platform. *arXiv preprint arXiv:1705.06476*, 2017. 26

[43] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management, 2017. 31

[44] Oseberg F. *react-chatbot-kit docs [Internet].*, 2021. URL https://fredrikoseberg.github.io/react-chatbot-kit-docs/. [Accessed 8th June 2021]. 34

Appendix A

Ethical Checklist

	Yes	No
Section 1: HUMANS		
Does your project involve human participants?	X	
Section 2: PROTECTION OF PERSONAL DATA		
Does your project involve personal data collection and/or processing?	X	
Does it involve the collection and/or processing of sensitive personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?	X	
Does it involve processing of genetic information?		X
Does it involve tracking or observation of participants? It should be noted that this issue is not limited to surveillance or localization data. It also applies to Wan data such as IP address, MACs, cookies etc.	X	
Does your project involve further processing of previously collected personal data (secondary use)? For example Does your project involve merging existing data sets?		X
Section 3: ANIMALS		
Does your project involve animals?		X
Section 4: DEVELOPING COUNTRIES		
Does your project involve developing countries?		X
If your project involves low and/or lower-middle income countries, are any benefit-sharing actions planned?		X
Could the situation in the country put the individuals taking part in the project at risk?		X
Section 5: ENVIRONMENTAL PROTECTION AND SAFETY		
Does your project involve the use of elements that may cause harm to the environment, animals or plants?		X
Does your project involve the use of elements that may cause harm to humans, including project staff?		X
Section 6: DUAL USE		
Does your project have the potential for military applications?		X
Does your project have an exclusive civilian application focus?	X	
Will your project use or produce goods or information that will require export licenses in accordance with legislation on dual use items?		X
Does your project affect current standards in military ethics – e.g., global ban on weapons of mass destruction, issues of proportionality, discrimination of combatants and accountability in drone and autonomous robotics developments, incendiary or laser weapons?		X

	Yes	No
Section 7: MISUSE		
Does your project have the potential for malevolent/criminal/terrorist abuse?		X
Does your project involve information on/or the use of biological-, chemical-, nuclear/radiological-security sensitive materials and explosives, and means of their delivery?		X
Does your project involve the development of technologies or the creation of information that could have severe negative impacts on human rights standards (e.g. privacy, stigmatization, discrimination), if misapplied?		X
Does your project have the potential for terrorist or criminal abuse e.g. infrastructural vulnerability studies, cybersecurity related project?		X
SECTION 8: LEGAL ISSUES		
Will your project use or produce software for which there are copyright licensing implications?		X
Will your project use or produce goods or information for which there are data protection, or other legal implications?		X
SECTION 9: OTHER ETHICS ISSUES		
Are there any other ethics issues that should be taken into consideration?		X

Appendix B

Self-Attachment Protocols and Descriptions

Short descriptions for the 20 Self-Attachment protocols are provided here. The descriptions have been adapted from workshops used for Self-Attachment therapy sessions.

Type 1 exercise: Connecting with the Child

Try to imagine the happy childhood photo/avatar and reflect on relevant positive affects, then imagine the unhappy photo and relevant negative affects. Repeat many times until this is easy to do. Try to imagine the child as you were is near (either happy or unhappy state), and then imagine that we are embracing/cuddling the child. You can also imagine playing with the child.

Type 2 exercise: Laughing at our two childhood pictures

Begin by laughing at the childhood pictures, then think about why we laugh at these pictures. Remember that we do not laugh at them to ridicule. This process will allow us to teach our childhood self to laugh.

Type 3 exercise: Falling in love with the child

While looking at the happy childhood photo, recite selected happy love songs and imagine that you are establishing a deep emotional bond with the childhood self. Then sing with a loud voice, gradually using your whole body as if dancing with the child and having a loving dialogue.

Type 4 exercise: Vow to adopt the child as own child

You imaginatively adopt your childhood self as your own child, loudly pledging to consistently support your child in any way possible. The pledge must be life-long and must be reinforced over time through practising Self-Attachment protocols.

Type 5 exercise: Maintaining a loving relationship with the child

Choose a short phrase e.g. "You are my beloved" and repeatedly utter it while focusing on the happy and unhappy childhood photos. Recite one or two happy love songs, loudly repeating these using your whole body.

Type 6 developmental exercises: An exercise to process the painful childhood events

With closed eyes, recall a painful scene from childhood e.g. emotional or physical abuse in as much detail as possible, and associate the face of the child you were with your unhappy photo. After recalling this event and the related emotions, imagine your adult self approaching and embracing the child like

a parent embracing a child in distress.
While your eyes are still closed, continue to imagine supporting and cuddling the child, loudly supporting them (Examples: “Why are you hitting my child?” and “My darling, I will not let them hurt you any more.”). Massage your face while doing so, which we interpret as cuddling the child.

Type 7 exercise: Protocols for creating zest for life

Using a mirror, imagine the reflection is your childhood self and loudly recite to it your selected happy love songs, using your entire body. Repeat songs and poems in many different circumstances e.g. while walking on the street or doing housework, to be able to integrate them into your life.

Type 8 exercise: Loosening facial and body muscles

You should loosen your muscles at least twice a day as you sing with your face and entire body, as if playing, dancing, laughing and having fun with the child as parents do with children.

Type 9 exercise: Protocols for attachment and love of nature

To create an attachment with nature, you should visit a park or forest and spend time admiring nature, e.g. admiring a beautiful tree, as if seeing its branches and leaves for the first time. Repeat continuously and with different trees until you feel you have formed an attachment with nature. This will help to modulate your emotions and you will want to spend more time with nature each day.

Type 10 exercises: Laughing at, and with one’s self

Begin laughing with yourself about a small accomplishment e.g. in sports, housework, or any other task, however small or unimportant. With every small accomplishment, you should smile as if victorious, and gradually change this smile to laughter, and make this laughter last longer and longer. By practising this you will be able to smile and laugh without ridicule about anything you have said or done in the past while maintaining compassion for your childhood self.

Type 11 exercise: Processing current negative emotions

With closed eyes, imagine the unhappy photo and project the unhappy emotions, e.g. anger, sorrow, towards the photo that represents the child. As with Type 6, we make contact with our adult self to attend to and care for the child to support the child and modulate the child’s negative emotions.
While projecting these negative emotions, loudly reassure the child and massage your own face, which we interpret as cuddling the child. Continue this until you have contained the negative emotions, at which point you can switch to focusing on the happy photo.

Type 12 exercise: Continuous laughter

At a time when you are alone, you should open your mouth slightly, loosen your face muscles, form a Duchenne smile and slowly repeat one of the following phrases as if laughing: eh, eh, eh, eh; ah, ah, ah, ah; oh, oh, oh, oh; uh, uh, uh, uh; or ye, ye, ye, ye.
If a subject is needed for laughter, you can think about the silliness of the exercise. This exercise is a good antidote for stress.

Type 13 exercise: Changing our perspective for getting over negative emotions

To break free of the gravitational field of powerful negative patterns that emerge when we are stuck in the storeroom of negative emotions, or the “psychological abyss”, stare at the black vase in the Gestalt vase picture (below). When you see the white faces, laugh out loud.
Having created a positive powerful pattern of love with the child through previous exercises, you can now depart from the field of negative patterns by singing your happy love song to enter the gravitational field of love for the child instead.

This is like changing our interpretation of the above image and instead of seeing a black vase of negative emotions discovering two white faces, you see the child and the adult self who are now looking at each other.

Figure B.1: Picture of a Gestalt vase, [9]

Type 14 exercise: Protocols for socializing the child

By repeating protocols 1-13 you can reduce negative emotions and increase positive affects. You should gradually be able to perform these exercises with eyes open and can integrate them into your daily life. You should be able to extend compassion for the child to other people. The adult self should become aware of any narcissistic tendencies or anti-social feelings of the child e.g. envy, jealousy, greed, hatred, mistrust, malevolence, controlling behavior and revengefulness.
The adult self can behave like a parent to contain these emotions and discourage anti-social feelings and attitudes of the child by expressing affection to the child and simulating cuddles by massaging your face.
The adult self should try to direct the child's anger and negative energy towards playing, creativity and development. As the child's positive affects increase and his/her negative affects decrease, by expressing positive emotions he/she can attract more positive reactions from others, and in turn gain a more positive outlook toward others.

Type 15 exercise: Recognizing and controlling narcissism and the internal persecutor

The adult self becomes aware of the facets of the trauma triangle: internal persecutor, victim, and rescuer. The adult self examines the effects of the triangle (narcissism, lack of creativity) in daily life and previous experiences.
Your adult self should then review an important life experience and your social and political views as an adult, with awareness of how the internal persecutor operates. Your adult self should then create a list of examples from own experiences about how the internal persecutor operates, and carefully analyse these for examples of being drawn to trauma, being traumatized by the internal persecutor, and projecting the internal persecutor.
You should be able to then re-evaluate your own experiences, control the internal persecutor and narcissism and be able to develop creativity.

Type 16 exercise: Creating an optimal inner model

With awareness of the internal persecutor, we will recognise emotions of the child that were learned from parents or through interactions with them. With the guidance of the adult self, who can transfer compassion for the child to others, the child will learn to avoid projecting the internal persecutor (which would lead to them becoming the victim or rescuer).

Type 17 exercise: Solving personal crisis

As you continue to practice the protocol for modulating negative affects and the protocol for laughter, ask your child the following:

- How can you see the crisis as a way of becoming stronger? (ha ha ha)
- How can you interpret the crisis as a way of reaching your high goal? (ha ha ha)
- Has the internal persecutor been projecting onto others again?

The adult self asks the following questions:

- What is the similarity between this crisis and ones I have faced before?
- How is it similar to the family crisis I experienced as a child?
- Aren't the other person's positive attributes greater than his/her negative ones?
- How would a mature person interpret the crisis in comparison to my child?
- Can I see it from the perspective of someone else?
- Can I put myself in their place and understand their affects?
- Given my new inner working model can I find a way to calm the people involved in the crisis so we can find a better solution for it?
- If not, can I respectfully maintain my distance and end the argument?

Type 18 exercise

(i): Laughing at the harmless contradiction of deep-rooted beliefs

"To those human beings who are of any concern to me I wish suffering, desolation, sickness, ill-treatment, indignities—I wish that they should not remain unfamiliar with profound self-contempt, the torture of self-mistrust, the wretchedness of the vanquished: I have no pity for them, because I wish them the only thing that can prove today whether one is worth anything or not—that one endures."

This is meaningful with, "What doesn't kill me makes me stronger." Nietzsche's wish is funny and a harmless contradiction of our deep-rooted beliefs. As we read the quote above, we remember our past sufferings and begin to laugh out loud when we get to "...I wish suffering..."

(i) continued: Laughing at trauma

First, visualize a painful event that took place in the distant past that you have struggled with for a long time, and despite its painfulness try to see a positive impact it has had. We start with a painful event that happened in the distant past, so that by now we have been able to adjust our negative affects toward it. After repeated daily exercises, once we have experienced the forceful effectiveness of laughing at distant problems, we can gradually begin to laugh at more recent painful memories.

(ii): Laughing at trauma

In expectation of hearing a funny joke we loosen our facial muscles, slightly open our mouths, and to grasp the incongruity in the joke we move our eyebrows up as a sign of surprise. As we repeat the sentences out loud, we slowly begin to laugh as we wait for the second part. And once we get to the first sentence of the second part, which is in complete contrast to our beliefs, we laugh out loud.

Not only should you: bear it, accept it, try to deal with it, tolerate its memory, try harder to endure its memory, adapt yourself to its memory, analyze and understand it and by doing so modulate your negative emotions and learn lessons for the future, try to soften your thoughts, depressive emotions, and anxieties, try to ...

Like Nietzsche's wish consider it a cherished treasure (ha ha ha...), treasure it with great love (ha ha ha...), welcome its challenges with all your heart (ha ha ha...), consider it a good omen with all your heart (ha ha ha...), consider its challenges a great fortune (ha ha ha...), celebrate its memory (ha ha ha...), celebrate its memory with great joy (ha ha ha...), consider it a true love (ha ha ha...), consider it a true love with great passion and intimacy (ha ha ha...) ...

After repeated practice of the laughing exercises you can begin to apply it to things that worry you in the present and the future.

Type 19 exercise: Changing ideological frameworks for creativity

We challenge our usual ideological framework to weaken one-sided patterns and encourage spontaneity and the examination of issues from multiple perspectives. Practice with subjects that you have deep-rooted beliefs and are excited about e.g. anything from political/social issues to ideas on marriage and sexuality. For instance, examine the topic of racism and consider whether you have any latent racism and consider this subject in the dual role of proponent and opponent.
Repeat with topics where you may have stronger views e.g. marriage and sexual orientation. If you are politically in the center, consider the subject both from a leftist and rightist point of view and try to understand both sides of the issue and see the subject from three perspectives.

Type 20 exercise: Affirmations

Put together a list of instructive sayings by different important figures. Choose ones that have an impact on you from the start and can provide you with strength in the long path for reaching your ultimate goal. Read them out loud.

A few examples:

- "My formula for greatness in a human being is Amor Fati: that one wants nothing to be other than it is, not in the future, not in the past, not in all eternity." (Nietzsche)
- "I assess the power of a will by how much resistance, pain, torture it endures and knows how to turn it to its advantage." (Nietzsche)
- Life is not easy. At times we inevitably suffer from hopelessness and paranoia unless if we have an ideal goal that helps us surpass suffering, weakness, and betrayals." (Bronstein)

Appendix C

Survey Question Templates

Appendix D

Example Model Conversations

Positive Conversation
Model:How are you feeling? Please tell me in as much detail as you can, including relevant situations that made you feel this way.
User:I'm feeling very good. Because I'm in charge of how I feel today.
Model:Based on what you have said, I believe you are: Happy. Is this correct?
User:Yes
Model:That's good! Let me suggest a protocol you can attempt.
Model: Here are my suggestions below. Based on these suggestions, which protocol would you like to attempt? Please choose one of the suggestions or type your own protocol number (1-20) below. (Suggestions: 17: Solving Personal Crises [Week 7] , 2: Laughing at our Two Childhood Pictures [Week 1], 4: Vow to Adopt the Child as Your Own Child [Week 2])
User:2: Laughing at our Two Childhood Pictures [Week 1]
Model:Please try to perform this protocol now. Press Continue when you have finished and you are ready to proceed.
User:Continue
Model:Do you feel better or worse having taken this protocol?
User:Better
Model:Would you like to try another protocol?
User:No
Model:Thank you for taking part! Please refresh the page to start again.

Figure D.1: Positive conversation within the model. Some content has been adapted from conversations that occurred within the 5-day trial.

Negative (Attempting Positive Protocols)
Model:How are you feeling? Please tell me in as much detail as you can, including relevant situations that made you feel this way.
User:I feel depressed and frustrated
Model:Based on what you have said, I believe you are: Sad. Is this correct?
User:Yes
Model:Thank you for letting me know. Now we will ask some questions to understand your situation.
Model: The following protocols may help: Type 3 (Singing a jolly love song to the child), Type 7 (Laughter for creating zest for life), Type 10 (Laughing at, and with one's self), Type 12 (Continuous Laughter), Type 18 (Laughing at misfortunes). Are you able to take the challenge now to sing your favourite jolly love song or to laugh as in the following protocols to reduce your negative emotions?
User:Yes
Model:That's good! Please choose one of these protocols to continue, or type your own protocol number (1-20) below.
(Suggestions: 3: Falling in Love with the Child [Week 2], 7: Protocols for Creating Zest for Life [Week 4], 10: Laughing at, and with One's Self [Week 5], 12: Continuous Laughter [Week 6], 18: Laughing at the Harmless Contradiction of Deep-Rooted Beliefs/Laughing at Trauma [Week 8])
User:18: Laughing at the Harmless Contradiction of Deep-Rooted Beliefs/Laughing at Trauma [Week 8]
Model:Please try to perform this protocol now. Press Continue when you have finished and you are ready to proceed.
User:Continue
Model:Do you feel better or worse having taken this protocol?
User:Better
Model:Would you like to try another protocol?
User:No
Model:Thank you for taking part! Please refresh the page to start again.

Figure D.2: Negative conversation (positive protocols attempted).

Negative (Not Attempting Positive Protocols) Model:How are you feeling? Please tell me in as much detail as you can, including relevant situations that made you feel this way.
User: *provides no input*
Model:Thank you. Now please select an emotion below that describes how you feel, based on what you have said.
User:Angry
Model:Thank you for letting me know. Now we will ask some questions to understand your situation.
Model: The following protocols may help: Type 3 (Singing a jolly love song to the child), Type 7 (Laughter for creating zest for life), Type 10 (Laughing at, and with one's self), Type 12 (Continuous Laughter), Type 18 (Laughing at misfortunes). Are you able to take the challenge now to sing your favourite jolly love song or to laugh as in the following protocols to reduce your negative emotions?
User:No
Model:Was your emotion caused by a specific event (or events)?
User:Yes
Model:Was this caused by a recent or distant event?
User:Recent
Model:Have you recently attempted Protocol 11: Processing Current Negative Emotions [Week 5] and found this reignited emotions of old events?
User:No
Model:Have you strongly felt or expressed any of the following emotions towards someone: Envy, jealousy, greed, hatred, mistrust, malevolence, or revengefulness?
User:Yes
Model:Here are my suggestions below. Based on these suggestions, which protocol would you like to attempt? Please choose one of the suggestions or type your own protocol number (1-20) below. (Suggestions: 11: Processing Current Negative Emotions [Week 5], 13: Changing Our Perspective for Getting Over Negative Emotions [Week 6])
User:11: Processing Current Negative Emotions [Week 5]
Model:Please try to perform this protocol now. Press Continue when you have finished and you are ready to proceed.
User:Continue
Model:Do you feel better or worse having taken this protocol?
User:Better
Model:Would you like to try another protocol?
User:No
Model:Thank you for taking part! Please refresh the page to start again.

Figure D.3: Negative conversation (not attempting positive protocols).

Appendix E

Full Trial Questionnaire Results

E.0.1 Participant Results

Question	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
The model displayed empathy in its responses throughout my conversations.	1/7 (14%)	3/7 (43%)	1/7 (14%)	2/7 (29%)	0/7 (0%)
The interface was easy to use.	0/7 (0%)	0/7 (0%)	2/7 (29%)	3/7 (43%)	2/7 (29%)
The interface was useful.	0/7 (0%)	1/7 (14%)	2/7 (29%)	3/7 (43%)	2/7 (29%)
The model provided suitable protocols as suggestions when I displayed a negative emotion.	0/7 (0%)	1/7 (14%)	2/7 (29%)	3/7 (43%)	1/7 (14%)
The model provided suitable protocols as suggestions when I displayed a positive or neutral emotion.	0/7 (0%)	1/7 (14%)	3/7 (43%)	2/7 (29%)	1/7 (14%)
The model was engaging during conversations.	0/7 (0%)	4/7 (57%)	1/7 (14%)	1/7 (14%)	1/7 (14%)
The model was more useful at providing appropriate protocol suggestions than me choosing from the workshop materials.	0/7 (0%)	3/7 (43%)	2/7 (29%)	1/7 (14%)	1/7 (14%)
When I chose to recall a negative event before using the platform, the chatbot provided useful suggestions to assist me with.	0/7 (0%)	1/7 (14%)	3/7 (43%)	3/7 (43%)	0/7 (0%)

Figure E.1: Results of multiple-choice questions from the 7 participants who completed the post-trial questionnaire. Percentages rounded to the nearest number. The last question also contained an N/A option, which was not selected.

E.0.2 Clinician Results

Question
Q: Please describe the overall emotional impact of your experience:
1) The program was very poor in guessing my emotions after writing my feelings. For instance, I wrote a long paragraph about a totally negative emotion, then the program was asking: 'you are feeling happy, is that right?'
2) The question is not quite clear. Assuming you mean the overall impact of using this platform in improving my emotional state, practicing the protocols generally have positive impact on me. It was the same using this platform. I was able to get more out of the protocols using this platform, by the mere fact that it provided a compact and focused structure for my practice. I wouldn't be able to say with confidence how much of this impact is specific to this tool and how much is merely due to the more focused practice of SAT protocols.
3) I did not find it very helpful. Only a good review of protocols for me
4) I really enjoyed chatting with it. It felt like I have a friend who has all the solutions.
5) It is a useful tool to remind right protocols for specific feelings. It was easy to use.
6) I really liked this as a useful tool to practice the protocols.
7) I feel that the model is more rule based (contain if else) rather than a trained machine. The model continuously suggests the same sentences and same protocol over and over again through the conversation and also exact same sentences/grammar in different conversations.

Figure E.2: Results of first written question from the 7 participants who completed the post-trial questionnaire.

Question
Q: Do you have any further suggestions for how the platform can be improved?
1) as a first draft, it is fine, however there is a huge room to improve
2) I really enjoyed using this platform, regardless of the level of its technical effectiveness (see my response above). I believe this tool has a great potential and might be the best way to guide participants for both the short- and long-term practice of SAT. I recognize that this is the very first prototype and it was certainly helpful even at this stage. While my feedback below is based on my experience with the current version, my suggestions are oriented towards perfecting the platform in the long run: • The chatbot is doing a good job at establishing basic empathy. It has a lot of space for further improvement though. A few ways it can be improved: I suggest using actual names to create a better feeling of connection. SATBot can be replaced by a human name and the bot can definitely use the participant's name to communicate with them. Another aspect to show empathy is how the follow-up questions are formulated after the participant initially tells you about their feeling. You can work (and draw on the literature for verbal empathy) on how the bot can respond to each specific feeling. • The bot starts by asking about details of how we feel, but it does not seem to interpret it with sufficient effort. In most cases, the second question was to ask me to choose my emotion among the predetermined options. Again, I understand that this is not simple and will take a lot of work, but ideally, it would be best if the system can interpret the answer and ask a followup question for further clarification, instead of disregarding the detailed response and asking the same question in a more rudimentary format. • The choices for emotions and also the choices for reporting the outcome are limited and limiting. There are a few other basic emotions that could be very common for the practitioners of SAT and were not reflected in the current set of choices (for example shame, guilt, frustration, ...). Also there could be a lot more to being 'neutral'. In general, I wasn't able to express my state of mind or emotions with accuracy, which in turn affected the accuracy of the recommendations I received from the chatbot. • Related to my last comment, in general I felt the tool was not well designed to recommend 'developmental' or 'growth-oriented' protocols as much as I wanted. These (mostly protocols from 11-19) are a very important group of protocols for long-term practice and would need their own specific prompts and recommendation algorithm.
3) It was very limited. Needed to be more comprehensive in interpreting emotions we share with chatbot
4) Maybe more emotional state phrases can be put in for it so it wouldn't force you to choose from happy, sad, neutral, angry ones.
5) I think we should have more than just negative and positive emotions. We can have broader range of emotions.
6) Yes. I think the conversation aspect can improve. It gets repetitive quickly.
7) The model should be trained. The model should understand concepts and not just the words/vocabularies.

Figure E.3: Results of second written question from the 7 participants who completed the post-trial questionnaire.

Question
Q: Please provide any additional comments you have. 1) It would better instead of sending the description of each practice as a separate file, somehow include it in the platform to have them all in front of the users ayes to make it easier to use 2) Some other very useful addition to this tool would be to: • Enable it to remind the participant of each protocol, e.g. be able to respond to something like 'Could you remind me what protocol 13 was about?' • Provide step-by-step guidance of how to do a recommended protocol. Instead of asking us to 'go, do it and come back', this tool in its perfection can actually be a guide and mentor on how to apply the protocol to each specific problem.
3) [added from email, intended for questionnaire] I believe the whole idea behind this project is very good and if this model becomes more developed, that would be definitely very helpful. However, at this level,I did not find this protocol very helpful as it interpretes my emotions into a few very specific categories. For example, when I write I feel disappointed it asks me do you feel sad? to me , sad is not necessarily equivalent to sad or angry. On the other hand, since I did the whole workshop in the past, it helped me to refresh my memory about protocols. I believe there was a mismatch with the name of the protocols and weeks (not sure though) BTW, i only did it 3 times and I assume it became less than 30 minutes for each session. Unfortunately I could not spend more time.
4) N/A
5) N/A
6) Thanks for sharing it.
7) It is also nice that the model gives some hints or a step by step guideline for each protocol so the user does not need another documents to review the protocol each time during the conversation.

Figure E.4: Results of third written question from the 7 participants who completed the post-trial questionnaire.

Question	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree	N/A
The model displayed empathy in its responses throughout my conversations.	0/3 (0%)	2/3 (66%)	0/3 (0%)	1/3 (33%)	0/3 (0%)	-
The interface was easy to use.	0/3 (0%)	0/3 (0%)	0/3 (0%)	1/3 (33%)	2/3 (66%)	-
The interface was useful.	0/3 (0%)	0/3 (0%)	1/3 (33%)	2/3 (66%)	0/3 (0%)	-
The model provided suitable protocols as suggestions when I displayed a negative emotion.	0/3 (0%)	0/3 (0%)	1/3 (33%)	2/3 (66%)	0/3 (0%)	-
The model provided suitable protocols as suggestions when I displayed a positive or neutral emotion.	0/3 (0%)	1/3 (33%)	0/3 (0%)	1/3 (33%)	1/3 (33%)	-
The model was engaging during conversations.	0/3 (0%)	1/3 (33%)	1/3 (33%)	1/3 (33%)	0/3 (0%)	-
The model was more useful at providing appropriate protocol suggestions than me choosing from the workshop materials.	0/3 (0%)	1/3 (33%)	1/3 (33%)	1/3 (33%)	0/3 (0%)	-
When I chose to recall a negative event before using the platform, the chatbot provided useful suggestions to assist me with.	0/3 (0%)	0/3 (0%)	2/3 (66%)	0/3 (0%)	0/3 (0%)	1/3 (33%)

Figure E.5: Results of multiple-choice questions from the 3 clinicians who completed the post-trial questionnaire. Percentages rounded to the nearest number. Only the last question contained an N/A option.

Question
Q: Please describe the overall emotional impact of your experience:
1) No emotion evoked.
2) When carrying out the protocol 16 exercise, I became very upset when getting in touch with my inner child; I felt that my adult self was unequipped to deal with the strength of negative emotions and provide relief through compassion. I was unclear about the language in exercise 17. I felt positively empowered by doing the exercise 19.
3) It was interesting. As someone previously unfamiliar with the protocols I found some more intuitive and easier to engage with than others. I was sceptical about the added value of the chatbot in recommending them but it did actually help.

Figure E.6: Results of first written question from the 3 clinicians who completed the post-trial questionnaire.

Question
Do you have any further suggestions for how the platform can be improved?
1) [added from email, intended for questionnaire] All looks good to me, as intended for a general (non-clinical) population already familiar with the SAT procedures. The Chat Bot responded but when I said "I am sad 'cos I lost my job" it did not reflect that. Is the goal to make the system more responsive in this regard? I assume yes! In due course - when moving towards a clinical (depressed/ distressed?) cohort - more consideration of risk estimation, monitoring and mitigation will be needed. For instance, the PHQ-9 has a question "Have you had thoughts that you would be better off dead or of hurting yourself in some way". Clinicians typically focus on this and explore suicidality next. Assuming the project is going towards the clinical interface we will need to consider the risk issues. Consequently, clearer pathways towards seeking support e.g. if a participant becomes distressed, say, at 3am - or 11am- where do they go? I should add these are soluble issues. In addition to exclusion/ inclusion parameters e.g. avoiding recruitment of the high risk/ impulsive / vulnerable, providing details on helplines, crisis team etc can mitigate. Another issue - more anecdotal based on my own experience - is that depressed people can become irritated or frustrated when asked to imagine happy memories or joyful activities. More fundamentally - and empirically based- is that the core cognitive theory of depression posits that it endures due to autobiographical memory biases for recall of personally relevant hedonic memories (depressed folk can recall more easily general memories " I was always sad - or sometimes happy - as a child". But not " I remember getting a beautiful kitten for my 8th Birthday and I was soooo happy!!" Again, this is for future consideration and one thing I can help with by placing the proposal in this broader context. For more info. see this 2016 article I edited: Frontiers | The neuroscience of positive memory deficits in depression | Psychology (frontiersin.org) <https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01295/full> This places the dopaminergic modulation of memory encoding centre stage in depression. Hypo activation leads to weaker encoding in Long Term Memory (in simple terms, speaking as a clinical psychologist!)
2) Suggestions on which exercises to use came after one question about how I was feeling. I did not think that all the suggestions were appropriate. Further probing / follow on questions may have resulted in suggesting different protocols.
3) Would it be possible to embed the protocol prompts or add hyperlinks within the chatbot function? It was annoying to need to consult a separate document to access the protocols.

Figure E.7: Results of second written questions from the 3 clinicians who completed the post-trial questionnaire.

Question
Please provide any additional comments you have. 1) N/A
2) I think the simple process of coming up with suggestions after asking one question can be improved /expanded.
3) The request to provide text information about how we were feeling at the beginning did not seem to influence the suggestions made, which seemed to be based on the selection of 1 of 5 affects. The range of emotions was limited.

Figure E.8: Results of third written question from the 3 clinicians who completed the post-trial questionnaire.